Teacher accountability policy and sociocultural context: A cross-country study focusing on Finland and Singapore

HWA, YUE-YI

St Catharine's College,
University of Cambridge

Supervisors:
Panayiotis Antoniou & Ricardo Sabates

Faculty of Education

September 2019

This dissertation is submitted for the degree of Doctor of Philosophy.

Preface

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the relevant Degree Committee.

HWA YUE-YI

Length: 79,626 words
(excluding front matter, appendices, reference list, footnotes, tables, and figures)

Teacher accountability policy and sociocultural context: A cross-country study focusing on Finland and Singapore

Hwa, Yue-Yi

Abstract

In this thesis, I address two polarised debates in education policy: how teachers should be held accountable, and whether 'best practices' from high-performing education systems should be adopted in other countries. I construct a conceptual framework that maps the intended effects of teacher accountability instruments on student outcomes, via changes in teacher motivation. In this framework, the efficacy of any teacher accountability instrument depends partly on its compatibility with sociocultural context. This is partly because an accountability instrument will only influence a teacher's motivation if the teacher regards the instrument as sufficiently meaningful, legitimate, or otherwise persuasive—and perceptions of meaning and legitimacy can be shaped by sociocultural patterns. To test this framework, I draw on two sets of empirical sources. Firstly, I use multilevel modelling to analyse cross-country survey data on education (e.g. PISA) and culture (e.g. the World Values Survey). I find that the relationship between teacher accountability instruments and student outcomes in these datasets varies with one aspect of sociocultural context, i.e. the strength of adherence to civic norms. Secondly, I analyse semi-structured interviews that I conducted with 12 lower secondary school teachers in Finland and 12 in secondary school teachers in Singapore. I find that teacher accountability instruments can have considerable effects (and side effects) on teacher motivation. However, interview participants' responses to accountability instruments are strongly influenced by sociocultural context, and Finland's and Singapore's contrasting but comparably effective approaches to teacher accountability are each compatible with their respective sociocultural contexts. Based on these findings, I argue that the efficacy of teacher accountability instruments is contingent on sociocultural context (among other factors). Consequently, an accountability instrument from a top-ranked education system may have null or negative effects if transplanted elsewhere. Instead, teacher accountability policymaking needs to accommodate local sociocultural patterns. To my knowledge, this is the first study to combine cross-country educational and cultural surveys to explore the relationship between teacher accountability and sociocultural context. It is also the first study to conduct a fieldwork-based comparison of teacher accountability in Finland and Singapore.

Table of contents

List of tables

List of figures

List of abbreviations

2PL	two-parameter logistic (an IRT model)
EPMS	Enhanced Performance Management System (Singapore)
ESCS	economic, social, and cultural status (a PISA measure)
EVS	European Values Study
GRM	graded response model (an IRT model)
IEA	International Association for the Evaluation of Educational Achievement
ILSA	international large-scale assessment
IRT	item-response theory
KIKY	*Kilpailukykysopimus* (‘competitiveness pact’, Finland)
OECD	Organisation for Economic Co-operation and Development
PISA	Programme for International Student Assessment
TALIS	Teaching and Learning International Survey
TIMSS	Trends in International Mathematics and Science Study
WVS	World Values Survey

Acknowledgements

This thesis would not have been possible without the support of many people and institutions. I am deeply grateful to all of them.

The 24 teachers who spoke with me in Singapore and Finland offered not only their experiences and expertise as a vital source of research insight, but also a great deal of warmth and good humour.

My thesis supervisor, Panayiotis Antoniou, gave me the freedom to explore my research interests in the directions I chose, but was always willing to offer guidance that challenged and furthered my thinking whenever I needed it. Ricardo Sabates was first a supportive faculty advisor, going far beyond the expectations of the role, before taking over as my supervisor in the last few months, during which he gave invaluable close readings of every chapter draft. Despite only serving as my advisor over the past few months, Rob Gruijters generously gave feedback on Chapter 6. I also benefitted greatly from input from Geoff Hayward at my registration viva.

Jaakko Kauko hosted me as a visiting researcher at the University of Tampere's Faculty of Education while I was conducting fieldwork in Finland in September 2019. Besides connecting me with research contacts, he took the time to discuss local contexts and epistemological conundrums with me. He also translated part of my participant information sheet and commented on my interview strategy as well as on a draft of Chapter 6. In Singapore, Warren Mark Liew shared his observations of local education and educational research, and recommended useful references.

Melanie Ehren kindly invited me to contribute a chapter to the forthcoming volume *Trust, Accountability, and Capacity in Education System Reform: Global Perspectives*, which she is co-editing with Jacqueline Baxter. Her feedback on drafts of the chapter aided theory development in this thesis.

My fellow PhD candidates at the Faculty of Education enriched this process with their acumen, wry wisdom, and camaraderie. I owe a particular debt of gratitude to Tiong Ngee Derk and Charleen Ning Chiong, who gave perceptive input on some of my drafts.

I piloted the interview guide with a Singaporean teacher, a Singaporean ministry official, and a Finnish principal, all of whom gave valuable feedback.

Further input into the interview guide, conceptual framework, and other aspects of the project came from wide-ranging and thoughtful observations from the Teach For All alumni community of practice in education policy. Regular conversations throughout 2017–18 with Liz Pasco Carmona, Lucy Crehan, and Jim O'Conner aided theory refinement and supplied joy.

I also received constructive feedback from participants at the Kaleidoscope Conference, University of Cambridge in May 2018; the EARLI SIG 18 & 23 Conference in August 2018; and CIES in April 2019, especially from the Dissertation Mentoring Workshop.

I have been extremely fortunate in receiving a PhD scholarship from the Gates Cambridge Trust. Besides this generous support from the Trust, I received grants for fieldwork and conferences from the Tunku Abdul Rahman Fund and the Graduate Overseas Bursary Fund at St Catharine's College, as well as from the Faculty of Education and the CIES New Scholars Committee.

I would not have been able to complete this thesis without encouragement from family and friends. Particularly pertinent support came from Thammika Songkaeo, Tracy Tan, and Inez Tan, who successively lent me their guest beds during my Singapore fieldwork in July 2018; and from Shirley Li, who gave feedback on my conceptual framework. Also, I will always value the support of the many people in my immediate and extended networks who helped me to recruit interview participants in Finland and Singapore. (They remain nameless to protect participants' identities.)

Throughout the PhD, Ujval Sidhu-Brar never failed to make me laugh, even on days when I thought I couldn't.

Chapter 1: Introduction

1.1 Why study teacher accountability and sociocultural context?

If you were a secondary school teacher in Singapore, one of your annual responsibilities would be formulating your plans for the year in three areas: student, professional, and organisational outcomes. At the beginning of the year, you would formally discuss these plans with your reporting officer (a teacher holding management-level responsibilities) and record them in a work review form. You would revisit this work review form at the middle and the end of the year in subsequent meetings with your reporting officer, who would also observe at least one of your lessons. Additionally, you would probably receive a scheme of work from your subject head, detailing the topics, objectives, and types of assignments that should be covered each week, and identifying a few points during the year when your students will face internal or external standardised assessments. Toward the end of the year, your reporting officer would confer with other senior teachers in the school at closed-door ranking panel meetings, during which they would award you a grade based on your performance in comparison to that of other teachers at your level of seniority. This performance grade would affect the size of your annual bonus and the speed of your promotion, among other things.

If you were a lower secondary school teacher in Finland, your principal might ask you to submit a basic plan of how you intend to fulfil national curricular objectives throughout the year—but you would be free to modify those plans as the need arose. In the autumn, you would likely have a developmental discussion with the principal, where you would discuss your general well-being and how your work is going. In most schools, however, there would not be any follow-up after this discussion. You would be expected to assess your students' learning, but you would have ample leeway in how you do so. Whether or not any adults are aware of what happens in your classroom would depend on how busy your principal is; how closely you collaborate with other teachers in your subject group (which can vary considerably); how often the special education teacher works alongside you; and what your students tell their parents. The biggest material reward you could expect is a small salary supplement if you take on additional school responsibilities. Conversely, you would not face any performance-based penalties, besides written warnings for failing to deliver your designated lesson hours, or getting fired if you abuse your students or show up at school drunk.

Both the Finnish and the Singaporean approaches to teacher accountability have been lauded internationally as 'best practices', under headlines such as 'Scotland eyes Singapore in "radical" overhaul of teaching career paths' (Hepburn, 2017) or 'Highly trained, respected and free: why Finland's teachers are different' (Crouch, 2015). Yet these two 'best' approaches are clearly disparate. This disparity may be puzzling—but only if education policymaking is viewed as a universal, context-neutral endeavour. Instead, if education policy in general and teacher accountability policy in particular are regarded as a matter of 'what works for whom in what circumstances and in what respects, and how' (Pawson & Tilley, 2004, p. 2), then this disparity is not only unsurprising, but also desirable: if Finland's and Singapore's educational contexts differ, then their accountability policies should reflect salient differences.

In this thesis, I focus on an aspect of context that is central to the efficacy of teacher accountability, i.e. sociocultural context. Although it typically receives less scholarly attention than other similarly important aspects of context, such as material constraints and teachers' pedagogical competence, sociocultural context can help to explain how Finland's and Singapore's divergent teacher accountability approaches are comparably successful.

My analysis also looks closely at teacher motivation: a concept that is implicit in any teacher accountability policy that aims to prompt teachers to work harder or work differently, i.e. to raise or reorient their motivation. Motivation is curiously under-theorised in the teacher accountability literature, despite its inclusion in some prominent frameworks for educational accountability (e.g. Pritchett, 2015; UNESCO, 2017, fig. 1.2). Also, despite an extensive body of psychological research on motivation in work and in education, there is far less research on teacher motivation than on student motivation (Urdan, 2014). Nonetheless, I believe that teacher accountability policies are more likely to influence teacher motivation as intended by policymakers if the policies are tailored to fit sociocultural particularities.

In the following chapters, I investigate whether there is any evidence for this claim, using multilevel modelling of cross-country educational and cultural surveys alongside interviews with teachers in Finland and Singapore. Even though Finland and Singapore are among the most admired and idealised high-performing education systems, this is, to my knowledge, the first comparative study of teacher accountability and sociocultural context in these two countries.[1]

The context of this study

The impetus for this research project comes from two prominent but problematic discussions on education. Firstly, there is the perennially contentious issue of teacher accountability for student outcomes. The debate on teacher accountability is longstanding and global. For example, 'payment by results' was attempted in English schools as early as the 1860s (Jabbar, 2013; Rapple, 1994). Recently, some scholars have noted a growing emphasis on accountability in education policy (Tulowitzki, 2016; Verger & Parcerisa, 2017a) and in public management more generally (Muller, 2018; Pollitt & Bouckaert, 2017). In 2015, routine teacher appraisals directly affected teachers' pay in countries ranging from Chile to Hungary to Singapore (OECD, 2016f).

Still, the popularity of teacher performance management and other associated policies has been neither uniform nor universal. A review of studies on educational decentralisation in the developing world found extensive evidence of decentralisation reforms, which are often associated with accountability reforms, in Central America, but almost none in Northern Africa, East Asia, and the Pacific region (Edwards & DeMatthews, 2014). Also, scholars have noted increasing incidence of external exams, school inspections, and other accountability instruments across European countries—but also significant variation between countries, often tied to historical path dependencies (Altrichter & Kemethofer, 2015; Herbst & Wojciuk, 2017; Hudson, 2007; Maroy, 2009; Mattei, 2012).

It is fairly uncontroversial to say that teachers should be accountable, in a broad sense, for the pivotal work that they do in developing individual well-being, cultural socialisation, and national economic growth—and for the funding that they receive to do so. But there is much less agreement about how teacher accountability instruments can, and should, facilitate optimal student outcomes. While some prominent voices advocate for test-based accountability (e.g. Hanushek, 2019), such policies also face fierce opposition. To illustrate, Finnish educationist Pasi

[1] I searched in ERIC and Scopus on 20 August 2019 for peer-reviewed studies containing the terms 'Finland', 'Singapore', and 'accountability' in their titles, abstracts, or keywords. These searches yielded seven and three results respectively, none of which focused on accountability, or examined in rich empirical detail the relationship between policy implementation and sociocultural context. The closest results were two desk-based studies of curricula in high-performing countries, which discussed some aspects of accountability, e.g. student assessment (Creese, Gonzalez, & Isaacs, 2016; Hollins & Reiss, 2016).

Sahlberg (2012, 2015a, 2016) disparages the advocacy of school choice, standardisation, and test-based accountability by dubbing it the Global Education Reform Movement, or GERM—a term that has gained some traction among GERM opponents (e.g. Robertson, 2015).[2]

However, Sahlberg himself is a central figure in the second set of discussions composing the backdrop of this thesis. Sahlberg's book, *Finnish Lessons: What Can the World Learn from Educational Change in Finland?* (2012), along with titles such as *Surpassing Shanghai: An Agenda for American Education Built on the World's Leading Systems* (Tucker, 2011), are part of the popular discourse on best practices in high-performing education systems. Other examples include two widely publicised reports by McKinsey & Company, which used cross-country student assessment data to identify, respectively, the 'best performing' and 'most improved' school systems (Barber & Mourshed, 2007; Mourshed, Chijioke, & Barber, 2010), and distilled the policies used in these countries into sets of 'best practices … [that] work irrespective of the culture in which they are applied … [and] demonstrate that substantial improvement in outcomes is possible in a short period of time' (Barber & Mourshed, 2007, p. 5).

This discourse is driven by league tables from cross-country student assessments such as the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS). In many participating countries, these assessments trigger competitive anxiety about how well their education system stacks up to its peers, as well as feverish curiosity about which policies can be adopted from countries at the top of the tables (Breakspear, 2012; Grek, 2009). Claiming that a policy initiative mimics that of high-performing systems can be a persuasive political strategy (Steiner-Khamsi, 2014). For example, policymakers in England have justified autonomous academies and high-stakes accountability on the basis that these policies mirror high-performing East Asian education systems—despite the fact that these English policies look very different from the East Asian models that they purportedly mimic (You, 2017).

[2] It is worth noting that both Finland's and Singapore's approaches to teacher accountability diverge from the stereotypical form of high-stakes, test-based accountability associated with 'neoliberal' trends in education policy. Singapore may seem a prototype of neoliberalism, with emphases on meritocracy and on the economic productivity of education—but some elements of Singaporean public education, such as restricted freedom of expression and extensive central planning, deviate from classical free-market principles (Chua, 2018; Weninger, 2016). In contrast, education in Finland may be portrayed as antithetical to neoliberalism, with its refusal to rank schools or impose consequences based on test scores (Conway & Murphy, 2013; Hurley, 2013; Simola, 2014). However, the OECD has strong influence in Finland (Simola, 2014), and policies adopted in the 1990s give parents free choice among public schools in the country, and put principals in charge of school finances (Rinne, Kivirauma, & Simola, 2002; Webb et al., 2004).

Against this backdrop, I look at teacher accountability practices in high-performing Finland and Singapore, and I also explore teacher accountability using multi-country PISA and TIMSS datasets. My outlook is informed by prior studies emphasising the social and relational nature of the teaching profession (e.g. Ingersoll, 2003), and by the growing body of policy analysis models that emphasise compatibility with the implementation context (see Section 2.2 below).[3] Beyond academic research, my interest in how sociocultural context affects the implementation of teacher accountability policy emerged during my two years as a secondary school teacher in Malaysia. In the school where I taught, local cultural priorities meant that the responsibility to protect the school from losing face was more important than the responsibility to be honest when filling in paperwork for faceless bureaucrats—such that many administrative instruments for teacher accountability neither collected accurate information nor influenced classroom practice. While sociocultural priorities inhibited the efficacy of some accountability instruments in this low-performing Malaysian school, such priorities can also enable effective teacher accountability processes, as I will show subsequently.

1.2 Contributions to research and policy

In this thesis, I explore the extent to which the efficacy of teacher accountability instruments depends on sociocultural context. This exploration uses multilevel modelling of international large-scale assessments (ILSAs) such as PISA, alongside international surveys of cultural values. To my knowledge, based on a systematic literature search that I conducted (see Section 2.3, as well as footnote 13 in Section 3.3), this is the first cross-country ILSA analysis of sociocultural factors influencing teacher accountability.

Additionally, I interview 12 teachers in Finland and 12 in Singapore, resulting in a detailed comparison of teacher accountability and sociocultural context. Although there have been other fieldwork-based studies arguing for the centrality of sociocultural context to Finland's education policy choices (e.g. Andere, 2014; Chung, 2009), I believe this thesis to be the first comparative analysis of educational accountability in these two countries, as noted above. This matters because Finland's and Singapore's education systems are often cited as references for policymaking (as observed by Chung, 2009; Dobbins & Martens, 2012; Takayama, Waldow, &

[3] I realised retrospectively that this study also echoes some of the arguments of neo-institutionalism (H.-D. Meyer & Rowan, 2006; J. W. Meyer & Rowan, 1977), e.g. that institutions are socially constructed and that (cultural) institutions influence behaviour. However, I did not draw specifically on this school of thought while working on this thesis.

Sung, 2013 on Finland; and Clapham & Vickers, 2018; de Roock & Espeña, 2018 on Singapore). Notwithstanding their popularity as policy references, Finland and Singapore pose some constraints to comparative education researchers, which may heighten the risk of superficial and inappropriate policy borrowing. In Finland's case, this is partly a language matter, since national policy documents, classroom interactions, and educational research largely use Finnish and, to a lesser extent, Swedish (Y. Li & Dervin, 2018). In Singapore's case, this is partly because Singapore has a single institute for all teacher training (Low & Tan, 2017; Teh, Hogan, & Dimmock, 2013), which is closely aligned to the authoritarian state via the Ministry of Education, and which is the academic home of all university-based education researchers. In both cases, another constraint is that foreign visitors' observations are often restricted to whatever their host institutions decide to show them (as observed by Y. Li & Dervin, 2018 on Finland; see Reimers & O'Donnell, 2016 for an example of a tightly planned visit for U.S. educators in Singapore). Although I conducted all interviews in English, thus requiring Finnish participants to use a second language, the fact that I recruited participants via personal, educational, and local contacts rather than through government sampling frames facilitated candid discussions of teachers' firsthand experiences.

Another strength of this research project is that the combination of large-scale statistics and granular teacher interviews allows for insight into whether relationships exist between teacher accountability and sociocultural context at an aggregate, cross-country level, and also into how these relationships manifest in particular national contexts. While the cross-country statistics can give a black-box indication of whether it is likely to be worthwhile for policymakers to investigate sociocultural context when designing accountability policy, the interview analysis can suggest specific areas and processes that may warrant investigation.

Furthermore, I develop and attempt to validate a conceptual framework for mapping the intended effects of teacher accountability instruments, and for considering potential pitfalls along these intended pathways. Despite increasing emphasis in the grey literature on the importance of implementation contexts in education policy and accountability (e.g. Schleicher, 2018; UNESCO, 2017), there is little clear guidance on how education policymakers should take sociocultural patterns into account. My framework is preliminary and only addresses a limited subset of education accountability. Nonetheless, based on formative feedback from policy practitioners in the Teach For All alumni community of practice in education policy, I have been encouraged to hope that this framework has the potential to prompt constructive reflections on teacher accountability policymaking among practitioners.

1.3 Outline of the chapters

In the next chapter, I review the literature on teacher accountability and sociocultural context. Specifically, I discuss the scope of relevant prior studies in educational research as well as in adjacent fields such as public policy and psychology. This includes a systematic search for literature about teacher accountability and sociocultural context. Based on this literature, I develop a conceptual framework for this study and propose three research questions emerging from the framework.

In Chapter 3, I lay out the methodology of this study. I discuss the theoretical perspective that I adopt—a realist ontology alongside a constructivist epistemology—as well as its implications for my empirical research. I then give an overview of how my data sources fit together with different aspects of the conceptual framework and research questions. Next, I describe and justify my approach to data collection and analysis for the secondary statistical analysis and the field interviews in Finland and Singapore, before discussing the ethical implications and limitations of this research project.

Chapters 4, 5, and 6 present the results for each of the three research questions. In Chapter 4, I investigate the extent to which the relationship between teacher accountability instruments and student outcomes varies with sociocultural context, using secondary statistical data. This includes a discussion of how these results link to the literature reviewed in Chapter 2, and a consideration of alternative explanations for the empirical results. A similar discussion of links to the literature and of alternative explanations is also included in each of the other two results chapters. In Chapter 5, I investigate the extent to which teacher accountability instruments influence teacher motivation—and, by so doing, affect student outcomes—using the statistical datasets as well as the field interviews with teachers in Finland and Singapore. In Chapter 6, I investigate the extent to which the influence of teacher accountability instruments on teacher motivation depends on sociocultural context, again using both the statistical datasets and the teacher interviews

Finally, Chapter 7 concludes the thesis, with a brief overall discussion of the extent to which the empirical evidence supports my conceptual framework. I also lay out some caveats about my argument, suggest directions for future research, and offer final reflections.

Chapter 2: Literature review and conceptual framework

In this chapter, I discuss prior research on the relationships between teacher accountability instruments, sociocultural context, and teacher motivation. In Section 2.1, I discuss different conceptions of teacher accountability and explain my definition of teacher accountability instruments. Next, I describe the varied effects of such instruments and illustrate how these variations can be due to differences in policy design and implementation contexts. In Section 2.2, I give a brief overview of strands of policy analysis and education research that pay attention to the influence of context generally and of sociocultural context particularly. I then define sociocultural context as discussed in this study. In Section 2.3, I describe a systematic literature search that I conducted on teacher accountability and sociocultural context, and outline evidence from this search suggesting that sociocultural context influences a range of processes within teacher accountability. In Section 2.4, I begin by defining motivation and outlining some theories of motivation from psychology and management studies. I then discuss the relationship between accountability and motivation; and illustrate how sociocultural context might influence teacher motivation and, relatedly, teachers' responses to accountability instruments. Finally, I lay out my conceptual framework in Section 2.5 and propose three research questions in Section 2.6.

2.1 Teacher accountability policy: multiple definitions, mixed results

Defining teacher accountability instruments

Conceptions of accountability, whether for teachers or for other actors, vary across time and space (Broadfoot & Osborn, 1993; Hopmann, 2008; Koppell, 2005; UNESCO, 2017). As Mulgan (2000) outlines, 'accountability' is used to denote not only answerability to an external authority, but also an internal sense of responsibility, an attribute of governments that respond to the wishes of the electorate, or any institutional constraints on the behaviour of actors in public organisations, among other things.

In this project, I conceptualise accountability as a principal-agent relationship (Bovens, Schillemans, & Goodin, 2014; Gailmard, 2014; Pritchett, 2015; see also World Bank, 2003). Within principal-agent accountability, one prominent strand of work entails formal rational-choice modelling of electoral politics, where politicians are accountable to voters (e.g. Besley, 2007, Ch. 3), or of bureaucracy, where bureaucrats are accountable to political authorities (e.g.

Dixit, 2002). However, I draw on principal-agent conceptualisations of accountability in the broader sense that analysing an accountability relationship entails identifying *who* (agent) is accountable to *whom* (principal), and *what* is being accounted for (numerically or otherwise) in this relationship. This accords with 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 judgement, and the actor may face consequences. (p. 450)

This conception encompasses a wide range of accountability relationships; regardless of whether the power distribution between principals and agents is hierarchical or horizontal (Bovens, 2007); whether agents' responsibility is specified using rigid metrics or contingent expectations (Honig & Pritchett, 2019); and whether the relationship prioritises ex ante selection of agents or ex post consequences for agents' fulfilment of responsibilities (Mansbridge, 2014); among other variations.

Notwithstanding the breadth of this relational conceptualisation of accountability, it is distinct from the concepts of governance and quality assurance. Some aspects of these concepts overlap with principal-agent accountability, as posited in theoretical frameworks that categorise accountability as a subcomponent within governance (van Kersbergen & van Waarden, 2004) or within quality assurance (Middlehurst & Woodhouse, 1995). However, governance and quality assurance do not encapsulate all forms of accountability. Like accountability, governance is a diffuse concept with varied meanings (Levi-Faur, 2012; Peters, 2012; van Kersbergen & van Waarden, 2004; Weiss, 2010). Also, like accountability, governance is often conceptualised in terms of its subtypes, such as good governance and global governance (Weiss, 2010) or hierarchical governance and horizontal governance (Levy, 2018; see also Greany & Higham, 2018, on hierarchies, markets, and networks in school governance). Nonetheless, governance is often situated in analyses of the changing roles of state and nonstate actors following the incorporation of markets and networks into public service delivery, amid globalisation and disenchantment with hierarchical bureaucracy (Bevir, 2007; Hudson, 2007; van Kersbergen & van Waarden, 2004; Zumbansen, 2012). Some conceptualisations of governance relate closely to principal-agent accountability, but others do not. Despite some overlaps, governance places primacy on processes and structures of governing (Fukuyama, 2013; Levi-Faur, 2012), whereas accountability is defined by the principal-agent relationship, at least for the purposes of this thesis. Moreover, some conceptualisations of governance prioritise the actors that do the governing (e.g. Fukuyama, 2013, on governance as government capacity), while my conceptualisation of accountability places equal weight on principal and agent.

In turn, quality assurance relates to processes for managing the quality of operations and outputs, whether these processes are internally determined or driven by external monitoring bodies (J. Williams, 2016). Quality assurance is conceptualised narrowly in some fields (e.g. regulatory compliance in pharmaceutical laboratories; Gawadi, 1996; Visschedijk, Henrdiks, & Nuyts, 2005), but takes a broader range of conceptualisations in education (e.g. Harvey & Green, 1993; Simola et al., 2009), where it is a particularly prominent concept in higher education research (as manifest in the journal *Quality in Higher Education*, currently in its 26th volume), and to a lesser extent in discussions of 'quality assurance and evaluation' in the governance of some European education systems (Grek et al., 2009; Simola et al, 2017). Still, quality assurance implies formalised standards ('quality') defined by the principal for measuring the agent's performance, whereas an accountability relationship can be based on informal, tacit, or dynamic expectations. More importantly, governance and quality assurance may overlap with managerial, top-down accountability relationships—or, in some subtypes of governance, with accountability relationships between equal-status partner organisations (e.g. Greany & Higham, 2018)—but these concepts typically exclude relationships such as informal accountability to local communities or ad hoc horizontal accountability between individual colleagues.

If accountability is defined as a principal-agent relationship, then it entails not only processes of account-giving and judgement emphasised in Bovens' (2007) definition of accountability, but also the processes of selecting the agent (Mansbridge, 2009, 2014) and of setting the standards for the agent's fulfilment of responsibilities. Selection of agents is discussed under 'Pathways from teacher accountability instruments to student outcomes' in Section 2.5. For standard-setting, note the following four characteristics. Firstly, standard-setting concerns not only the results or outputs of the agent's actions, but also standards for the conditions under which they perform those actions. If a teacher is not given a standard of instructional resources, pedagogical training, and decision-making discretion adequate for fulfilling any stipulated standards for student learning, then it would be unfair to hold them accountable for those outcome standards, because the terms of the accountability relationship render such outcomes beyond the teacher's control (Wagner, 1989). Secondly, these standards need not be formally codified, nor rigidly standardised. Tacit social expectations can carry at least as much weight as explicit legal guidelines. Thirdly, standard-setting can be particularly important when the agents under consideration are teachers, given that teachers are beholden to a range of stakeholders—students, communities, headteachers, education authorities—and that the teaching profession serves multiple and often ambiguous goals (Murnane & Cohen, 1986; Wagner, 1989; for empirical examples, see Atuhurra & Kaffenberger, 2019, and Kurniasih, Utari, & Akhmadi, 2018). This applies to public service jobs more generally (Lipsky, 2010; see also Hölmstrom & Milgrom, 1991). As Romzek and Dubnick's (1987) argue, '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). Finally, such processes of managing expectations and setting standards involve social perceptions of what matters most in education, and are often influenced by power dynamics between different stakeholders (Day & Klein, 1987; see also Rittel & Webber, 1973, on the challenge of formulating objectives in resolving wicked problems). For example, the shifting forms of teacher accountability in the United States have been shaped by shifting conceptions of educational equity (McDermott, 2011), and by problem-solving models from higher-status fields such as business and national defence (Mehta, 2013). More generally, the determination of what counts as a desirable outcome and how it should be counted is a fundamental and contestable part of teacher accountability.

To gain some analytic traction within the expansive domain of principal-agent accountability, in this thesis I focus on the instruments within teacher accountability relationships. Bringing together Bovens' (2007) emphasis on answerability and judgement in principal-agent relationships and Romzek and Dubnick's (1987) attention to managing expectations within such relationships, 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 states that accountability instruments 'aim to orient teacher practice', because teacher practice does not change mechanistically in response to accountability instruments. Rather, teachers have some agency over how they respond to such instruments, as discussed in Section 2.4. Note also that this definition includes instruments targeting not only teachers' individual actions but also their collective practice, since department- or school-level evaluation and incentives can also affect teachers' professional experiences and play a key role in managing teachers' performance (e.g. Ingersoll, Merrill, & May, 2016). It further includes teacher accountability instruments that may not be officially codified but nonetheless influence teachers' work within their respective accountability relationships, such as social networks carrying observations from Israeli Bedouin classrooms to the tribal sheikh (Mizel, 2009) and regular telephone calls from Norwegian parents to teachers (Czerniawski, 2011). Additionally, this definition includes instruments for which the stakeholders in question are teachers themselves, as in professional accountability (e.g. Fullan, Rincón-Gallardo, & Hargreaves, 2015; Romzek & Dubnick, 1987). Such professional accountability relationships can be conceptualised as principal-agent relationships in which individual teachers are accountable to the collective body of fellow teachers, who hold a legitimate stake in each individual's work because of how such work affects their collective reputations, morale, and working conditions. The three mechanisms by which accountability instruments can attempt to influence teacher practice—communicating information, setting standards, and allocating consequences—are discussed further in Section 2.5.

The inconsistent effects of teacher accountability instruments

Teacher accountability instruments have a patchy track record in improving student outcomes (Ganimian & Murnane, 2016; Kozlowski & Lauen, 2019; National Research Council, 2011). In some experimental and quasi-experimental studies, performance-related incentives for teachers helped to raise student outcomes (Duflo, Hanna, & Ryan, 2012; Glewwe, Ilias, & Kremer, 2010; Lavy, 2009; Muralidharan & Sundararaman, 2011). However, other such studies found no significant effects on student outcomes (Fryer, 2013; Glazerman & Seifullah, 2012; Springer et al., 2010, 2012).

Besides these direct effects on student outcomes, teacher accountability instruments can have numerous side effects (Thiel, Schweizer, & Bellmann, 2017; Zhao, 2018). Many tests with high stakes for students, teachers, or schools inadvertently embody Campbell's Law that:

> The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor. (Campbell, 1979, p. 85)

In particular, high-stakes tests have prompted strategic, and arguably detrimental, behaviour among teachers and school leaders, such as focusing narrowly on subjects and learning objectives that are prioritised in assessments (Altrichter & Kemethofer, 2015; Ng & Chan, 2008; Weninger, 2016), or disproportionately allocating resources to test preparation (Thomas, 2013). More pernicious actions include diverting educational resources toward students hovering below performance thresholds (Booher-Jennings, 2005), reclassifying students as having special education needs so that they would be excluded from accountability calculations (ibid), and outright cheating (Vogell, 2011). Furthermore, test-based accountability can unhealthily heighten students' competitiveness, stress levels, and passivity in learning (Luna, 2015; Redden & Low, 2012; A. Walker, Qian, & Zhang, 2011). It can also raise teacher stress and compromise school climate (von der Embse, Pendergast, Segool, Saeki, & Ryan, 2016).

One reason why teacher accountability instruments can have divergent effects is that apparently similar accountability policies can differ in pivotal ways (Verger & Parcerisa, 2017a). In short, policy design matters (Pritchett, 2017). Analyses of effective performance-based pay schemes for teachers posit that programme design—such as using multiple measures of teacher performance (Lavy, 2009) and awarding bonuses based on clear criteria that teachers recognise as fair (Murnane & Cohen, 1986)—plays a decisive role in minimising perverse incentives. Additionally, the presentation and framing of a policy can be influential. One experiment found that framing financial incentives as losses rather than gains can elicit greater increases in teacher productivity, even if the incentives are of equal magnitude and are based on identical performance criteria (Fryer & Levitt, 2010).

Another reason behind the differential efficacy of teacher accountability instruments is that implementation contexts differ. One such area of difference is the larger policy environment. Some studies suggest that accountability instruments may be most effective when other policy instruments give teachers and schools enough latitude, whether in administrative autonomy or resource levels, to meet accountability standards. In terms of autonomy, Woessmann (2005) found that central exit examinations are associated with higher scores in international assessments when the former are combined with certain forms of school decision-making autonomy. Ingersoll, Merrill, and May (2016) found that autonomy can mitigate a negative side effect of accountability instruments: among American schools that faced sanctions for falling below accountability standards, most saw higher teacher turnover than in schools that met the standards, but sanctioned schools that gave teachers significant classroom autonomy had turnover rates on par with their higher-performing counterparts. Ineffectual delegation can also compromise efficacy, as in a Pakistani province where the supervisors who inspected schools were not empowered to administer the teacher transfers, rewards, or punishments that they believed were necessary, while the authorities who were thus empowered rarely acted on supervisors' recommendations (Jaffer, 2010). In terms of resources, a performance-based pay intervention in Uganda only improved student attendance and achievement in schools that had textbooks (Gilligan, Karachiwalla, Kasirye, Lucas, & Neal, 2018). A Tanzanian experiment found that performance-based pay on its own had ambiguous effects on student achievement, whereas performance-based pay combined with additional school funding raised student performance (Mbiti et al., 2019). Similarly, accountability instruments can falter due to inadequate resources for implementing the instruments themselves, such as when school inspectorates lack the personnel or transportation to visit all the schools under their purview (De Grauwe & Lugaz, 2007; Ehren, Eddy-Spicer, Bangpan, & Reid, 2016; Schwartz, 2000).

Beyond the policy environment, other contextual features can also affect the efficacy of accountability instruments. For example, some forms of test-based accountability appear to have a more positive effect on student achievement in lower-performing countries than in higher-performing ones (Bergbauer, Hanushek, & Woessmann, 2018). At a more granular level, individuals and small groups can be key to increasing the benefits of teacher accountability while minimising potential harm. Such individuals and groups may be especially important for cultivating non-bureaucratic forms of accountability, such as professional accountability or community accountability, as R. Iyengar (2012) shows in two Indian villages and Skrla, Mckenzie, Scheurich, and Dickerson (2011) demonstrate in a Texas school district. Conversely, interest groups can also precipitate the failure of teacher accountability instruments, as with partisan capture for patronage in a scheme for community-managed schools in Honduras (Altschuler, 2013) and a merit pay programme in Mexico (Douglas, 2014). Moreover, as I show below, the implementation and outcomes of teacher accountability instruments can be influenced considerably by sociocultural factors.

2.2 Why sociocultural context matters

The role of sociocultural context in policy implementation

In the previous section, I showed how some aspects of the implementation context can affect teacher accountability policy. In this section, I focus on a particular aspect of context—the sociocultural—while considering not only education policy, but also public policy more broadly.

It is worth noting, firstly, that a growing number of policy analysis models emphasise the theories of change that connect policy instruments to their desired outcomes, and the contextual features that influence these change processes (e.g. Andrews, Pritchett, & Woolcock, 2017; and Bates & Glennerster, 2017 in public policy generally; and Cambridge Assessment, 2018; McDonnell & Elmore, 1987; and Monaghan & King, 2018 in education policy). Pawson and

Tilley's (1997) landmark work on realist policy evaluation included not only the mantra 'what works for whom in what circumstances' (p. 210 and elsewhere), but also the formulation 'mechanism + context = outcome' (p. xv and elsewhere), which emphasises that policy outcomes depend not only on appropriate change processes (mechanisms), but also contextual suitability. Similarly, Cartwright and Hardie (2012) argue that policies can only play their intended causal roles if the necessary support factors are in place. In a framework pitched at policymakers rather than scholars, M.J. Williams (2017) maps the inputs, activities, and outputs of a policy programme onto its intended intermediate outcomes and final outcomes; thus helping policymakers to consider whether the contextual assumptions at each stage hold true in the actual implementation context. One thread running through all of these models is the argument that a policy will only work if it is compatible with its context.

Alongside this growing attention to theories of change and implementation contexts, there has been increasing attention to the role of culture in human behaviour change. In economics, Alesina and Giuliano (2015) recently emphasised the importance of studying bidirectional causal relationships between culture and institutions, rather than seeing culture solely as a cause or solely as an effect. Woolcock (2018) likewise observes that development studies has shifted away from earlier discussions about 'backward' cultures as determinants of poverty, and from modernisation theories which assume that economic, political, and cultural development occur in lockstep, toward regarding culture as a set of context-specific tools for engaging the world. Similarly, Collier (2017) argues that incorporating culture into formal models as an endogenous and influential element allows for better explanations of socioeconomic and political phenomena. Beyond political economy and development, there is a long-established body of work in cross-cultural psychology examining differences in thought, values, and behaviour between cultural groups (Berry, 2011; Gelfand et al., 2011; Markus & Kitayama, 1991; Triandis, 2004). Adjacent to such psychological research, some strands of organisational studies investigate the interaction between culture and optimal organisation design, notably in Hofstede's (1980) influential survey programme (see Nardon & Steers, 2009 for a review of other cultural models in organisational research).

Additionally, there are established cross-cultural research programmes in some areas of educational research. Notable studies of cultural differences in teaching and learning include Tobin, Wu, and Davidson's (1989) *Preschool in Three Cultures* on Japan, China, and the United States; Stevenson and Stigler's (1992) *The Learning Gap,* comparing the same three countries (see also Stigler and Hiebert's, 1999, subsequent work with the TIMSS video study); Alexander's (2001) *Culture and Pedagogy*, examining primary education in England, France, India, Russia and the United States; and Li's (2012) *Cultural Foundations of Learning: East and West* on how the distinct learning philosophies of East Asia and the West manifest in different pedagogies and psychologies of learning. All of these studies emphasise the influence of cultural values, priorities, and practices on teaching and learning. Others have demonstrated how the implementation of imported curricula and pedagogies may be hampered or, at least, modified by local cultural pressures (Clapham & Vickers, 2018; Heng & Song, 2020; S. Kim, 2017).

However, as I show in the next section, there has been little systematic comparative work on the relationship between culture and teacher accountability. Before I examine the evidence for such cultural influence over teacher accountability policy, I first define sociocultural context as conceptualised in this thesis.

Defining sociocultural context

Culture, like accountability, is a diffuse and contested concept. To gain some analytic traction, I focus not on culture as a broad concept, but on sociocultural context. In defining sociocultural context, I draw 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 (see also Adams & Markus, 2001; Kitayama, 2002; Markus & Kitayama, 1991). 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, this definition is aligned not only with the realist ontological stance of this thesis (discussed in Section 3.1), but also with a strand of psychological research. Additionally, this definition coheres with the secondary sources on cultural data that I analyse. It emphasises that sociocultural context is a characteristic of groups, not of individuals; but also that it need not be distributed uniformly throughout a group (as in the seminal work of Wallace, 1970, in anthropology). Both of these properties of culture are affirmed by key theorists associated with the two sociocultural datasets that I analyse: Ronald Inglehart and Christian Welzel, the most prominent analysts of the World Values Survey, and Geert Hofstede, who developed the Values Survey Module. Theoretically, both Inglehart and Welzel (2005) as well as Hofstede (2001) draw a distinction between the collective frames of meaning that influence the social environment, and the unique beliefs and values that shape individual actions. Methodologically, their quantitative sociocultural constructs derive from the proportion of people in each country who subscribe to particular views—thus, the constructs are necessarily collective, but are based on degrees of difference rather than presumed uniformity within the group.

As a system of dominant patterns, sociocultural context has emergent properties that can differ from the individual tendencies within it. For example, a teacher may believe that unequal distributions of power are inherently unjust—but if she lives in a context where such inequalities are largely accepted, then she is likely to conform to hierarchical norms in school management, whatever her personal inclinations.

2.3 Teacher accountability and sociocultural context: a systematic literature search

I conducted a systematic literature search in April 2017, using the Institute of Educational Sciences' ERIC and Elsevier's Scopus databases. (I chose Scopus over Web of Science to maximise geographic coverage, as the former covers more non-Western journals [Elsevier, 2016; Vieira & Gomes, 2009].) In each database, I searched for records containing the terms (1) 'accountability' and (2) 'teacher' or 'school', as well as (3) at least one of following terms:

- 'culture', 'social', 'societal', or 'sociocultural', to explore existing research about the interaction of sociocultural context and teacher accountability;[4]
- 'cross-country', 'across countries', 'cross national', or 'national differences', to investigate differences across education systems;
- 'production function', 'input-output', or 'productivity', to identify other studies using the family of statistical models that I intend to use; or
- 'PISA', 'TIMSS', or 'TALIS', to ascertain the scope of prior teacher accountability studies that used large-scale international survey data.

[4] I used a wildcard indicator to include different word endings, e.g. 'cultural' as well as 'culture'. This set of search terms initially yielded over 1,000 results in each database. To make the results list more manageable, I limited the search to studies published from 2008 to 2017 (inclusive) and only looked for the search terms in abstracts (rather than abstracts, titles, and keywords). These restrictions were not applied to the other sets of search terms.

All searches were restricted to peer-reviewed studies.

After eliminating duplicates and non-English-language publications, I was left with 1,740 records. I skimmed the abstracts from these records over the course of two weeks, identifying studies that either compared teacher accountability instruments across countries; or discussed the interaction of teacher accountability instruments and sociocultural context, whether at the national, regional, or school level. In order to maintain a manageable scope, I excluded studies on early childhood education and higher education. In total, 89 journal articles, 2 book chapters, and 4 books fit these inclusion criteria. In the interests of conciseness, the discussion below focuses on studies discussing the relationship between teacher accountability and sociocultural context. Other studies from this systematic search have also informed this research project, and many are cited elsewhere in this chapter.

Based on the systematic literature search, there has been little empirical cross-country research focusing specifically on the relationship between sociocultural context and teacher accountability policy (although there have been numerous context-sensitive single-country case studies, such as Easley & Tulowitzki, 2016, which synthesises 12 geographically diverse single-country studies). Nevertheless, I found studies indicating, across a range of countries, that sociocultural context influences the design of teacher accountability policy, the range of stakeholders to whom teachers are accountable, and the implementation of teacher accountability instruments. I discuss each in turn.

First, there is some cross-country evidence that sociocultural context can shape the formulation of teacher accountability policy. Hopmann (2008) attributes the different emphases of the accountability policies enacted in the United States, Nordic countries, Germany, and Austria to 'deeply engrained ways of understanding the relation between the public and its institutions' (p. 425). For example, a deep-seated trust in teacher quality led Nordic countries to accommodate accountability pressure by introducing decentralisation to better support teachers' work; while a widespread perception in the United States that education was in crisis precipitated high-stakes tests based on national standards (ibid). Similarly, others have linked contemporary accountability policies and schooling structures in some European countries to longstanding traditions in education and governance (see Mattei, 2012, on England, Germany, France, and Italy; Møller & Skedsmo, 2013, on Norway; Osborn, 2006, on Denmark, France, and England). Contemporary sociopolitical forces can also exert influence. For example, post-World War II decentralisation policies were influenced by allegiances to liberalism and socialism: in South Korea, decentralisation discourse emphasised fostering liberal democracy through schooling, while decentralisation in China and Tanzania was linked to socialist rural cooperatives (Edwards & DeMatthews, 2014).

Second, some studies suggest that sociocultural context can influence the configuration of stakeholders to whom teachers are accountable, as well as the tenor of these accountability relationships. For example, Tanzanian teachers feel strongly accountable to parents, but this accountability is rooted in a sense of shared moral responsibility for children's futures, which implies that parents hold reciprocal responsibility for supporting their children's education—and which contrasts with the neoliberal view of parents as the customers of the school (Barrett, 2005). In a fascinating insider study, Mizel (2009) discusses how Israeli Bedouin teachers face conflicting accountability demands from the Ministry—which expects them to be apolitical employees—and from the tribal sheikh—who expects them to participate in the political leadership of the community. Other studies have found between-country differences in the stakeholders to whom teachers feel accountable, as well as the processes and goals emphasised within these relationships (Czerniawski, 2011, on Norway, Germany, and England; Farrand, 1988, on Mexico, England, and France; Müller & Hernández, 2010, on Finland, England, and Ireland).

Third, sociocultural context can affect teachers' beliefs about priorities and principles, which in turn affects the implementation and efficacy of teacher accountability policy. For example, a major teacher quality reform in Indonesia included plans to train teachers as peer evaluators, but teachers in this highly hierarchical society questioned colleagues' authority to evaluate their work, believing that such authority should only be held by supervisors or head teachers (Broekman, 2016). Similarly, some teachers in India challenged community accountability structures because they expect to be treated as high-status professionals who should be beyond the purview of low-status villagers (Narwana, 2015). Mizel's (2009) study of Israeli Bedouin schooling found numerous sociocultural challenges to formal teacher accountability. One issue was the prioritisation of cultural preservation over academic achievement. Despite the Ministry's expectation that the schools would adhere to the official curriculum, curricular standards and documentation were not prioritised in practice: principals were required to report to the tribal sheikh if students transgressed behavioural norms, but neither principals nor sheikh emphasised pedagogical planning and reporting. Thus, sociocultural beliefs—whether about education, hierarchy, or social status—can threaten the efficacy of planned accountability instruments.

In addition to offering examples of how sociocultural context can affect teacher accountability, articles from the systematic literature search also indicate two particular aspects of sociocultural context that may play a role in accountability: social capital and power distance. Social capital—a community-based complement to other forms of productivity-enhancing capital—has been defined in numerous ways (e.g. Bourdieu, 1986; Coleman, 1988; Putnam, 1995). These definitions share an emphasis on aspects of social relationships that facilitate the actions of those who possess such capital, whether at the individual level (e.g. facilitating access to opportunities) or the group level (e.g. facilitating cooperation). From the systematic search, two studies examined the relationship between social capital and accountability, through interviews in 14 villages in a central Indian district (R. Iyengar, 2012) and regressions of student outcomes against voter turnout for Missouri school board elections (Webber, 2010). Both conclude that there is a link between social capital and accountability-related improvement, though this relationship may only apply to certain student outcomes and under certain enabling conditions. For example, in 12 out of the 14 villages in R. Iyengar's (2012) study, informal social networks did not lead to community participation in school accountability. However, such community participation did exist in two of the villages, where village elders and a local NGO, respectively, had mobilised informal networks for accountability. Beyond these examples from the systematic search, social trust—a key component of social capital—may strengthen the benefits of accountability instruments when it strengthens teachers' and stakeholders' commitment to shared expectations (Bryk & Schneider, 2002; Cerna, 2014). Conversely, if teachers are used to being trusted as professionals, introducing new accountability instruments may imply suspicion, which may harm teacher motivation and student outcomes (O'Neill, 2002; Sahlberg, 2010).

While social capital is a measure of interpersonal interactions, power distance is a measure of hierarchy. Drawing on Mulder's (1977) work on power, Hofstede (2001) defines the power distance between a boss (B) and a subordinate (S) as 'the difference between the extent to which B can determine the behaviour of S and the extent to which S can determine the behaviour of B' (p. 83). On a societal level, power distance thus measures acceptance of hierarchical distributions of power. Although no articles from the systematic literature search examined power distance explicitly, Broekman's (2016) and Narwana's (2015) studies give instances in which strong social hierarchies hampered teacher accountability, because teachers believed that, respectively, fellow teachers in Indonesia and low-status villagers in India could not legitimately appraise them, as noted above. Beyond the systematic search, H.-D. Meyer and Schiller (2013) argue that high-performing countries in PISA have a high GDP as well as either high individualism and low power distance, or low individualism and high power distance. More pertinently, some studies in psychology and management argue that different levels of power distance often lead to different modes of accountability (Gelfand, Lim, & Raver, 2004; Velayutham & Perera, 2004). On one hand, strong social hierarchies may hamper accountability instruments if the instruments require teachers to be appraised by stakeholders of equal or lower social status, as noted above (Broekman, 2016; Narwana, 2015). On the other, a greater acceptance of hierarchy can enhance the efficacy of accountability instruments because it promotes compliance (Jaques, 1990).

Another sociocultural factor that may influence teacher accountability processes is uncertainty avoidance, which was also constructed in Hofstede's influential research programme. He derived his measure of uncertainty avoidance from questionnaire items about rule-orientedness, preferences for job stability, and stress levels (Hofstede, 2001, p. 150). Although uncertainty avoidance was not mentioned in the studies from the systematic literature search, Velayutham and Perera (2004) include uncertainty avoidance among the societal values that affect the emotional states associated with accountability. Specifically, higher levels of uncertainty avoidance may boost the efficacy of rule-based forms of teacher accountability, since stronger preferences for stability may encourage conformity to accountability standards (see also Hood, 2011 on policy strategies for blame avoidance).

Overall, these studies offer numerous examples of the interplay between teacher accountability and sociocultural context. In the next section, I propose a pathway through which sociocultural context can influence the efficacy of policy instruments for teacher accountability.

2.4 Linking teacher motivation, sociocultural context, and accountability instruments

As discussed above, analyses of public policies should consider the processes of change that connect policy programmes to their intended outcomes. In this thesis, I focus on the ways in which teacher accountability instruments influence teachers and their work. I frame this analysis using the concept of teacher motivation, for reasons I explain below.

Defining and explaining teacher motivation

Whether implicitly or explicitly, accountability instruments targeting teacher practice assume that student outcomes can improve when teachers work harder, i.e. with raised motivation, and/or work differently, i.e. with motivation reoriented toward other activities or goals (c.f. Kozlowski & Lauen, 2019; Wagner, 1989). (Of course, some accountability instruments collect information on teacher practice for the sake of informing stakeholder decisions rather than influencing teacher practice directly. I discuss this pathway further in Section 2.5.) Hence, in this project, I focus on *teacher motivation* as the pivot of accountability instruments that successfully change teacher practice. For conceptual clarity, I adopt Schunk, Pintrich, and Meece's (2010) definition of motivation as *'the process whereby goal-directed activity is instigated and sustained'* (p.4).

However, as with accountability and with culture, motivation is a complex and contested concept. Psychological research has developed multiple empirically validated theories for explaining motivational processes (Deci, 1992). Current theories agree that motivation is a cognitive phenomenon, involving subjective mental processes that exert causal influence on action (Schunk et al., 2010). Nonetheless, these theories diverge considerably. Different theories emphasise different motivational factors and have different implications for the design of teacher accountability instruments. Additionally, different fields of inquiry tend to emphasise different theories or variants of theories. I discuss a few such theories below.

In education, one prominent set of theories focuses on the goals to which motivation is directed. This includes earlier needs-satisfaction theories such as Murray's (1938) taxonomy of primary and secondary needs and Maslow's (1954) hierarchy of higher and lower needs; as well as more recent theories such as work by Dweck and her collaborators (Dweck & Leggett, 1988; Elliott & Dweck, 1988) that connects learning goals and performance goals to distinct theories of intelligence and patterns of behaviour. These goal-related theories suggest that teachers' motivational responses to accountability instruments may depend on factors such as whether they view their students' (and their own) intelligence as fixed or malleable, and the degree to which the instruments support the satisfaction of their particular needs.

Another set of theories focuses on different types of motivation. The most prominent articulation here is Ryan and Deci's (2000a, 2000b, 2000c) self-determination theory, which distinguishes between intrinsic motivation (doing an activity for its own sake) and extrinsic motivation (doing an activity for the sake of some external reward, constraint, or other compulsion). One of the propositions of self-determination theory is that intrinsic motivation is sustained when the actor feels autonomous, competent, and meaningfully related to others. Drawing on this theory, Jang (2019) found that the content of teacher motivation influenced classroom practices, as reported by both the teachers and their students. Specifically, when teachers' self-reported instructional goals were intrinsic (i.e. students' personal and relational growth), their classroom styles supported pupil autonomy, whereas teachers whose instructional goals were extrinsic (i.e. students' test scores or the teacher's professional success) adopted more controlling styles. More relevant to teacher accountability is Deci and Ryan's (2000) proposition that external rewards—such as performance-based pay for teachers—may reduce intrinsic motivation if they compromise actors' sense of autonomy (see also related arguments in other fields, e.g. Deming, 1993, in management; Stout, 2010, on law and public policy).

Besides educational psychology, another area that is relevant to this project is research into management, organisational behaviour, and work motivation. Theories of motivation that are prominent in this field include Herzberg's motivation-hygiene theory (Herzberg, 1966, 1968; Herzberg, Mausner, & Snyderman, 1959). Herzberg's theory overlaps with Murray's (1938) and Maslow's (1954) needs-focused theories by distinguishing between basic needs stemming from biological necessity and pain avoidance, and the need for achievement and growth. In the workplace, the former are supported by hygiene factors (e.g. adequate compensation and job security), which can reduce job dissatisfaction but do not positively affect satisfaction; whereas the latter are supported by motivator factors (e.g. suitably challenging tasks and professional advancement), which can raise job satisfaction. Thus, Herzberg's theory would predict that performance-based bonuses—which are often regarded as an incentive for better teacher performance—could decrease teachers' dissatisfaction with their jobs if their compensation was previously inadequate. However, such bonuses would be less likely to have any impact on teachers' job satisfaction or their motivation to improve classroom practice.

In contrast, Vroom's (1964) expectancy theory—another prominent theory among management scholars and practitioners—would predict that salary bonuses could, under the right circumstances, boost teacher motivation. Vroom's theory centres on three motivational factors: expectancy, the belief that effort will lead to successful performance; instrumentality, the belief that successful performance will lead to desired outcomes; and valence, or how much value the actor expects to gain from the outcome. Thus, if a teacher believes that working harder will raise their students' test scores, and that raising these test scores will lead to a financial reward, and if they eagerly desire such financial gains, then establishing performance-based bonuses would raise this teacher's motivation.

Hence, conceptually, there are both overlaps and divergences between different theories of motivation. There are also overlaps and divergences between the research programmes studying motivation. To illustrate, Vroom's theory is cited frequently in management (Pinder, 1992), but rarely in education. However, an analogous framework in education psychology is expectancy-value theory (Wigfield & Eccles, 2000). Like Vroom's framework, expectancy-value theory draws on Lewin's (1935) and Atkinson's (1957) earlier work on expectancy and valence. Unlike Vroom's work, expectancy-value theory does not emphasise instrumentality, and it is often used as an analytic framework for teaching and learning processes (Schunk et al., 2010; see also Robertson-Kraft, 2014 for an example using expectancy-value theory to analyse a teacher evaluation system). However, Bandura's (1977) social cognitive theory, which is also popular in educational research, draws on different theoretical traditions but includes concepts that are similar to both instrumentality and expectancy. In Bandura's (1977) framework, there is a distinction between outcome expectancy, i.e. 'a person's estimate that a given behavior will lead to certain outcomes' (p. 193), which is analogous to Vroom's instrumentality; and efficacy expectations or self-efficacy, i.e. 'the conviction that one can successfully execute the behavior required to produce the outcomes' (p. 193), which is analogous to Vroom's expectancy.

The range of these theories indicates that motivational processes—and the analysis thereof—are complex. This complexity may help to account for the variable effects of teacher accountability instruments, since influencing teacher motivation may not be a straightforward process, as indicated by the diverse predictions of these theories of motivation. In this thesis, I do not subscribe to any single motivational theory. This is partly because I am not a psychologist by training and hence am not equipped to weigh the empirical validity of different theories. More importantly, all theories of motivation—like any other abstractions—are necessarily incomplete, and different theories may have greater explanatory power in different empirical situations. Next, I discuss how and why teacher motivation is pivotal to teacher accountability.

The relationship between teacher motivation and teacher accountability instruments

There are theoretical grounds for expecting a close connection between motivation and accountability. For example, Pawson and Tilley (1997, Chapter 3) argue that real change in any social process—including classroom teaching—results from changes in individual actors' choices, reasoning, and effort. As McLaughlin (1987) observes,

> change ultimately is a problem of the smallest unit. At each point in the policy process, a policy is transformed as individuals interpret and respond to it. What actually is delivered or provided under the aegis of a policy depends finally on the individual at the end of the line. (p. 174)

This emphasis on individual agency also appears in Wagner's (1989) argument that the likelihood of an actor fulfilling an accountability-related obligation depends on how they interpret this obligation morally, and whether circumstances are conducive to compliance (e.g. whether the obligation is compatible with prior obligations). More recently, Andrews, Pritchett, and Woolcock (2017) have drawn a distinction between accounting (i.e. reporting standardised information about one's work for institutional accountability processes) and accounts (i.e. 'the justificatory narrative I tell myself which reconciles my actions with my identity', p. 114). They argue not only that organisational success requires complementarity between accounting and accounts; but also that success in complex fields (such as education) depends less on the accounting and more on the strength and alignment of individual and organisational accounts (see also Abelmann & Elmore, 1999, on the distinctions between responsibility, expectations, and accountability). All of this suggests that changes in individual teachers' motivation—i.e. the cognitive processes that drive goal-directed behaviour (Schunk et al., 2010, p. 4)—underlie any changes in teacher practice that are prompted by accountability instruments.

Notwithstanding these theoretical foundations, the empirical basis for linking teacher accountability and teacher motivation is not particularly strong. This may be due in part to the difficulty of measuring teacher motivation: since motivation is a cognitive process, both observational and self-report measurements will necessarily be imperfect proxies (Schunk et al., 2010, Chapter 1). Still, some analyses have found that teacher accountability instruments can raise teacher effort, as measured by different proxies: Lavy (2009) used teachers' self-reports of additional instructional time after regular school hours as a proxy for their effort; Karachiwalla and Park (2017) instead used teachers' annual performance evaluation scores (which were awarded by district-level committees based on student test scores and teacher attitude, attendance, and classroom preparation); and Macartney, McMillan, and Petronijevic (2018) designated incentive-invariant teacher effects on test scores as 'teacher ability' and teacher effects that did vary with incentives as 'teacher effort'. Although these three studies did find that accountability instruments can raise teacher effort, some analyses did not find any such association between teacher accountability instruments and teacher motivation. In Yuan et al.

(2013), randomised studies of three different pay-for-performance programmes did not find effects on the number of hours teachers worked, nor on their instructional practices. It is important to note that these effort measures—hours of work, instructional practices, evaluation scores, and test results—are also influenced by non-motivational factors. At best, they indicate symptoms of teacher motivation, rather than motivation itself. Hence, the evidence for a direct connection between teacher motivation and accountability instruments is relatively weak.

However, there is stronger evidence that teachers' subjective priorities and choices—which are, arguably, closely tied to motivation—intervene in between accountability instruments and teachers' manifested actions, as posited by the theorists discussed above. From the systematic literature search, in addition to the examples discussed above of sociocultural context influencing teachers' responses to accountability instruments (i.e. Broekman, 2016; Mizel, 2009; Narwana, 2015), Müller and Hernández's (2010) mixed-methods 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 regarded as their chief responsibility. More broadly, scholars have suggested that certain education reforms in Mexico and in France failed to take root partly because these reforms were not aligned with teachers' conceptions of their responsibilities (Farrand, 1988; Osborn, 2006).

Beyond the systematic search, Bjork (2016) observed that Japanese teachers were submitting official reports that indicated compliance with new curricular directives—even though their actual classroom activities conformed more to established practice, local expectations, and parental demands than to national-level policy. Similarly, Ingram, Louis, and Schroeder (2004) found that one barrier to American teachers' uptake of data-driven accountability instruments is that teachers' conceptions of classroom efficacy are much broader than the test scores that these instruments emphasise. Both of these studies suggest that accountability instruments can be hampered by a mismatch between their goals and the goals that teachers are already motivated to pursue. More generally, a review of 29 empirical studies on accountability policy and teachers' workplace relations found that teachers' responses to accountability policies depend on their ethical and professional stances (Mausethagen, 2013). In another review article, Rowan (1990) proposes that where teaching is viewed as a routine task, schools should be organised following a control-oriented strategy; whereas a commitment-building strategy would be more effective where teaching is viewed as complex. Looking beyond education, Lerner and Tetlock's (1999) review of psychological research on how accountability affects social choices found that accountability instruments only prompt desirable increases in cognitive effort when numerous contingent factors coincide. One such factor is that the actor must view the source of the accountability as legitimate.

Taken together, these studies suggest that teachers' subjective perspectives matter greatly to the implementation and efficacy of accountability instruments (see also Spillane, 2009; Verger & Parcerisa, 2017a). As Abelmann and Elmore (1999) argue in their study of accountability in schools:

> The distinguishing characteristic of responsibility … is that it is personal and individual in nature and it stems from the values and beliefs of individuals. … organizational and external influences *may* play a part in teachers' perceptions of their role, but … individual values are *certainly* influential. (p. 3, emphasis original)

Hence, an accountability instrument will only influence a teacher as intended if the teacher regards the instrument as sufficiently persuasive in some way. In my systematic literature search and other explorations of the literature, I did not find any established bodies of research about the characteristics that are likely to render an accountability instrument persuasive to its targeted actors. Nonetheless, given the range of motivational theories discussed above as well as the complexity of teachers as individuals and of teaching as a profession, the bases of such persuasion are probably diverse—including moral principles, the desire for financial gain, and the fear of legal repercussions.

Some studies suggest that policy directives in general are more likely to receive compliance when they are regarded as legitimate. For example, Tyler (2006) argues that when legitimacy 'exists in the thinking of people within groups, organisations, or societies, it leads them to feel personally obligated to defer to those authorities, institutions, and social arrangements' (p. 376). Empirically, perceptions of legitimacy depend on the belief that an institution's procedures are fair and that the outcomes it generates are desirable, among other factors (Tyler, 2006; Wallner, 2008). This is borne out in some studies of teacher accountability. In a study of school inspection feedback in Flanders, teachers were more likely to accept feedback when they viewed the inspectors as professionally and ethically credible (Quintelier, Vanhoof, & De Maeyer, 2018). Similarly, teachers experiencing an accountability policy change in Virginia were more likely to view the new evaluation policy as legitimate when they believed that it had valid and reliable instruments, fair procedures, and worthwhile outcomes—and those teachers who regarded the new policy as legitimate were more likely to improve their instructional practices in response (J. Kim, Sun, & Youngs, 2019). Note the parallel here with Vroom's (1964) expectancy theory of motivation, described above: Vroom argues that instrumentality, i.e. belief that successful performance will yield the desired reward, is a key factor in motivation; just as the studies outlined in this paragraph suggest that belief in the procedural justice of an institution is key to regarding it as legitimate and, thus, to accepting its authority.

A caveat: even though it is clear that teachers' subjective perspectives play a critical role in accountability processes, these subjective perspectives cannot be reduced solely to motivation. Teachers' perspectives can be conceptualised using various constructs; including teacher identity (Barrett, 2005; Czerniawski, 2011), to give an example from the systematic literature search. Nonetheless, in this project I focus on teacher motivation, for two reasons. From a policy standpoint, as noted above, many teacher accountability instruments implicitly or explicitly aim to raise or reorient teacher motivation (e.g. Adkins, 2004). From a conceptual standpoint, motivation by definition relates to goal-directed behaviour—which can be mapped onto to the standards and expectations implicit in any accountability instrument. Framed in this way, the crux of the matter becomes how to design accountability instruments that influence teacher motivation in desirable ways.

How sociocultural context influences teacher motivation

Teachers' responses to accountability instruments can differ considerably—even for the same instruments in the same locality (e.g. J. Kim et al., 2019; Quintelier et al., 2018). This underscores the fact that, as argued above, educational change via teacher accountability instruments requires change at the level of individual teachers, who are heterogeneous. Heterogeneity notwithstanding, within a given setting there will be some broad patterns in the accountability instruments that most teachers regard as compelling. This is because teachers' perspectives and, in turn, the overall efficacy of teacher accountability instruments are shaped by context. As Pawson and Tilley (1997) observe, 'subjects will only act upon the resources and choices offered by a program if they are in conducive settings' (p. 216)—and most settings have some dominant characteristics that will influence the actions of most actors within them.

Importantly, it is not only teachers who are heterogeneous, but also the contexts they inhabit. Such contextual differences can include differences in how motivation-related processes operate. Contemporary research in both psychology (e.g. Markus & Kitayama, 1991; Triandis, 2004) and organisational studies (e.g. Kirkman, Lowe, & Gibson, 2006) posits that motivation and its related processes may be influenced by different factors depending on the sociocultural context. In a landmark study, S.S. Iyengar and Lepper (1999) showed that Anglo-American children were most motivated when they could choose between different tasks, whereas their Asian-American counterparts were most motivated when tasks were ostensibly chosen by their mothers or their classmates. This runs counter to one interpretation of self-determination theory, which posits that independent choice raises intrinsic motivation universally.[5]

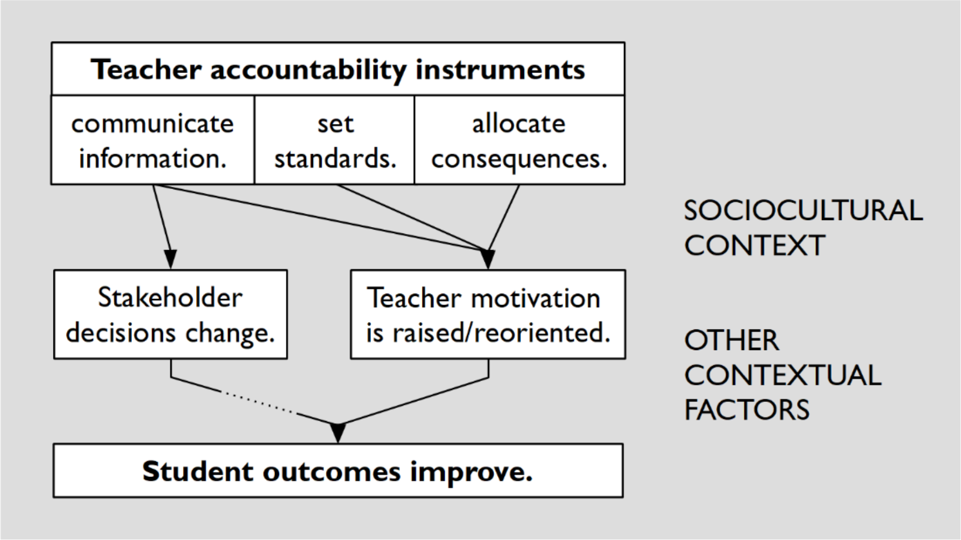
Looking at teachers specifically, empirical studies suggest that sociocultural context may influence teachers' professional priorities and goals (which would, in turn, shape teacher motivation). Besides the studies from the systematic literature search discussed above, Chistolini (2010) found cross-country differences in teachers' conceptions of 'the good teacher' in eight countries (Belgium, Cyprus, Italy, Libya, Poland, Slovakia, Turkey, and the United States). Broadfoot and Osborn (1993) likewise found systematic differences between English and French teachers' perceptions of their professional responsibilities, which were influenced by cultural and ideological assumptions. Furthermore, in a cross-cultural analysis of teacher motivation in Western and Chinese contexts, Ho and Hau (2014) concluded that some motivational processes function similarly across contexts (e.g. the association between high intrinsic motivation and positive teacher practices), whereas other aspects of motivation are culture-dependent (e.g. teachers' goals and values, as well as their beliefs about what good practice looks like). Similarly, Klassen, Usher, and Bong (2010) found that the associations between teacher job satisfaction, job stress, and collectivism (a cultural construct) differed between samples of teachers surveyed in the United States, Canada, and South Korea. In another study, Klassen et al. (2018) found that teachers and teacher educators in England, Finland, Malawi, and Oman differed in the non-cognitive attributes that they viewed as most important for effective teaching, and that these differences corresponded to differences in cultural context.

[5] Ryan and Deci (2000c) disagree with this interpretation. While arguing that a sense of autonomy is necessary to sustain intrinsic motivation, they emphasise that autonomy does not necessarily mean independent or isolated action, but rather 'the extent to which people genuinely and authentically *concur* with the forces that do influence their behaviour' (p. 328, emphasis original). This view is compatible with S.S. Iyengar and Lepper's (1999) findings.

2.5 Conceptual framework: mapping teacher accountability pathways and mechanisms

Taken together, the research discussed in the previous sections suggests that whether accountability instruments will prompt desirable or undesirable changes in teacher motivation depends partly on the compatibility between the instruments and teachers' culturally shaped conceptions of their work. To clarify these relationships conceptually and to facilitate their testing empirically, I now present a framework for mapping the intended outcomes of teacher accountability instruments. This framework, shown in Figure 2.1, is agnostic about the types and forms of student outcomes that are most desirable. My use of 'student outcomes' encapsulates any goals for students' development, whether individual—such as cognitive or socioemotional growth—or collective—such as equity—that stakeholders in the context deem desirable (Allen, 2016; Biesta, 2011). Also, it does not imply that any particular teacher accountability instrument is inherently superior. Instead, it emphasises the interplay between different elements and actors along the policy pathway.

Figure 2.1 *Conceptual framework for mapping the intended outcomes of teacher accountability instruments*


Teacher accountability instruments
communicate information.
set standards.
allocate consequences.
Stakeholder decisions change.
Teacher motivation is raised/reoriented.
Student outcomes improve.
SOCIOCULTURAL CONTEXT
OTHER CONTEXTUAL FACTORS

Rather than attempting to comprehensively diagram all the key factors in teacher accountability processes, this conceptual framework reflects my analytic interests in this thesis. I am interested in the implementation and efficacy of teacher accountability instruments, rather than their formulation. Accordingly, the starting point of the framework is the teacher accountability instruments themselves, instead of the cumulative processes of conceptualisation, negotiation, and habituation that underlie policy formulation and/or the emergence of informal accountability practices. Also, rather than enumerating the full spectrum of contextual factors that may affect teacher accountability instruments, I only delineate 'sociocultural context' from 'other contextual factors'. This is not to imply that sociocultural context exerts as much influence as the totality of other contextual factors, but merely that this thesis focuses on sociocultural context.

This framework was developed iteratively over the course of this project. I developed the overall pathways—i.e. from teacher accountability instruments to student outcomes via changes in teacher motivation or in stakeholder decisions—based on my reading of the literature, prior to conducting my fieldwork. At that point, I was also interested in identifying the mechanisms underlying such changes in teacher motivation, and I had tentatively outlined some potential mechanisms. However, it was only when analysing my field interviews that I became convinced that the three mechanisms shown in Figure 2.1—i.e. communicating information, setting standards, and allocating consequences—encompassed all the motivational effects that interview participants had mentioned. For example, prior to analysing the interviews, I had posited a separate mechanism wherein teachers' motivation may be increased through their desire to protect professional reputations or maintain collegial esteem, as implied in some discussions of professional accountability (e.g. Fullan, et al., 2015; Müller & Hernández, 2010) and informal accountability (e.g. Romzek, 2014). However, as I analysed the data, it became clear that the mechanism underlying such collegial accountability was the same mechanism underlying informational accountability instruments, as I discuss below and in Section 5.4. Hence their consolidation into a single mechanism. The alignment between these three mechanisms and the three categories in my definition of teacher accountability instruments is a happy coincidence, since the categories in this definition have been largely unchanged since I submitted my registration report in September 2017. To recall, in Section 2.1 I defined 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'.

Pathways from teacher accountability instruments to student outcomes

As shown in Figure 2.1, I propose two causal pathways by which teacher accountability instruments may improve student outcomes: either by (a) changing stakeholder decisions, or (b) raising teacher motivation or reorienting their motivation toward different tasks or goals. Such changes in stakeholder decisions and teacher motivation are what M.J. Williams (2017; see also Pollitt & Bouckaert, 2017, p. 15) calls intermediate outcomes, i.e. changes that are triggered by the policy instrument and that lead, in turn, to the desired final outcomes—in this case, improved student outcomes (however defined).

However, the pathway from stakeholder decisions to student outcomes is indirect. Even if a school leader decides to hire or fire a teacher (or a teacher decides to support or shun a colleague, or a guardian decides to enrol a child in a private school, or a policymaker decides to modify teacher compensation rules), further intermediate outcomes must occur before this decision can influence the classroom teaching and learning processes that ultimately impact student outcomes. Some of these intermediate outcomes may entail stakeholders instituting or altering teacher accountability instruments that aim to directly influence the teacher motivation pathway instead. Hence, I focus instead on the teacher motivation pathway, since teachers are frontline implementors of education policy.

It is important to note that none of the processes in Figure 2.1 are sure-fire. To illustrate, for the teacher motivation pathway, teachers' subjective perspectives are crucial determinants of whether an accountability instrument works as intended (as discussed in Section 2.4). As McLaughlin (1987) argues, successful policy implementation is not only a matter of local capacity, but also of will, i.e. 'the attitudes, motivation, and beliefs that underlie an implementor's response to a policy's goals or strategies' (p. 172). Teachers may not usually be active participants in formulating accountability policies, but they certainly have some agency over how they respond to policy instruments. For example, under a national reform in England emphasising performance-related pay, one of the three most common reasons given by headteachers for revising their pay policies was to raise teacher motivation (Sharp et al., 2017). However, only 27% of surveyed teachers agreed that these performance-related pay policies helped to motivate underperforming teachers, and only 38% agreed that these policies further motivated teachers who were already performing well (ibid)—indicating that many teachers did not respond to the pay policies as principals intended.

Additionally, note that ‘teacher motivation’ in the framework can be interpreted either individually or collectively. An accountability instrument could raise the motivation of individual teachers by prompting them to work harder, or it could raise the collective motivation of teachers in a given setting by influencing the configuration of people who enter and remain in the teaching profession. In prior studies, this distinction has been conceptualised as incentives vs. sorting (Lazear, 2000, 2003), motivation effects vs. selection effects (Podgursky & Springer, 2007) or changes in effort vs. compositional change (Biasi, 2018). Such selection effects can be triggered by various instruments, such as salary structures intended to render the profession unattractive to unmotivated teachers (e.g. Gerhart & Fang, 2017; Leaver, Ozier, Serneels, & Zeitlin, 2019), or entry criteria for pre-service training. Selection-oriented instruments fall within the conceptualisation of accountability as a principal-agent relationship, because identifying the agent is no less important than determining the agent’s obligations, as noted in Section 2.1.

Even if the level or direction of teachers’ individual or collective motivation changes positively, this may not translate into improved student outcomes. For this to happen, the changes in teacher motivation must, in turn, change teacher practice in ways that positively influence student learning. Although this is the implicit theory of change in many teacher accountability policies, the empirical evidence base for the causal influence of teacher motivation on student outcomes is not particularly strong. For example, a major review of research on teacher efficacy, which is a key construct in Bandura’s theory of motivation (see Section 2.4), found that only 6 out of the 218 reviewed studies of teacher efficacy examined its relationship with student outcomes, finding modest empirical support for this link (Klassen, Tze, Betts, & Gordon, 2011). Nonetheless, there is evidence that higher levels of teachers’ self-reported motivation are associated with better instructional practices, as reported by students (Holzberger, Philipp, & Kunter, 2014; Kunter et al., 2008). There is also some evidence that higher levels of teacher motivation are associated with higher levels of student motivation (Roth, Assor, Kanat-Maymon, & Kaplan, 2007) and with better student achievement (Caprara, Barbaranelli, Steca, & Malone, 2006). However, these relationships may be influenced by other factors, including teacher knowledge (Keller, Neumann, & Fischer, 2017).

Finally, it is also worth noting that even though the stakeholder decisions pathway and the teacher motivation pathway are clearly delineated conceptually, this distinction may be harder to identify empirically since teachers’ and other stakeholders’ interactions with accountability instruments are often entangled. For example, one analysis of school-level accountability grades in New York found that schools receiving unfavourable grades subsequently experienced lower teacher turnover, which suggests a rise in teacher motivation; but that this lower turnover is likely due to principals' actions, which instead indicates changes in stakeholder decisions in response to the accountability grades (Dizon-Ross, 2018). Entanglements notwithstanding, it is worthwhile to distinguish between the stakeholder and teacher pathways, both for analytic clarity and for effective policy design.

Mechanisms for influencing teacher motivation: information, standards, and consequences

Besides drawing a distinction between the stakeholder and teacher pathways of change, my framework also recognises three different mechanisms through which such change can be initiated. These mechanisms are: communicating information on teacher practice, thus informing stakeholders' decisions or activating a teacher's desire to compare favourably with a set of expectations; setting standards for teacher practice, thus directing the teacher's efforts toward certain expectations; and allocating consequences based on judgements of teacher practice, thus creating incentives to gain rewards and avoid penalties.

One reason why it is worth exploring such mechanisms empirically is that seemingly similar teacher accountability instruments can lead to dramatically different outcomes. As discussed in Section 2.1, this variability can stem not only from variations in policy design and implementation quality, but also from variations in context. Furthermore, the fact that radically different sets of teacher accountability instruments can contribute to equally desirable student outcomes—as in Finland's and Singapore's respective approaches to teacher accountability—indicates multiple pathways of change. This suggests that there is much to be learnt from examining the mechanisms underlying the effects of teacher accountability instruments.

However, any attempt to analyse such mechanisms also entails a lot of variability and contingency, not least because social scientists define mechanisms in numerous different ways. Elster's definition of mechanisms as 'frequently occurring and easily recognizable causal patterns that are triggered under generally unknown conditions or with indeterminate consequences' (1998, p. 45) is widely cited but far from definitive. Mahoney (2001), for example, lists over 20 different definitions from various scholarly publications. In addition to defining mechanisms, some scholars have created typologies of mechanisms across different settings or different change processes (e.g. Hedström & Swedberg, 1998; Westhorp, 2018). Variability notwithstanding, most definitions treat mechanisms as the crucial, causal links between inputs and outputs. Focusing on such mechanisms facilitates comparison between entities or events that may differ in complex ways yet share the same underlying causal processes (Steinmetz, 2004).[6]

Apart from the variability and contingency of teacher accountability instruments, another reason why it is important to look at the mechanisms underlying teacher accountability is to counterbalance an influential strand of policy rhetoric about accountability. This strand of rhetoric advocates for a particular form of accountability—establishing quantitative metrics of success, tracking performance on these metrics, and making these metrics public—even though such metric-based accountability instruments have a mixed track record, often failing to deliver the promised improvements while triggering undesirable side effects (Muller, 2018; O'Neill, 2002; Zhao, 2018). As Dubnick (2005) argues, in order to move beyond the assumption that accountability instruments necessarily lead to improved performance, we need to examine the mechanisms underlying this presumed relationship.

Such attention to mechanisms and theories of change is central to some of the policy analysis models discussed in Section 2.2 above. For example, Cartwright and Hardie (2012) argue that evidence-based policy requires the identification of causal roles and support factors. While support factors concern the context of policy implementation, causal roles concern the mechanisms of change. Policies that play causal roles are INUS conditions for a particular effect: '*I*nsufficient but *N*ecessary part of an *U*nnecessary but *S*ufficient condition for producing a contribution to the effect' (ibid, p. 25, emphasis original; see also Mackie, 1965). In other words, (a) there are multiple sets of circumstances that could (potentially) contribute to a desired policy outcome; and (b) a particular policy instrument will only contribute positively under the right circumstances. In Cartwright and Hardie's terminology, each of the three teacher accountability mechanisms that I propose is an INUS condition for improving educational outcomes. Depending on the context, they can (but do not necessarily) influence teacher motivation or stakeholder decisions in ways that may (or may not) improve teacher practice and student outcomes.

[6] Merton (1968) makes a similar argument in his discussion of middle-range theory (see also Pawson, 2000).

One of the three mechanisms entails *communicating information on teacher practice* to stakeholders. This informational mechanism can prompt stakeholders to change their decisions, as when the publication of school rankings prompts families to avoid enrolling students in poorly ranked schools (Nunes, Reis, & Seabra, 2015) or when school evaluation ratings prompt administrators to alter budgetary allocations (Craig, Imberman, & Perdue, 2015). However, accountability instruments that use the informational mechanism can also initiate change along the teacher motivation pathway. Such motivational change occurs by activating teachers' desires to be regarded favourably when information about their practice is compared to stakeholder expectations. This can occur through accountability instruments that directly convey feedback to teachers, whether on student test results (Muralidharan & Sundararaman, 2010) or on classroom practice (Garet et al., 2017). However, the desire to compare favourably with stakeholder expectations can be triggered merely by the consciousness that information is being collected about their practice, even if the information is not relayed to teachers themselves. Psychology experiments have shown that the mere awareness of being observed can raise cognitive effort (Lerner & Tetlock, 1999), probably due to the desire to safeguard or improve one's image (Tetlock, 1991). In education, this awareness can result from tacit monitoring by colleagues (Ahmad, 2016; Müller & Hernández, 2010, p. 313), with one school promoting such monitoring by designing classroom windows that render teachers visible to others in the hallway (Gill, Lerner, & Meosky, 2016).

Informational instruments for teacher accountability can vary greatly, encompassing compulsory standardised questionnaires on classroom activities and casual troubleshooting sessions with colleagues. As noted, the recipients of the information can be other stakeholders or teachers themselves, or both. (For example, Singapore's teacher appraisal system, which I describe in Section 5.2, provides developmental feedback to teachers while also providing school leaders and government officials with data for workforce planning.) The stakeholder expectations to which the information is compared may be codified or tacit, shared or unilateral, precise or vague, externally imposed or personally espoused.[7]

[7] I suspect that the direct, targeted action of human stakeholders may not be necessary to trigger the informational mechanism. Conceivably, a teacher who receives an automated electronic report showing their pupils' test scores as compared to a national distribution may be prompted to work harder or to work differently. Even the mere requirement for a teacher to report on test scores or lesson objectives may activate the consciousness that their practice is being compared to standards and, thus, may alter the level or direction of their motivation. However, even such impersonal instruments were designed by human stakeholders at some point. Moreover, all of the participants in my field interviews referred to specific people or groups of people when describing the workings of informational accountability instruments, so I do not focus on the impersonal possibilities in my analysis.

These expectations may come from accountability instruments that invoke the second mechanism, which is to *set standards for teacher practice*. Unlike the informational mechanism, which operates through teachers' awareness that their practice is being actively compared to a set of expectations, the standard-setting mechanism passively directs teacher motivation toward general or particular expectations set by stakeholders. As noted in Section 2.1, standard-setting is central to teacher accountability because education serves many competing and amorphous goals. Passivity notwithstanding, standards can be highly influential. For example, accountability standards in the United States have triggered substantial changes in kindergarten teachers' instructional priorities, even though their pupils are not subject to monitoring via standardised tests (Bassok, Latham, & Rorem, 2016). Within principal-agent relationships, standard-setting pertains not only to expectations for the results of teachers' work, but also to reciprocal expectations for the conditions under which they do that work, as noted in Section 2.1. If the principals—whether actual school principals, the national or subnational government, parents, and the local community—provide teachers with an appropriate standard of support, then teachers as the agents may feel a corresponding obligation to meet the standards set for their work, even if these standards are passive. Some instruments may play a dual role in setting standards for teacher practice while providing support in the form of guidance and structure for meeting those standards. In Finland, textbooks—which supplement the standards set by the national curriculum—influence teachers' pedagogical decisions extensively, even though teachers are usually free to choose which textbooks their classes will use (Crehan, 2016, pp. 59–60; Törnroos, 2005; Viholainen, Partanen, Piiroinen, Asikainen, & Hirvonen, 2015). Other standard-setting instruments include teacher codes of conduct, school-level goals, and criteria for entry into the teaching profession, such as licensure requirements or admissions screening for initial teacher training. (Refer to the previous subsection for a discussion of entry standards as accountability instruments that influence teacher practice collectively at the point of becoming an agent in the accountability relationship, rather than individually in-service).

The third mechanism in my conceptual framework is that accountability instruments can *allocate consequences based on stakeholders' judgements of teacher practice*, thus influencing teacher motivation via the desire to gain rewards and avoid penalties. These consequences can be individual (e.g. merit-based career progression) or collective (e.g. budgetary incentives for school-level improvements). As discussed in Section 2.1, such performance-based consequences have been shown to influence teachers considerably, whether in desirable ways, such as raising the amount of time that teachers invest in their work (Duflo et al., 2012; Lavy, 2009), or undesirable ones, such as perverse behaviours that aim to raise test scores rather than meaningfully improving student learning (Booher-Jennings, 2005; Vogell, 2011).

An accountability instrument may involve more than one mechanism, as with a self-evaluation form that collects information on teacher practice while concurrently specifying standards via the categories of information required. Also, many teacher accountability approaches deploy several interrelated instruments which use different mechanisms, as with pay-for-performance schemes that award bonuses based on information about student test scores (e.g. Chiang et al., 2015).

The role of context

For any teacher accountability instrument or set of instruments, its mechanism(s) will only work as intended if the context is conducive. The influence of contexts—cultural and historical, national and local, sectoral and institutional—in policy implementation has been emphasised in studies of public policy in general (Lipsky, 2010; McLaughlin, 1987; Weaver, 2010), as well as in education policy (Ball et al., 2012; Spillane, 2009) and in educational assessment and accountability in particular (Easley & Tulowitzki, 2016; Feniger & Lefstein, 2014).

Since teaching involves numerous interacting factors, a given context will only be conducive to the desired effects of an accountability instrument if multiple factors align. For example, even if a teacher accountability instrument successfully raises or reorients a teacher's motivation, this may not lead to any improvements in student outcomes if the teacher lacks the training to teach effectively. These contextual enablers and constraints operate at numerous levels. The success of accountability instruments can hinge on a single district administrator (e.g. Skrla et al., 2011) or a few village elders (e.g. R. Iyengar, 2012). Moreover, context can affect the efficacy of any of the steps along the pathways in Figure 2.1. A teacher accountability instrument may fail to improve student outcomes because it does not fit teachers' socioculturally embedded beliefs about educational priorities and thus does not influence their motivation (e.g. Mizel, 2009). Equally, a teacher accountability instrument may boost teacher motivation but fail to improve student outcomes because of other contextual constraints, such as inadequate resources, as noted in Section 2.1. This may have been the case in a performance-based pay experiment in Uganda, which found that teachers across treatment schools exerted more effort than those in the control group, but student outcomes only improved in treatment schools that had textbooks (Gilligan et al., 2018).

Among these various aspects and levels of context, I focus in this project on sociocultural context at the national level. I focus on sociocultural patterns because, as shown, they can have profound effects on teacher accountability processes, yet there has been little systematic comparative research on the relationship between teacher accountability policy and sociocultural context. Having chosen to investigate sociocultural factors, it made sense to also focus on national-level differences. Pragmatically, there are public-access secondary datasets measuring sociocultural variation at the national level (EVS, 2011; Hofstede, 2001; Inglehart et al., 2014b). Moreover, a significant amount of policy discourse about teacher accountability is oriented toward cross-country variation, not least because of ILSA league tables and their associated 'best practices' discussions.

How this relates to existing frameworks

There is hardly a shortage of frameworks for analysing different aspects of accountability (e.g. Dubnick, 2003; Koppell, 2005; Romzek & Dubnick, 1987; see Verger & Parcerisa, 2017b for an overview of typologies in educational accountability). Furthermore, as noted in Section 2.1, there is also no lack of policy analysis models that unpick the theories of change and enabling conditions implicit in policy programmes. However, my conceptual framework differs from existing frameworks in ways that, I believe, make it uniquely suited for thinking through the process of influencing teacher practice using accountability policy instruments and for identifying potential pitfalls along the intended causal pathway.

Established conceptual frameworks for accountability relationships within education include Pritchett's (2015; see also World Bank, 2003) accountability triangle, which identifies different accountability relationships between educational stakeholders (e.g. the relationship of client power from families and communities to frontline providers), with four design elements (i.e. delegation, finance, information, and motivation) within each accountability relationship. My framework zooms in on a single set of stakeholders, i.e. teachers, and focuses primarily on the element of motivation. Abelmann and Elmore (1999) develop a framework that is narrower than Pritchett's landscape of accountability stakeholders but broader than my framework for teacher accountability. Their Venn diagrams for analysing accountability within schools look at overlaps and mismatches between collective expectations, individual responsibility, and accountability (i.e. account-giving practices). I am likewise interested in such overlaps and mismatches, but whereas their emphasis is on diagramming the interplay of school-based actors' conceptions of accountability, mine is on examining how accountability instruments can change student outcomes by influencing teacher motivation. Firestone and Pennell's (1993) framework and mine are more similar in that both look at the relationship between teacher accountability instruments (which they call 'incentive policies') and teacher motivation ('teacher commitment'). However, they differ in that Firestone and Pennell identify in detail various elements within this relationship (e.g. variables for working conditions, such as job design, feedback, and collaboration), whereas I take a coarser-grained view of this relationship but also consider the link from teacher motivation to student outcomes.

Another difference between Firestone and Pennell's (1993) framework and mine is that I take a fairly broad view of what constitutes a teacher accountability instrument, whereas they look primarily at 'differential incentive policies', i.e. material or professional rewards that are allocated based on some assessment of performance. Bruns and Luque's (2014) threefold classification of incentives that motivate teachers (i.e. 'professional rewards, accountability pressure, and financial incentives', p. 224) likewise foregrounds incentives. In contrast, although two of the mechanisms in my framework relate to incentives; with an obvious link to incentives in the consequence-based mechanism and an indirect link in the informational mechanism, which can reward teachers emotionally through the awareness that they have satisfied the expectations that are stated or implied in an informational instrument. However, the standard-setting mechanism does not invoke incentives.

By affirming the passive but potentially influential role of standard-setting mechanisms in teacher practice, and by drawing a distinction between the standard-setting and informational mechanisms, I also diverge from Lerner and Tetlock's (1999) typology of four psychological mechanisms for accountability (see also Gill et al., 2016, for an application of their typology education). While my consequence-based mechanism coincides with their 'evaluation', and my informational mechanism encompasses their other three mechanisms (mere presence of another, identifiability, and reason-giving), they do not have an equivalent for my standard-setting mechanism.[8] On the other hand, McDonnell and Elmore's (1987) classification of policy instruments does include standard-setting instruments ('mandates') and consequence-based instruments ('inducements'), but it does not treat informational instruments separately. (Their classification also has additional categories for capacity-building and system-changing instruments that extend beyond the realm of teacher accountability.)

In short, my conceptual framework has a distinctive focus on teacher accountability as a subset of educational accountability. It traces the teacher accountability pathway from the instruments that generate accountability to the hoped-for student outcomes at the other end of the process. Along the way, it categorises different mechanisms by which accountability instruments can seek to influence teacher motivation—offering a menu for policy planning similar to McDonnell and Elmore's (1987) policy instruments and Hood's (1983; Hood & Margetts, 2007) tools of government, but in a small, specific policy area.

2.6 Research questions

To test the validity of this conceptual framework, I investigate three research questions:

1. *To what extent does the influence of teacher accountability instruments on student outcomes depend on sociocultural context?*
First, I look at the start and end points of the framework, and consider the evidence for whether there is a link between teacher accountability instruments and student outcomes to begin with, and the degree to which this link is affected by sociocultural context

2. *To what extent, and how, does teacher motivation mediate the influence of teacher accountability instruments on student outcomes?*
Next, I examine the pathway from teacher accountability instruments to student outcomes in order to determine whether accountability instruments influence teacher motivation as an intermediate step along this pathway. I also look at empirical evidence for the three teacher

[8] This is likely due to the different empirical approaches of our frameworks. Lerner and Tetlock (1999) parse out subtle psychological manipulations to be tested in laboratory settings, where the conditions that yield observable changes in the level and manifestation of participants' motivation within the span of an experiment would be of greater interest than the standards that direct pre-existing motivation towards particular goals and practices. In contrast, my data on accountability mechanisms come from interviews about teachers' daily work, in which standards play an important guiding role, and where it can be messier to distinguish between, for example, the effects identifiability versus that of the presence of an observer.

accountability mechanisms that I have proposed: setting standards, communicating information, and allocating consequences.

3. *To what extent does the influence of teacher accountability instruments on teacher motivation depend on sociocultural context?*

Finally, I consider the degree to which sociocultural context affects the relationship between teacher accountability instruments and teacher motivation. In the next chapter, I lay out my approach to answering these questions.

Chapter 3: Methodology

As discussed above, the aim of this research project is to investigate some of the ways in which sociocultural context may (or may not) influence the processes underlying teacher accountability policy. This investigation is designed as an attempt to validate a conceptual framework that maps potential pathways from teacher accountability instruments to their intended outcomes. Each of the three research questions proposed above in Section 2.6 focuses on a subset of the relationships within this conceptual framework.

In this chapter, I describe the methodology I use to address the proposed research questions. I begin in Section 3.1 by explaining why I take a realist-constructivist theoretical standpoint, and what this perspective implies for my research design. Next, in Section 3.2, I outline the overall research design, linking each research question to empirical sources and analytic methods. I then describe each of the two empirical sources—cross-country surveys of education and culture, and teacher interviews in Finland and Singapore—and how I analyse each source, in Sections 3.3 and 3.4 respectively. Finally, I discuss the ethical considerations and limitations of this project in Sections 3.5 and 3.6.

3.1 Theoretical perspective: mechanisms, realist research, and validity

Why adopt a realist ontology with a constructivist epistemology?

As described in Section 2.5, my conceptual framework posits three mechanisms by which teacher accountability instruments can attempt to influence the level or direction of teacher motivation and, thus, student outcomes. Much of the social scientific research on causal mechanisms falls under the banner of realism (Porpora, 2015). One of the most prominent articulations here is Pawson and Tilley's (1997) seminal book on realist policy evaluation, with its 'mechanism + context = outcome' (p. xv and elsewhere) formula and its emphasis on 'what works for whom in what circumstances' (p. 210 and elsewhere). Pawson and Tilley's work is particularly relevant given their conception of mechanisms as 'the choices and capacities which lead to regular patterns of social behavior' (ibid, p. 216). They argue that social programmes generate change by influencing actors' reasoning and, thus, behaviour (Pawson, 2013). While there are other forms of mechanisms that do not operate primarily via actors' decision-making (see Westhorp, 2018, for some examples), Pawson and Tilley's reasoning-focused conception of social behaviour suffices for this project, given my focus on teacher motivation, which is itself a cognitive process. I propose that teacher accountability instruments can generate improvements in student outcomes by influencing teachers' motivation and, thus, classroom practice.

Beyond Pawson and Tilley, scientific realism is a broad school of thought, spanning philosophy, sociology, and policy evaluation, among other disciplines (Maxwell, 2012; Pawson, 2018). The unifying element here is a realist ontology, i.e. the view that there are real entities that exist independently of our perceptions of them (Maxwell, 2012). One implication of this ontological stance is that causation is real. Realist researchers say that A causes B not because they observe that the occurrence of A is consistently followed by the occurrence of B, as in black-box understandings of causality that draw on Hume's 'constant conjunction' argument. Rather, realists deem A to cause B if A actually generates B—in other words, A is a mechanism that, under certain circumstances, can lead to outcome B (Pawson & Tilley, 1997). This notion of causality accords with the implicit assumption of teacher accountability policymaking: that such policy instruments can meaningfully change student outcomes. (This is not to deny that some teacher accountability policies may be instituted with the more cynical aim of superficially appeasing voter demands or deflecting blame for subpar student outcomes [see Hood, 2011]. However, since improving student outcomes is not only a more valuable goal, but also one which—hopefully—guides the bulk of teacher accountability policymaking, I focus on this goal rather than its more cynical counterparts.)

Alongside this realist ontology, I adopt a constructivist epistemology that is shared by many strands of realism. That is, although entities in the world exist independently of our perceptions, our understanding of these entities is necessarily contingent and fallible (Maxwell & Mittapalli, 2010). In Hammersley's (1998) words, 'No knowledge is certain, but knowledge claims can be judged in terms of their likely truth' (p. 66)—implying a constructivist epistemology—and, 'There are phenomena independent of us as researchers or readers of which we can have such knowledge' (p. 66)—implying a realist ontology. (This combination of a realist ontology and a constructivist epistemology is sometimes called 'critical realism'. However, some strands of critical realism include theoretical commitments that would add unnecessary complexity to this project, e.g. Bhaskar's dialectics [1993; see Gorski, 2013 for a short overview]. For simplicity's sake, I take a more general realist stance [Maxwell, 2012; Pawson & Tilley, 1997].) This realist-constructivist theoretical stance has been used extensively in health-related research (e.g. Luetsch, Twigg, Rowett, & Wong, 2019; Manzano-Santaella, 2011; Marchal, Dedzo, & Kegels, 2010), including in research on the intersections of health and education (e.g. Pearson et al., 2015 on health promotion programmes in schools; Wong, Greenhalgh, Westhorp, & Pawson, 2012 on medical education) and of healthcare provision and accountability (Maluka et al., 2011). More pertinently, this theoretical perspective has been used in a field study of teacher performance in Tanzania (Tao, 2013) and in systematic reviews of educational accountability in low- and middle-income countries (Eddy-Spicer, Ehren, Bangpan, Khatwa, & Perrone, 2016; Westhorp et al., 2014).

The subject matter of this thesis fits well with the combination of a realist ontology and constructivist epistemology. Ontologically, the focal points in this project—teacher accountability instruments, teacher motivation, and student outcomes; embedded within sociocultural contexts— are treated as real entities by actors in educational settings. For example, the measurement of student outcomes may be contentious, but the desirability of good student outcomes is widely accepted. Also, the fierce debates around teacher accountability suggest that its impact on teachers and students is palpably felt in real ways. Epistemologically, although these entities are real, they are also enmeshed in individual and social constructions that require interpretation. For example, even the most banal education policy text contains implicit normative beliefs about the purpose of education. Test scores may masquerade as objective measures, but they result from assessment items and marking schemes that comprise a series of subjective decisions about what students should know. Equally, research on teacher accountability is inevitably shaped by researchers' values and ideologies, whether tacit or explicit. Hence the aptness of a constructivist epistemology that acknowledges multiple perspectives on entities and relationships in the real world.

Realism, research design, and validity

These ontological and epistemological commitments have a few implications for my research design. For instance, I use quantitative and qualitative data sources to address different aspects of the research questions, and realism offers coherent theoretical basis for such research (Maxwell & Mittapalli, 2010). From a realist standpoint, quantitative and qualitative data sources can both provide (fallible) insight into real phenomena. A given data source or analytic method cannot be deemed better than the alternatives on the basis of data collection methods or analytic procedures (although these should be rigorous). Rather, what matters is whether the data source can meaningfully provide evidence for the particular theory under consideration. To illustrate, Porpora (2015) describes a research project in which '[t]o the extent that what we were asking here was a "how" question, what was required was a qualitative analysis', whereas for other aspects of the project, 'questions about relative frequency are quantitative in nature, answerable only by comparative counts' (p. 63). Similarly, in exploring RQ2, I use cross-country statistical analysis to discern to what extent teacher motivation mediates the relationship between teacher accountability instruments and student outcomes, whereas I use field interviews to investigate how this happens, as outlined below in Section 3.2.

Realism—at least, Pawson and Tilley's brand of realism—also had practical implications for how I use teacher interviews in Finland and Singapore to construct evidence for my conceptual framework. In particular, Pawson and Tilley (1997) argue that realism's commitment to building causal theories implies that research interviews should not be driven by the participant's ideas. Rather, 'the researcher's theory is the subject matter of the interview and the interviewee is there to confirm or falsify and, above all, to refine that theory' (p. 159). Thus, realist researchers should not approach interviews by operationalising their theories into interview questions that obliquely elicit relevant data from interview participants. Instead, researchers should teach their theories to participants, so that participants can directly address the aspects of the theory that do or do not fit their experiences and observations. I discuss how this approach to conceptual refinement influenced my field interviews in Section 3.4.

Furthermore, realist arguments about validity influenced my analytic approach. Given that I aim to validate a conceptual framework involving cognitive processes that cannot be directly observed (e.g. changes in teacher motivation), concerns about validity are paramount. Some realist theorists argue that validity does not derive primarily from the rigour of the methods used to construct empirical evidence. Rather, the validity of a descriptive or explanatory account hinges on its relationship to the entities that it aims to describe or explain (Maxwell, 2017). Consequently, a key way of strengthening the validity of a theory is by comparing it with alternative explanations (Pawson, 2013; Porpora, 2015; see also Elster, 2015 on lateral support for an explanation). For example, in answering RQ1, I propose that the efficacy of teacher accountability instruments is contingent on their compatibility with sociocultural context. However, it is also possible that this efficacy hinges primarily on non-sociocultural factors, such as the internal coherence of a given set of accountability instruments. If each of these possible explanations could independently account for the empirical evidence that I examine for RQ1, then this evidence on its own would not allow me to decide which explanation is more valid.

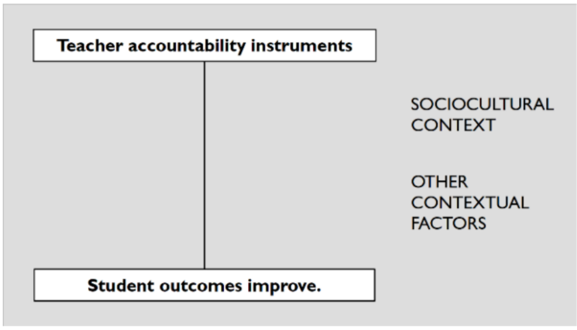
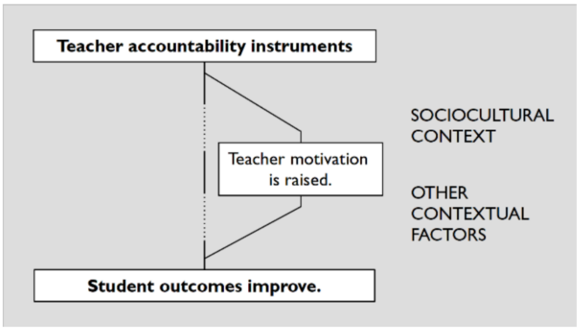
However, if there is further evidence of a separate but related phenomenon that fits the sociocultural explanation but not the alternative explanation, it is reasonable to conclude that the sociocultural explanation has greater validity, at least in this context. Hence, for each research question, I consider an alternative explanation that can also account for some of my empirical observations, and I discuss the extent to which these explanations can be refuted.

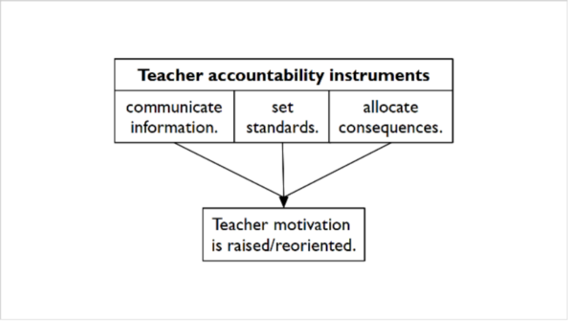
3.2 Research design overview

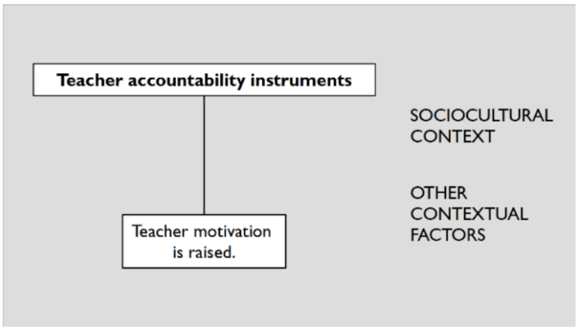
In this section, I justify my choice of research methods and overall strategy, as summarised in Table 3.1. Details about the selection, collection, and analysis of data will be discussed in subsequent sections. To use Pawson and Tilley's (1997) terminology, I examine the evidence for linking context with outcome in RQ1, mechanism with outcome in RQ2, and mechanism with context in RQ3.

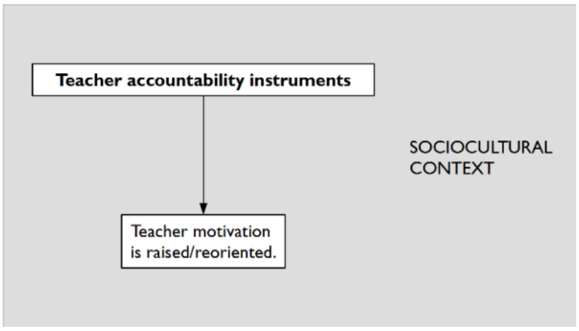
As noted in Section 2.6, the three research questions aim to test different relationships within the conceptual framework that I have proposed. The first research question (RQ1) seeks to establish whether or not there is a case for looking at socioculturally embedded mechanisms for teacher accountability in the first place. Hence, answering RQ1 entails investigating whether there is evidence for a relationship between teacher accountability instruments and student achievement, and whether this relationship is contingent on sociocultural context. Given that much of the variation in education policy structures and in sociocultural context manifests across countries rather than within them (Wagemaker, 2010; Woessmann, 2007), I investigate RQ1 using cross-country secondary datasets. Specifically, I draw on PISA 2015 data that operationalise aspects of accountability and student outcomes, alongside sociocultural indicators from the World Values Survey/European Values Study (WVS/EVS) and Hofstede's Values Survey Module. Using these cross-country surveys, I construct scales for teacher accountability instruments and for theoretically salient sociocultural constructs. I then use these scales in multilevel models that test whether sociocultural context *moderates* the relationship between teacher accountability and student outcomes—that is, whether different sociocultural conditions can either intensify or attenuate this relationship. This moderation relationship is tested statistically by including an interaction between the predictor (i.e. teacher accountability) and the potential moderator (i.e. sociocultural context) in the regression model. Note that I am not making any causal arguments at this stage of the analysis. Methodologically, the analysis of cross-sectional datasets without an identification strategy does not allow for causal inference. Theoretically—and more importantly—a causal explanation in realist research requires an explication of the mechanisms that generate change, which I examine next.

Table 3.1 *Correspondence between research questions, conceptual framework, and empirical sources*

<table>
<tr><th>Cross-country statistical analysis</th><th>Teacher interviews in Finland and Singapore</th></tr>
<tr><td colspan="2">RQ1. To what extent does the influence of teacher accountability instruments on student outcomes depend on sociocultural context?</td></tr>
<tr><td>
</td><td>N/A</td></tr>
<tr><td colspan="2">RQ2. To what extent, and how, does teacher motivation mediate the influence of teacher accountability instruments on student outcomes?</td></tr>
<tr><td>
Teacher accountability instruments
Teacher motivation is raised.
SOCIOCULTURAL CONTEXT
OTHER CONTEXTUAL FACTORS
Student outcomes improve.

</td><td>
Teacher accountability instruments
communicate information.
set standards.
allocate consequences.
Teacher motivation is raised/reoriented.

</td></tr>
<tr><td colspan="2">RQ3. To what extent does the influence of teacher accountability instruments on teacher motivation depend on sociocultural context?</td></tr>
<tr><td>
Teacher accountability instruments
SOCIOCULTURAL CONTEXT
OTHER CONTEXTUAL FACTORS
Teacher motivation is raised.

</td><td>
Teacher accountability instruments
SOCIOCULTURAL CONTEXT
Teacher motivation is raised/reoriented.

</td></tr>
</table>

Having considered the overall relationship between teacher accountability, student outcomes, and sociocultural context in RQ1, I then delve into the mechanisms that may underlie this relationship. In RQ2, I test the possibility that the relationship between teacher accountability instruments and student outcomes is *mediated* by teacher motivation, such that accountability instruments cause changes in teacher motivation, which in turn change student outcomes. Statistically, mediation is tested by examining (a) whether the main predictor (i.e. teacher accountability) significantly predicts the mediator (i.e. teacher motivation), and (b) whether the addition of the mediator to the regression model shifts predictive power from the main predictor to the mediator (i.e. the coefficient on teacher motivation is significant, and there is a corresponding reduction in the coefficient on accountability instruments; Hayes, 2013). Here, the main educational dataset is TIMSS 2015, which includes a teacher-level proxy for teacher motivation.[9] Next, having examined the extent of this mediation relationship, I set aside student outcomes and investigate how teacher accountability instruments may influence teacher motivation, using the interviews that I conducted with teachers in Finland and Singapore. As I explain in Section 3.4, I chose Finland and Singapore for comparative field research because they are a pair of countries that have high student achievement, contrasting approaches to teacher accountability, and distinctly different sociocultural contexts. In addressing RQ2, I use these interviews first to map out the extent to which teacher accountability instruments have affected the motivation of the interview participants and their colleagues, and then to parse out the mechanisms that generate this motivational influence.

Subsequently, to address RQ3, I look more closely at this motivational influence and its association with wider sociocultural patterns. Specifically, I consider the extent to which the relationship between teacher accountability instruments and teacher motivation depends on sociocultural context. Formally, RQ3 mirrors RQ1, except that the outcome of interest here is the intermediate outcome, i.e. teacher motivation, rather than the final outcome, i.e. student learning. Similar to the investigation of RQ1, I also investigate RQ3 using a series of moderated multilevel models. This time, the main educational dataset is TALIS 2013, which has teacher-reported indicators of teacher motivation as well as school-level indicators of teacher accountability. I also draw on my field interviews with teachers, to establish the degree to which the interview participants themselves believe that the influence of teacher accountability instruments on their motivation is contingent on sociocultural context. I also use these interviews to construct detailed descriptions of Finland's and Singapore's sociocultural contexts, thus adding nuance to the summary scales obtained from the standardised cross-country surveys.

[9] I use TIMSS 2015 as the main dataset because the proxies for teacher motivation in the PISA datasets are drawn from school-level (i.e. principal-reported) questionnaires. However, TIMSS 2015 does not include questionnaire items on teacher accountability. Instead, as I detail in Section 3.3, I match the TIMSS data to country-level weighted means of the PISA teacher accountability scales.

3.3 Statistical analysis of cross-country datasets

Since institutional setups and education policies are often quite homogeneous within countries, cross-country statistical datasets, such as the OECD's PISA, can be an important source of information for analysing the effects of different institutions and policies on student learning (Hanushek & Woessmann, 2011; Wagemaker, 2010). Such international large-scale assessments (ILSAs) are especially informative for policy areas where scale and complexity make randomised-control trials difficult, such as national-level accountability systems.

While ILSAs capture some cross-country variation in education policy, they lack indicators of other cross-country influences on student outcomes, such as culture (Hanushek & Woessmann, 2011). The influence of sociocultural context on educational outcomes is indeterminate but probably far-reaching, as suggested by studies in which students of East Asian heritage in Oceania attained PISA scores that were closer to their counterparts in Asia than to peers in their countries of residence (Feniger & Lefstein, 2014; Jerrim, 2015). One statistical strategy for dealing with cultural influence is to use country-level fixed effects, thus parsing out such unmeasured variation. However, since this sociocultural variation is a key variable of interest in this study, rather than a nuisance to be 'dummied out' of the regression, I employ an alternative strategy: incorporating this variation into the model using contextual data from other sources.[10]

While the national economic context, as proxied by GDP-related measures, is often incorporated into ILSA analyses, fewer analyses include data on the national 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 external

[10] As Clarke, Crawford, Steele, and Vignoles (2010) argue, the comparative quality of estimates from fixed-effects versus random-effects models depends on a number of factors. For example, random-effects models take into account the size of higher-level units (e.g. schools), thus reducing the influence of unusually small higher-level units that may otherwise distort the results. However, random-effects models also require the additional assumption that higher-level residuals (e.g. unobserved school-level variation that is associated with student outcomes) are uncorrelated with the covariates in the model. If this assumption is violated (e.g. students are sorted into schools based on non-random school-level covariates that are correlated with some other factors which are not measured in the dataset), then fixed-effects estimates may be superior. Nevertheless, these considerations are not central to my decision to use random-effects modelling. Instead, my decision is driven by the prior consideration that one of the key variables of interest—i.e. sociocultural context—is only measured at the country level in the available survey datasets. Hence, models with country-level fixed effects are unable to address my research questions. I recognise that my regression models may well incorporate bias from non-random sorting of students and schools into countries. To mitigate this bias statistically, I include socioeconomic status as a student-level covariate and GDP as country-level covariate to somewhat control for between-country sorting, and I run multiple sensitivity checks. To mitigate this bias analytically, I do not use these cross-sectional random-effects models to make strong causal claims, but rather to demonstrate broad associations.

contextual data in causal identification, by constructing instrumental-variable models that use historic data on Catholic population shares and on whether Catholicism was the state religion to provide a source of exogenous variation in private school enrolment. However, I do not use the sociocultural indicators as instruments because I am interested in sociocultural context in itself. (Besides, I am investigating a moderation relationship, and it is difficult to conceive of an instrument providing measurable source of exogenous variation in the interaction between teacher accountability instruments and sociocultural context.)

A second use of external contextual data in ILSA analysis is to identify country-level correlates of various student outcomes. Some scholars have looked at straightforward relationships between country-level correlates and student achievement, with little agreement between these studies. Specifically, Benoliel and Berkovich's (2018) analysis of PISA 2012 data and six sociocultural indicators from the WVS finds that the strongest positive sociocultural predictor of student achievement is conservatism, i.e. valuing security, the status quo, and traditional social roles, among other things. In contrast, He, van de Vijver, and Kulikova (2017) examine a wider set of sociocultural indicators alongside both PISA and TIMSS data, and find that conservatism and uncertainty avoidance have a significant *negative* association with student achievement, whereas there are significant positive associations between student achievement and sociocultural indicators related to modernity (such as autonomy, individualism, and favouring secular-rational authority). H.-D. Meyer and Schiller (2013), in turn, argue that there are two different clusters of high-performing PISA countries: a group of Western countries that have high scores for individualism but low scores on power distance (i.e. low acceptance of hierarchy), and a group of East Asian countries for which the converse is true. Others have used national contextual data to show that gender gaps in mathematics proficiency are influenced by societal gender norms (Guiso, Monte, Sapienza, & Zingales, 2008; Rodríguez-Planas & Nollenberger, 2018).[11] S.W. Han, Borgonovi, and Guerriero (2018) examine PISA data alongside WVS data and show that students are slightly more likely to aspire to teaching careers in countries where respect and responsibility are regarded important job characteristics.

Finally, cross-country datasets on education and culture can be combined to examine how contextual variables moderate—that is, either intensify or attenuate—the relationship between educational inputs and outcomes. Using data from both PISA and Hofstede's survey, Chiu and

[11] However, Fryer and Levitt (2010) suggest that the Guiso et al (2008) analysis may be vulnerable to the composition of the country-level sample.

Klassen (2010) find that some aspects of sociocultural context moderate the relationship between students' mathematics self-concept and their mathematics proficiency. Specifically, in countries that are more hierarchical, more uncertainty-averse, or more 'masculine',[12] a student's perception of their mathematical competence was less likely to correspond with their demonstrated mathematics proficiency in PISA. However, these moderation effects are small. 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, teacher motivation, and student achievement. While other studies have investigated whether the relationship between accountability instruments and student outcomes is moderated by other institutional features (e.g. school autonomy, as in Hanushek, Link, & Woessmann, 2013; and Woessmann, 2016) or by average country-level educational achievement (e.g. Bergbauer et al., 2018), this study is novel, to my knowledge, in investigating moderation effects from sociocultural context.[13]

I used a few computer programmes for these statistical analyses. I used SPSS 22 to clean and merge all the datasets, since the OECD and the IEA provide SPSS macros for manipulating ILSA data. I also used SPSS to construct scales for social capital using factor analysis, as reported in Section 4.2. However, I subsequently decided to derive teacher accountability scores using item-response theory (IRT) modelling, a function which SPSS 22 lacks. Consequently, I ran the IRT models in Stata, as reported in Section 4.1. Finally, I used MLwiN for the multilevel regression models, as noted in Section 3.3. This latter decision was because MLwiN has specialist

[12] In Hofstede's (2001; Chiu & Klassen, 2010) framework, more masculine cultures place more value on employment that brings higher earnings, recognition, opportunities for advancement, and greater levels of challenge; whereas more feminine cultures place more value on employment that involves job security, the opportunity to live in a desirable area, and good relationships with colleagues and with managers.

[13] In addition to the systematic literature search on teacher accountability and sociocultural context described in Section 2.3, I checked specifically for prior statistical analyses by searching in Scopus on 13 August 2019 for studies containing the term 'accountability' as well as either (a) 'World Values Survey' or 'European Values Study' (10 results); (b) 'Inglehart' or 'Welzel', i.e. the two most prominent WVS/EVS researchers (1 result); or (c) 'Hofstede' (10 results, including duplicates from the two other searches). None of the studies from this search investigated the extent to which sociocultural context moderates the effect of accountability instruments on an outcome. Topically, the closest matches were two cross-country statistical analyses examining correlations between corruption perception scales and sociocultural scales (Akbar & Vujić, 2014; Stanfill et al., 2016) and one cross-country statistical analysis on the relationship between education levels and preferences for democracy (A. Chong & Gradstein, 2015). In terms of statistical approaches, the closest match was a multilevel moderated mediation exploring whether the influence of development (predictor) on gay and lesbian rights (outcome) was mediated by tolerance toward the LGBTQ community (derived from the WVS) and moderated by political regime type (Hildebrandt, Trüdinger, & Wyss, 2018).

functions that facilitate multilevel modelling, and I had received training in MLwiN at the multilevel modelling course that I completed at the Essex Summer School in 2017.

Data and sampling

All statistical analyses in this study used publicly available secondary datasets. These datasets cover a range of educational and sociocultural constructs that are of theoretical interest to this study, as shown in Table 3.2. The table also shows how each dataset is used in addressing the different research questions, which I discuss in subsequent subsections. 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.

Table 3.2 *Levels of measurement across secondary datasets, and alignment of datasets with research questions*

	PISA 2015	PISA 2012	TIMSS 2015	TALIS 2013	WVS /EVS	Hofstede	Penn World Table
Teacher accountability	School	School		School			
Student outcomes	Pupil	Pupil	Pupil				
Teacher motivation		School	Teacher	Teacher			
Sociocultural context					Individual → Country	Individual → Country	
Control variables	Pupil, School	Pupil, School	Pupil, Teacher, School	Teacher, School			Country
RQ1	Main dataset	Sensitivity checks	Sensitivity checks	—	✓	✓	✓
RQ2	*	Sensitivity checks	Main dataset	—	✓	✓	✓
RQ3	*	Sensitivity checks	Sensitivity checks	Main dataset	✓	✓	✓

*PISA 2015 outcome data are not used to investigate RQ2 and RQ3; however, PISA 2015 country-level weighted means for teacher accountability are used in the RQ2 and RQ3 regressions with TIMSS 2015 outcome data.

Missing values are excluded listwise. There is little consensus about the proportion of missing data that will result in biased estimates, with suggested cut-offs ranging from 5% to 20% (Schlomer, Bauman, & Card, 2010). In my datasets, the level of missingness is around the lower cut-off for the proportion of total missing values across all cases (i.e. 5.8% missing values for PISA 2015, 3.8% for TIMSS 2015, and 5.9% for TALIS 2013) and comfortably within the higher suggested cut-off for cases with any missing values (i.e. 10.0% for PISA 2015, 12.5% for TIMSS 2015, and 15.9% for TALIS 2013). Moreover, there are no missing values for the main outcome variable, student proficiency; nor at the country level, where variables of key policy interest are located (i.e. country-level teacher accountability and sociocultural context). At the levels of analysis that do have missing data, the regressions are more than adequately powered due to the large samples of students, teachers, and schools. Rather than listwise deletion, two recent multilevel analyses of PISA data used group-mean imputation of missing values (Bergbauer et al., 2018; Heller-Sahlgren, 2018). However, when I attempted group-mean imputation with a PISA 2015 regression, this did not change coefficient estimates, but it did reduce the standard errors—which I believe to be an erroneous reduction because the group-mean imputation inadvertently lessened the variance of the data. Another technique, multiple imputation of missing values, may introduce additional bias to my analysis because the data appear to be 'not missing at random' (NMAR) in indeterminate ways. Little's test rules out the possibility of the data being 'missing completely at random' (MCAR), and the lack of correlation between dummy variables for missingness and the explanatory variables in the model suggests that 'missing at random' (MAR) does not apply (Schlomer et al, 2010). Given that both of these methods for imputing missing data may introduce additional bias and measurement error, I follow Cheema's (2014) guidelines and use listwise deletion for efficiency's sake.

One of the ILSA datasets that I analyse is the 2015 wave of PISA, the OECD's Programme for International Student Assessment (OECD, 2016d). 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 related to 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. These countries are identified in Table 3.3. As shown in the table, there were 14 countries that participated in PISA 2015 but were not included in this dataset because they had not participated in the sociocultural surveys. However, in some sensitivity checks, I include countries for which data are available for only a subset of the sociocultural indicators, even if they do not have the full complement of WVS/EVS and Hofstede indicators. (I discuss sensitivity checks in more detail below.)

Table 3.3 *Country participation in the educational and sociocultural surveys*

Country	**PISA 2015**	**PISA 2012**	**TIMSS 2015 (8th grade)**	**TALIS 2013**	**In at least one educational survey and also in**	
					WVS 5/WVS 6 /EVS 4	Hofstede
Albania					2008	
Algeria	○				2013	
Argentina	*Buenos Aires*	●			2006, 2013	✓
Australia	●	●	●	●	2005, 2012	✓
Austria	●	●			2008	✓
Bahrain			○		2014	
Belgium	●	●		*Flanders*	2009	✓
Botswana			○			
Brazil	●	●		●	2006, 2014	✓
Bulgaria	●	●		●	2005, 2008	✓
Canada	●	●	●	*Alberta*	2006	✓
Chile	●	●	●	●	2006, 2011	✓
China	*B-S-J-G*	*Shanghai*		*Shanghai*	2007, 2012	✓
Colombia	●	○			2005, 2012	✓
Costa Rica	○	○				✓
Croatia	●	●		●	2008	✓
Cyprus		○		○	2008, 2012	
Czech Republic	●	●		●	2008	✓
Denmark	●	●		●	2008	✓
Dominican Republic	○					✓*
Egypt			○		2008, 2013	✓*
Estonia	●	●		●	2008, 2011	✓
Finland	●	●		●	2005, 2009	✓
France	●	●		●	2006, 2008	✓
Georgia	○		○	○	2009, 2014	
Germany	●	●			2009, 2013	✓
Greece	●	●			2008	✓
Hong Kong	●	○	●		2005, 2013	✓
Hungary	●	●	●		2009, 2009	✓
Iceland	●	●			2010	✓*
Indonesia	●	●			2006	✓
Iran			○		2007	✓
Ireland	●	●	●		2008	✓
Israel	○	○	○	○		✓
Italy	●	●	●	●	2005, 2009	✓
Japan	●	●	●	●	2005, 2010	✓
Jordan	●	●	●		2007, 2014	✓*
Kazakhstan		○	○		2011	
Kosovo	○				2008	
Kuwait			○		2014	✓*
Latvia	●	●		●	2008	✓
Lebanon	●		●		2013	✓*
Liechtenstein		○				
Lithuania	●	●	●		2008	✓
Luxembourg	●	●			2008	✓
Macao	○	○				
Macedonia	○				2008	
Malaysia		●	○		2006, 2012	✓
Malta	●		●		2008	✓
Mexico	●	●		●	2005, 2012	✓
Moldova	○				2006, 2008	
Montenegro	○	○			2008	
Morocco			○		2007, 2011	✓
Netherlands	●	●		●	2008, 2012	✓
New Zealand	○	○	○	○	2004, 2011	✓
Norway	●	●	●	●	2007, 2008	✓
Oman			○			
Peru	●	○			2006, 2012	✓
Poland	●	●		●	2008, 2012	✓
Portugal	●	●		●	2008	✓
Qatar	○	○	○		2010	
Romania	●	●		●	2008, 2012	✓
Russia	●	●	●	●	2008, 2011	✓
Saudi Arabia			○			
Serbia		●			2008	✓
Singapore	●	○	●	●	2012	✓
Slovak Republic	●	●		●	2008	✓
Slovenia	●	●	●		2008, 2011	✓
South Africa			○		2006, 2013	✓
South Korea	●	●	●	●	2005, 2010	✓
Spain	●	●		●	2008, 2012	✓
Sweden	●	●	●	●	2010, 2011	✓
Switzerland	●	●			2007, 2008	✓
Taiwan	●	●	●		2006, 2012	✓
Thailand	●	●	●		2007, 2013	✓
Trinidad and Tobago	●				2006, 2011	✓
Tunisia	○	○			2013	
Turkey	●	●	●		2009, 2011	✓
United Arab Emirates	○	○	○	*(Abu Dhabi)*		✓*
U.K. Northern Ireland	●	●			2008	✓
England	●	●	●	●	2005, 2010	
Scotland	●	●				
Wales	●	●				
United States of America	●	●	●		2006, 2011	✓
Uruguay	●	●			2006, 2011	✓
Vietnam	●	●			2006	✓
Total	**57** (+14)	**56** (+11)	**23** (+16)	**29** (+5)	75	66

● Included in the main dataset for the ILSA cycle.

○ Participated in the ILSA cycle but not included in the main dataset due to unavailability of sociocultural, accountability, or GDP data.

* Data source is not Hofstede's personal work, but a replication that he has endorsed.

Note. Regional (rather than national) survey participation in the survey is denoted by the name of the participating regions (where B-S-J-G stands for Beijing-Shanghai-Jiangsu-Guangdong). The WVS5/WVS6/EVS4 column shows the two most recent year(s) of data collection, where applicable. The four nations of the United Kingdom were treated as separate countries due to uneven participation across the surveys, as shown. Countries that took part in a survey but were excluded because of data quality or availability issues are not shown (PISA 2015: Albania, Cyprus, Kazakhstan, Malaysia; PISA 2012: Albania; TALIS 2013: Iceland, Malaysia, Serbia, United States). Some countries were excluded from the main datasets because they lacked data on particular WVS/EVS questions despite having participated in the pertinent WVS/EVS wave. The TIMSS 2015 main dataset is paired with country-level teacher accountability data from PISA 2015. Penn World Table data are available for all listed countries except Kosovo and Liechtenstein. Sources: EVS (2011); Feenstra, Inklaar, & Timmer (2016); Hofstede (2015); itim International (2017); Inglehart et al. (2014a, 2014b); Martin et al. (2016a); OECD (2014a, 2014c, 2016d).

Additionally, I also use data from PISA 2012 (OECD, 2014a). The main PISA 2012 analyses are based on a dataset of 375,207 pupils from 14,840 schools across 52 countries. In addition to offering a different dataset for sensitivity checks, the PISA 2012 questionnaires have a richer set of items on accountability than PISA 2015. While the 2015 cycle asked principals about processes for collecting and disseminating teacher performance information, and for setting standards for teacher performance (i.e. the first two accountability mechanisms in my conceptual framework), the 2012 cycle included these two components as well as questions about the consequences of teacher appraisals (i.e. the third teacher accountability mechanism). For more details on these accountability-related questionnaire items, see Section 4.1.

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, Mullis, Foy, et al., 2016a). 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. 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, Luedemann, Schuetz, & West, 2009). For the main TIMSS 2015 analysis that uses teacher accountability data from PISA 2015, the dataset includes 118,363 pupils taught by 6,147 mathematics teachers in 3,761 across 23 countries. (Note that TIMSS is administered to students in both Grades 4 and 8. However, there may be significant differences in accountability instruments between the schools attended by PISA 15-year-olds and those attended by TIMSS Grade 4 students, so I focus on the TIMSS Grade 8 sample. Hence, to analyse TIMSS student outcome data matched with country-level PISA teacher accountability data, I assume that accountability instruments affecting teachers of Grade 8 students in TIMSS are sufficiently similar to the accountability instruments affecting teachers of 15-year-old students in PISA. In PISA 2015, 31.1% of surveyed 15-year-olds were enrolled in Grade 9 and 52.2% were enrolled in Grade 10 (OECD, 2016c). Although in some countries Grades 8 and Grades 9-10 may fall under different levels of schooling, e.g. middle/junior high school vs. high school in the United States, in many countries they fall within the same level, e.g. secondary or lower secondary school. Moreover, neither the TIMSS Grade 8 students nor the PISA 15-year-old students are likely to face the secondary school exit examinations that have substantial accountability implications.)

The final educational dataset is the 2013 wave of the OECD's Teaching and Learning International Survey (TALIS; OECD, 2014c). TALIS is administered to teachers and principals as a set of self-report questionnaires on educational contexts. Although it does not include data on student outcomes, it has data on the other two educational constructs in the conceptual framework, i.e. teacher accountability instruments and teacher motivation. The main TALIS 2013 analyses look at 79,252 teachers in 5,259 schools across 29 countries.

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 values and 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, 2016a; WVS Association, n.d.). Additionally, I use two sociocultural indices 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 in organisational behaviour (see Taras, Kirkman, & Steel, 2010, for a review). Besides sociocultural context, I also include country-level GDP data from the Penn World Table 9.0 (Feenstra et al., 2016).

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 time lags between the sociocultural and educational surveys. For example, most of the Hofstede data were collected over 40 years prior to the educational surveys that I analyse. Although the WVS/EVS data are much more recent, this assumption of sociocultural stability over time comes into play nonetheless, because I match data from each PISA/TIMSS/TALIS wave with sociocultural data from multiple WVS/EVS waves, in order to maximise the number of countries included in each regression. This is possible because many of the questionnaire items across the WVS and EVS have been matched through a shared dictionary (EVS, 2015), and cross-wave data have been analysed together in prior scholarly studies (e.g. Inglehart &

Welzel, 2005; Norris & Inglehart, 2004). In so doing, I assume that these different waves of country-level sociocultural data are comparably accurate despite their different time lags. For example, collection dates for the sociocultural data that I match to PISA 2015 countries range from 2006 (for some WVS 5 countries) to 2014 (for some WVS 6 countries). Even within a single WVS wave, data collection across countries takes place over five years. These sociocultural data will only be reliable proxies if the cultural patterns that they capture are reasonably consistent over time. Hofstede (2001) makes such an assumption, arguing that national cultures are relatively stable, barring external shocks. Inglehart and Welzel (2005), the most prominent WVS/EVS scholars, broadly agree. They further argue that many post-industrial societies have become more culturally similar over time, but that such cultural shifts are subject to strong path dependencies. Empirically, Merritt (2000) attempted to replicate four of Hofstede's scales in a 1993–1997 survey of airline pilots in 19 countries, and found that two of the replicated scales correlated significantly with Hofstede's (including power distance, for which r=0.74), while the others did not (including uncertainty avoidance, for which r=0.31). While my field interviews did not involve adequate sample sizes for such replication, in the secondary statistical analysis I examine the stability of national-level WVS/EVS averages across different waves of the study (see Section 4.2). Additionally, I asked interview participants whether the summary sociocultural data accurately reflect their experience of their respective sociocultural contexts (see Section 6.2).[14]

In matching the country-level sociocultural data to educational data at the student, teacher, and school levels, 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 surveys were administered a few months later than TIMSS 2015 surveys, with some overlap in survey periods (March 2015 to December 2015 for PISA, compared to October 2014 to May 2015 for TIMSS). Thus, for some

[14] I am particularly conscious of the possibility of cultural change in between the Hofstede surveys and the ILSAs because Singapore's uncertainty avoidance score in the Hofstede dataset seems distinctly inaccurate to me, based on my experience of living in Singapore between 2003 and 2006. In the Hofstede dataset, Singapore has the lowest uncertainty avoidance score among all countries, falling over 2.5 standard deviations below the mean. This does not accord with Singapore's notoriously *kiasu* (a Chinese dialect term meaning 'afraid of losing') and *kiasi* ('afraid of death') society. I suspect that the difference is due to the fact that (a) only 58 Singaporeans took part in Hofstede's 1971–73 survey, making it the smallest sample size among the 40 countries that participated in Hofstede's initial 1967–69 and 1971–73 surveys; and (b) these 58 participants were all IBM employees in the early 1970s, at which point IBM Singapore was a relatively small branch of technology firm (with just 100 employees in 1970; IBM, 2019) in the early days of Singapore's industrialisation boom, which suggests that they may have been less risk-averse than their contemporaries. Unfortunately, I have not been able to find comparable recent surveys in Singapore covering constructs like the Hofstede scales that I analyse, so I do not have the means to ascertain the extent of such sample bias and/or change over time.

countries within these datasets, there may be a slight violation of time ordering when I combine PISA accountability data from PISA 2015 with student outcome data from TIMSS 2015. This time-ordering issue, however minor, reinforces the importance of paying attention to sensitivity checks from the other datasets.

As mentioned above, the number of countries in each regression depends on the overlap in country participation between the educational, cultural, and economic datasets in question, and on whether the items composing each sociocultural construct were included in each country's version of the WVS questionnaires. (Specifically, Bahrain, Kuwait, and Qatar did not administer some of the items about civic networks; Bahrain, Egypt, Kuwait, and Qatar did not administer some of the items about confidence in institutions; Qatar did not administer one of the items about civic norms; and New Zealand did not administer one of the items about trust. Consequently, these countries are not included in the main ILSA regressions. Additionally, Colombia, Hong Kong, and Peru did not administer some pertinent items in WVS 4, but subsequently included these items in WVS6. Consequently, they are excluded from the main PISA 2012 regressions, but included in the main PISA 2015 regressions.) Some PISA 2015 regressions have as many as 64 countries, while one set of TIMSS 2015 regressions has just 22. However, besides the TIMSS regressions, the other regressions reported in this thesis are comfortably above the minimum sample size of 25 countries that Bryan and Jenkins (2013) recommend for unbiased estimates of country-level parameters in multilevel models.[15] Also, every regression includes both high- and low-performing education systems, from both the majority and minority worlds. While the TALIS regressions do not include any African countries, the other regressions span all six continents.

Operationalisation

For summary statistics and descriptions of all included variables, see Table 3.4.

[15] However, this is far from being a hard-and-fast rule, as Bryan and Jenkins (2013) note in their review of recommended sample sizes. For example, Maas and Hox (2005) recommend a sample size of at least 50 countries, while Raudenbush and Bryk (2002) suggest that the required number of country-level observations depends on the number of country-level predictors included in the regression model. The possible inadequacy of some of my country-level sample sizes reinforces the importance of the sensitivity checks that I describe below. For an analysis of whether the relatively small sample sizes for TIMSS are likely to result in overfitting and multicollinearity, see Appendix A.

Table 3.4 *Descriptive statistics*

Variable	Description	N	Level	Mean	SD
PISA 2015					
Science (plausible values 1–10)	OECD-generated estimate of science proficiency; using IRT scaling.	476 852	pupils	462.78	105.09
ESCS (economic, social, and cultural status)	OECD-generated score based on pupil-reported parental education, parental occupation, and home possessions; using PCA.	466 458	pupils	-0.28	1.10
School autonomy	OECD-generated score based on the proportion of responsibility for 12 tasks held by the principal, teachers, or school governing body (cf. regional or national actors); principal-reported	15 735	schools	0.68	0.26
Accountability	Latent trait parameter estimate for the intensity of formal teacher accountability in the school, based on 14 principal-reported items; using a 2PL IRT model pooled between PISA 2012 and 2015.	15 509	schools	0.19	0.77
$\overline{\text{Accountability}}$	Country-level weighted mean of Accountability.	72	countries	0.16	0.42
PISA 2012					
Mathematics (plausible values 1–5)	OECD-generated estimate of mathematics proficiency; using IRT scaling.	473 670	pupils	475.06	105.30
ESCS (economic, social, and cultural status)	OECD-generated score based on pupil-reported parental education, parental occupation, and home possessions; using PCA.	466 632	pupils	-0.28	1.12
School autonomy	OECD-generated score based on principal reports of 12 tasks for which the principal, teachers, or school governing body have considerable responsibility; using IRT partial-credit scaling.	17 669	schools	-0.08	1.06
Teacher morale	OECD-generated score based on 4 principal-reported items on teacher morale; using IRT partial-credit scaling. Higher values = better morale.	17 559	schools	-0.00	0.98
Accountability	Latent trait parameter estimate for the intensity of formal teacher accountability in the school, based on 20 principal-reported items; using a 2PL IRT model pooled between PISA 2012 and 2015.	17 628	schools	-0.05	0.86
$\overline{\text{Accountability}}$	Country-level weighted mean of Accountability.	67	countries	-0.08	0.52
TIMSS 2015					
Mathematics (plausible values 1–5)	IEA-generated estimates of mathematics proficiency; using IRT scaling.	257 030	pupils	481.25	109.70
Home educational resources	IEA-generated score (standardised for this analysis) based on pupil-reported parental education, home study supports, and books at home; using IRT partial-credit scaling.	252 187	pupils	10.25	1.82
Teaching experience	Years of teaching experience; teacher-reported; centred at 15 years.	11 270	teachers	0.53	10.73
Job satisfaction	IEA-generated score (standardised for this analysis) based on 7 teacher-reported items on job satisfaction; using IRT partial-credit scaling. Higher values = greater satisfaction.	11 307	teachers	-0.02	0.99
School resources	IEA-generated score (standardised for this analysis) based on 13 principal-reported items on the degree to which school instructional capacity is affected by resource inadequacies (both general and mathematics-specific). Higher values = fewer inadequacies.	7 219	schools	0.03	1.01

Table 3.4 (continued)

Variable	Description	N	Level	Mean	SD
TALIS 2013					
Teaching experience	Years of teaching experience; teacher-reported; centred at 15 years.	101 526	teachers	1.54	10.56
Job satisfaction	OECD-generated score (standardised for this analysis) computed as the mean of two CFA scores, each based on 4 teacher-reported items; for satisfaction with (a) current work environment and (b) the teaching profession, respectively.	105 345	teachers	-0.02	1.00
School autonomy	Proportion of 11 tasks for which the principal, school management team, teachers, and/or school governing body have significant responsibility; principal-reported.	5 891	schools	0.72	0.25
Accountability	Latent trait parameter estimate for the intensity of formal teacher accountability in the school, based on 13 principal-reported items; using an IRT graded-response model.	6 578	schools	-0.03	0.91
$\overline{\text{Accountability}}$	Country-level weighted mean of Accountability.	35	countries	0.04	0.58
World Values Survey /European Values Study					
Confidence in institutions*	Factor score based on 12 items on confidence in various public/private institutions.	99	countries	0.00	0.98
Civic networks*	Factor score based on 7 items on member in various voluntary organisations.	101	countries	0.00	0.99
Civic norms*	Factor score based on 4 items on how justifiable it is to breach civic norms.	103	countries	0.00	0.94
Social trust*	Factor score based on 2 items on how trustworthy most people are.	103	countries	0.00	0.87
Hofstede's Values Survey Module					
Power distance	Hofstede-generated score (standardised for this analysis) based on 3 items on relationships between bosses and subordinates.	104	countries	0.00	1.00
Uncertainty avoidance	Hofstede-generated score (standardised for this analysis) based on 3 items on rule orientedness, desire for job stability, and stress levels.	104	countries	0.00	1.00
Penn World Table					
GDP**	Per capita expenditure-side real GDP at current PPPs; in 2011 US$10,000s, centred at $30,000.	183	countries	-1.01	2.12

Note. IRT=item response theory. 2PL=two parameter logistic. PCA=principal component analysis. CFA=confirmatory factor analysis. Sources: EVS (2011); Feenstra et al. (2016); Hofstede (2001); Inglehart et al. (2014a, 2014b); Martin, Mullis, & Hooper, 2016; OECD (2014b; 2014d; 2016e; 2016f); and my calculations.
*Value used depends on assessment year. Descriptive statistics are shown for values used with 2015 assessment data.
**Descriptive statistics shown are for 2014. Note: mean is greater for countries in the regression samples.

Teacher accountability instruments. To operationalise teacher accountability instruments—defined in this study 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*—I draw on school principals' reports in PISA and TALIS questionnaires of whether particular teacher accountability instruments are present in the school.

I operationalise this by using item-response theory (IRT) modelling to construct scales for the extensiveness of teacher accountability instruments in any given school. For PISA 2012 and 2015, the scale draws on 21 principal-reported questionnaire items. These items fell into four categories: *how teachers are monitored* (4 items, e.g. lesson observations by school leaders); *what quality assurance approaches are used* (7 items, e.g. external evaluation); *how student achievement data are shared* (3 items, e.g. achievement data are provided directly to parents), and *what consequences might result from teacher appraisals* (7 items, PISA 2012 only, e.g. a change in salary). For TALIS 2013, the 13 principal-reported items fell into two categories: *the frequency of formal teacher appraisal by different stakeholders* (5 items, e.g. formal appraisal by the principal) and *the frequency of different consequences of teacher appraisals* (8 items, e.g. a change in the likelihood of career advancement). Further details on the construction of these teacher accountability scales are available in Section 4.1.

In any statistical analysis of questionnaire data, both theoretical and methodological considerations need to be weighed in the decision about whether to represent a construct using raw questionnaire items (e.g. binary indicators for which of several approaches are used to monitor teacher practice), a set of scales (e.g. test-based teacher accountability policies and test-based school accountability as in S. W. Han, 2018), or a single aggregate scale (e.g. a single scale for teacher accountability instruments). Some forms of aggregation are widely accepted as summary representations of a construct, such as test scores that aggregate responses across multiple questionnaire items; or socioeconomic status scales that aggregate responses about different indicators of household privilege, as with the PISA index of student's economic, social, and cultural status (OECD, 2017). For teacher accountability instruments in this statistical analysis, none of the options—raw questionnaire items, a set of scales, nor a single aggregate scale—would be unambiguously advantageous. Having weighed the reasons for and against different measurements, I decided to use a single aggregate scale for teacher accountability instruments in each analysis, for both theoretical and methodological reasons.

One important argument against such an aggregate scale is that aggregation complicates interpretation. This can be due to both theoretical reasons, such as if the aggregated questionnaire items do not share enough similarity to be legitimately conceptualised as indicators of the same construct, and methodological reasons, because modelling assumptions of the aggregation method add a further layer of interpretation to the analysis. In an analysis that aspires to be policy-relevant, such interpretive considerations are especially important. For example, using separate variables for individual teacher accountability instruments or using a set of scales that each represent a category of teacher accountability instruments may be more informative for accountability policy design than a single scale for the overall extensiveness of teacher accountability instruments.

However, in the context of this analysis, it is unlikely that those interpretive advantages of disaggregated items or multiple scales would have been realised. Firstly, using individual questionnaire items would not reflect educational realities because teacher accountability instruments do not usually operate in isolation. Rather, they function as systems (Pritchett, 2015; UNESCO, 2017). For example, merit pay systems typically operate together with performance standards as well as systems for observing lessons and/or tracking test scores. Consequently, I do not focus on the effects of individual accountability instruments, but rather on systems of accountability instruments working in concert.

Secondly, using a set of scales for different types of teacher accountability instruments would require theoretically rigorous and policy relevant reasons for sorting the questionnaire items into their respective categories. My conceptual framework lays out three theoretically delineated categories of teacher accountability instruments based on their underlying mechanisms: communicating information, setting standards, and allocating consequences. However, it is unclear how the accountability instruments in the ILSA survey data would map onto these mechanisms, especially when considering the biases and range of respondents' interpretations within such self-report questionnaire items. For some items and the instruments they represent, this mapping may be relatively straightforward. For example, if a school principal reports that tests or assessment of student achievement are used to monitor teacher practice in their school (OECD 2013a, 2016b), it is relatively safe to assume that if this test-based monitoring influences teacher motivation, it does so through a combination of the informational and standard-setting mechanisms. However, this mapping is less straightforward for some other items/instruments. For example, if a lesson observation by the principal or senior staff (OECD 2013a, 2016b) influences teacher motivation, it likely does so through the informational mechanism; but whether or not it concurrently sets standards for teacher practice could vary depending on whether the observation is based on a rubric, or focuses informally on any elements of teaching mentioned by the observer either before or after the observation, or does not involve any such communication to the teacher. A further issue arising from the nature of the questionnaire data is that these data are based on principals' self-reported responses that may not tally with teachers' experiences: a principal may believe that they have publicly recognised a teacher following an appraisal (OECD, 2013a), but the teacher may not have experienced this recognition as such, and thus this recognition may not have oriented the teacher's motivation toward the goals that the principal regards as desirable, as in the consequence-allocation mechanism. Moreover, as discussed in Chapter 2, within any given category or type of accountability instruments, the context-specific design of an instrument matters tremendously for its efficacy (Pawson & Tilley, 1997; Pritchett, 2017; Williams, 2017). Hence, even if an analysis using multiple scales found that more extensive use of informational teacher accountability instruments (or, to use a category from the original questionnaires, e.g. changes in teachers' responsibilities, careers, or compensation following an appraisal or feedback; OECD, 2013a) was associated with better student outcomes in certain sociocultural contexts, this finding would offer little information to policymakers on the crucial questions of which particular informational instruments should be chosen, and how exactly they should be designed. Indeed, one minor argument against using a set of scales for different categories of accountability instruments it that the statistical analysis would then lend itself to interpretations that have the appearance of adequate nuance while giving a highly reductive picture of the context-sensitive orientation of this research project.

Although a single aggregate scale constitutes an even more reductive representation of teacher accountability instruments, I regard this aggregate scale as the best option for operationalising teacher accountability instruments for several reasons. Firstly, despite reducing teacher accountability instruments to a single dimension of greater or lesser extensiveness, it reflects the fact that accountability instruments operate interactively rather than in isolation. Secondly, it does not require ambiguous decisions about fitting self-report questionnaire items into various categories, and instead can be constructed using principled higher-level decisions about whether each questionnaire item fits within the overall definition of teacher accountability instruments. (For a discussion of how well these items fit empirically in a single scale, see Section 4.1.) Thirdly, although (like the other two measurement options) this aggregate scale cannot yield granular policy implications about which accountability instruments are best for which contexts, it has the potential to make a clearer case against the acontextual 'more accountability and more best practices = better student outcomes' arguments that lurk in the background of research on accountability in high-performing education systems, as outlined in Section 1.1. That is, if my operationalisation of teacher accountability instruments adopts the reductive more/less logic of such arguments, and if my analysis uses the same ILSA datasets that are typically used to identify 'best practices' from high-performing countries, and if this analysis can demonstrate the context-dependence of even this reductive aggregate measure of teacher accountability instruments, then the analysis would lack policy relevance at an instrument-specific level, but it would gain policy relevance in making a case against a prominent argument in education policy debates. Finally, an additional methodological benefit is that these single aggregate measures eliminate the multicollinearity that would result from including multiple, correlated variables for accountability in the same regression model. They also aid model convergence when looking at country-level weighted means for teacher accountability, given the limited number of countries available in the data. These scales offer an overall snapshot of teacher accountability instruments, facilitating cross-country comparison.

Student outcomes. In both PISA and TIMSS, student proficiency scores are generated by their respective administering institutions (the OECD and the IEA) using item-response theory (IRT) modelling. PISA and TIMSS assess multiple subjects, but individual students' proficiency scores are highly correlated across subjects. Accordingly, for the sake of simplicity I focus on one subject per assessment wave. PISA includes questions on reading, mathematics, and science, but every wave emphasises one of the three subjects, assessing it 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 model 1 (outlined below and reported in Section 4.3) using each of the other PISA and TIMSS subjects as outcome variables. Across the subjects, there were no differences in the direction or significance of key variables, thus justifying my decision to focus on one subject per assessment cycle.

Teacher motivation. To proxy for teacher motivation, I use three scales that were constructed by the respective survey administrators and are included in the public-access datasets. Two of the datasets offered teacher-reported scales for job satisfaction. The TALIS 2013 job satisfaction scale was computed as the mean of two confirmatory factor-analysis scores, each based on four questionnaire items, for satisfaction with (a) the current work environment (e.g. 'I would recommend my school as a good place to work') and (b) the teaching profession (e.g. 'If I could decide again, I would still choose to work as a teacher'), respectively (OECD, 2014d, pp. 204–14). The TIMSS 2015 job satisfaction scale was generated using IRT partial-credit scaling, based on seven questionnaire items (e.g. 'I am proud of the work that I do'; Martin, Mullis, & Hooper, 2016, pp. 15.298–15.302). While there is thematic overlap between the TALIS and TIMSS items for teacher job satisfaction, none of the items are identical. Although the PISA datasets do not include teacher-level questionnaires,[16] the PISA 2012 principal questionnaires did include items pertaining to school-level teacher morale. This teacher morale scale was generated using IRT partial-credit scaling of five principal-reported items (e.g. 'Teachers work with enthusiasm'; OECD, 2014b, p. 349). The differences between these three scales, as well as how valid they may be as proxies for teacher motivation, are discussed in Section 5.1.

National sociocultural context. Given the multidimensionality of culture and society, I draw on 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, & Han, 2009; Inglehart & Welzel, 2005; Markus & Conner, 2013; Thompson, Ellis, & Wildavsky, 1990), I focus on these constructs because they represent cultural patterns that are theoretically expected to moderate the effects of accountability instruments, as discussed in Section 2.3.

First, from the WVS/EVS datasets, I use factor analysis to construct four scale variables for aspects of social capital. (Inglehart and Welzel [2005] also use factor analysis to construct their WVS scales for traditional vs. secular-rational values and for survival vs. self-expression values. However, while Inglehart and Welzel chose items for inclusion in their factor variables based on how much cross-national variation the items collectively explained, I have chosen items for inclusion based on their conceptual relevance.) As noted in Section 2.3, there are both theoretical and empirical reasons for expecting social capital to moderate the effects of teacher accountability instruments on teachers and students. For this statistical analysis, 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); membership in *civic networks* (7 items, e.g. religious organisations); how justifiable it is to breach *civic norms* (4 items, reverse coded, e.g. cheating on taxes); and *social trust* (2 items: whether most people can be trusted; to what degree most people would try to take advantage of you). Further details on the construction of these social capital scales are available in Section 4.2.

[16] PISA 2015 included an optional teacher questionnaire, but only 19 countries administered it.

Alongside these WVS/EVS factor variables, I include two pre-existing sociocultural scales from Hofstede's IBM dataset. The first scale, power distance, is a measure of hierarchy. As discussed in Section 2.3, power distance measures the acceptance of hierarchical distributions of power. I also use Hofstede's index of uncertainty avoidance, which indicates a tendency toward anxiety and preferences for stability, even if that entails organisational rigidity (Hofstede, 2001). Hofstede calculates the power distance and uncertainty avoidance indices through linear combinations of country-level average responses to the pertinent questionnaire items (Hofstede, 2001, pp. 86, 150). For this analysis, I standardise power distance and uncertainty avoidance scores to match the mean and spread of the social capital factor scores.

Control variables. In every regression model, I include a control variable at each level of analysis. 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. At the teacher level, in TALIS and TIMSS regressions, I control for years of teaching experience. At the school level, for PISA and TALIS I control for the degree of school autonomy in decision-making. TIMSS 2015 did not include questionnaire items on decision-making autonomy, so instead I use a pre-existing TIMSS scale for the degree to which instructional capacity is constrained by inadequate resources (whether general resources and mathematics-specific ones), as reported by the principal. Finally, at the country level, I control for per-capita GDP. 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.

This is a relatively parsimonious set of control variables, because my aim is not to capture as much variability in student achievement (or teacher motivation) as possible; but rather to determine whether the proposed conceptual framework holds empirically. Accordingly, I focus on the relationships mapped out in the framework—between teacher accountability, teacher motivation, and student outcomes—and the extent to which these relationships are affected by measurable aspects of context. Control variables are included only if there are theoretical or empirical grounds for expecting them to affect the relationship between teacher accountability, teacher motivation, sociocultural context, and student outcomes.

Specifically, I control for pupil socioeconomic status, since the effects of teacher accountability instruments may depend on how privileged pupils are. For example, since less privileged students tend to be concentrated in lower-performing schools, Diamond and Spillane (2004) argue that the perverse effects of high-stakes accountability instruments, such as curriculum narrowing and diverting resources to students at the pass/fail margin, may disproportionately affect less privileged students. At the teacher level, I control for years of teaching experience. Since teaching experience has been associated with different levels of both student outcomes (Hanushek & Rivkin, 2006; Murnane & Phillips, 1981) and of motivational factors such as teacher self-efficacy and job satisfaction (Chiong, Menzies, & Parameshwaran, 2017; Klassen & Chiu, 2010; Liu & Ramsey, 2008; Wolters & Daugherty, 2007), it may affect the mediation relationship between teacher accountability, teacher motivation, and student outcomes.

Beyond these individual-level factors, I also control for school autonomy, since teacher accountability instruments may be more effective when schools have more freedom to change their practices in response to accountability incentives (see Woessmann, 2016 for a related empirical analysis). As mentioned above, the TIMSS 2015 questionnaires did not include items related to school autonomy. Instead, I use a principal-reported measure of the extent to which school instructional capacity is constrained by resource shortages. While resource constraints differ considerably from decision-making constraints, both affect schools' capacity for autonomous action. Thus, the measure of school resource shortages may not be a close equivalent for school autonomy, but it is the best available substitute. Furthermore, Kim, Sun, and Youngs (2019) found that teachers were more likely to perceive a teacher evaluation programme as legitimate if they had adequate time and resources to complete the evaluation process; thus lending support to the inclusion of school resources as a control in its own right.

Finally, I control for national GDP per capita. Although I do not subscribe to modernisation theories that associate developed countries with 'modern' values and developing countries with 'traditional' values (e.g. Inglehart & Welzel, 2005), it is empirically true in my dataset that GDP is moderately correlated with some of the sociocultural constructs (see Table A.2 in Appendix A). 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.

In interpreting parameter estimates from these control variables, it is important to bear in mind that the pupil- and school-level measures differ across the datasets. As noted above and shown in Table 3.4, the PISA and TIMSS proxies for student socioeconomic background use different questionnaire items and scaling procedures. For the school-level controls, the TIMSS variable concerns school resources whereas the PISA and TALIS variables concern school autonomy. Additionally, the three PISA and TALIS school autonomy scales all differ from each other, even though all three are based on principal-reported items. Although the PISA 2012 and 2015 scales were based on identical questionnaire items, the 2012 scale was constructed using IRT partial-credit scaling, whereas the 2015 scale was constructed using an approach that weighted each item equally [OECD 2014b, pp. 312–313, 346–347; 2016e, pp. 243–244]. Although the TALIS 2013 school questionnaire included a set of 11 items that were similar to the 12 PISA school autonomy items, the dataset did not have a pre-existing school autonomy scale. Instead, I used these TALIS school autonomy items (i.e. the TC2G18A to TC2G18K items) to construct a simple scale based on the proportion of the 11 items for which at least one set of school-level actors had significant responsibility.

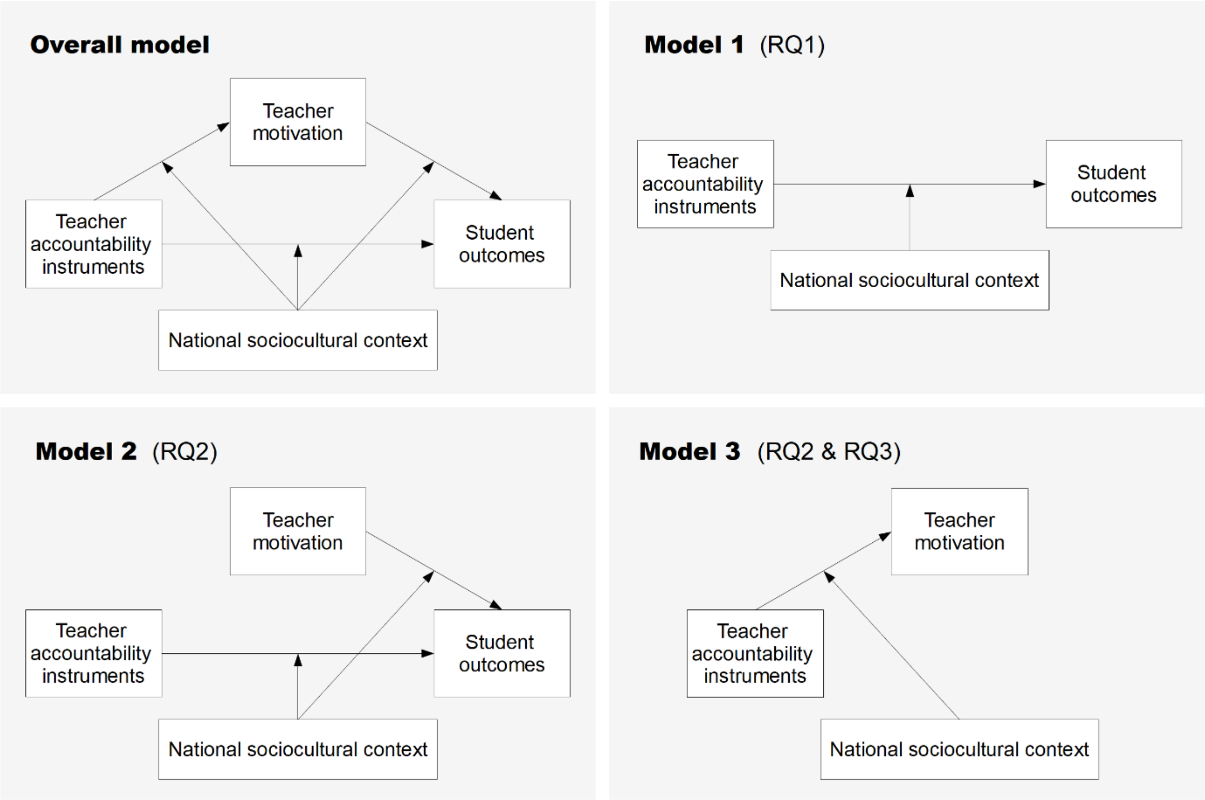
Modelling

This statistical analysis is framed as a mediated moderation, as shown in the first panel of Figure 3.1. That is, I posit firstly that the influence of teacher accountability instruments on student outcomes is transmitted indirectly via their influence on teacher motivation, which in turn influences student outcomes; thus, teacher motivation *mediates* the relationship between accountability instruments and student outcomes. Secondly, I posit that the relationship between teacher accountability instruments and student outcomes (as well as the relationships within the mediation pathway) can either intensified or weakened by sociocultural context; thus, sociocultural context *moderates* the relationship between accountability instruments and student outcomes.[17]

[17] It is also possible that teacher motivation may itself moderate—that is, intensify or weaken—the relationship between accountability instruments and student outcomes. For example, accountability instruments may have stronger effects on student outcomes when teachers are more motivated to begin with, and weaker effects when teachers are less motivated. In fact, some of the teachers whom I interviewed in Finland and Singapore observed that demotivated teachers may pay little attention to accountability instruments, unlike their more motivated colleagues. I tested the possibility of teacher motivation moderating the effects of accountability instruments, by adding an interaction between accountability and motivation to the main TIMSS 2015 and PISA 2012 models. In both cases, the coefficients on these interaction terms were insignificant. This may well be due to measurement issues with teacher motivation, which I address in Sections 3.6 and 5.1. Nonetheless, such differential effects of accountability instruments contingent on teachers' pre-existing motivation levels are not the primary focus of this study. Rather, I am mainly interested in differential effects of accountability instruments contingent on variation in national-level sociocultural context. Accordingly, the results presented in this thesis do not include any accountability*motivation interaction terms.

To test these relationships, I estimate three different sets of models, which correspond to different research questions, as shown in Figure 3.1. First, for RQ1, on *the extent to which the relationship between teacher accountability instruments and student outcomes depends on sociocultural context,* I estimate a moderation model for the effects of teacher accountability instruments on student outcomes, moderated by the six sociocultural constructs (model 1). Next, to test the mediation pathway in RQ2, on *the extent to which teacher motivation mediates the relationship between teacher accountability instruments and student outcomes*, I estimate two different models, in which the outcome variables are, respectively, student outcomes (model 2) and teacher motivation (model 3). To address RQ3, on *the extent to which the relationship between teacher accountability instruments and teacher motivation depends on sociocultural context,* I run another set of model 3 estimations. Whereas for RQ2 I interpret the model 3 results with an eye to the mediation relationship, for RQ3 I focus on the degree to which each of the six sociocultural constructs moderates the effect of teacher accountability on teacher motivation.

Figure 3.1 *Conceptual diagrams for the overall analysis and each statistical model*


Overall model
Teacher motivation
Teacher accountability instruments
Student outcomes
National sociocultural context
Model 1 (RQ1)
Teacher accountability instruments
Student outcomes
National sociocultural context
Model 2 (RQ2)
Teacher motivation
Teacher accountability instruments
Student outcomes
National sociocultural context
Model 3 (RQ2 & RQ3)
Teacher motivation
Teacher accountability instruments
National sociocultural context

In all of these regressions, I use statistical techniques that are appropriate for the sampling and assessment designs of the educational surveys (Jerrim, Lopez-Agudo, Marcenaro-Gutierrez, & Shure, 2017; Martin, Mullis, & Hooper, 2016; OECD, 2014b, 2014d, 2017; Rutkowski, Gonzalez, Joncas, & von Davier, 2010). PISA, TIMSS, and TALIS all use stratified random sampling such that students and teachers are clustered in classes and/or schools. The multilevel models account for this clustering and avoid the inaccurately low standard errors that would result from straightforward single-level analyses. Another feature of the sampling design is that some countries choose to oversample certain strata (e.g. certain regions or certain school types) to obtain higher-resolution data on these subpopulations. To simulate nationally representative populations, each observation is weighted with school-level weights and conditional student- and teacher-level weights, as provided by dataset administrators and standardised by MLwiN (Gebhardt, 2009; Laukaityte & Wiberg, 2017a; Pillinger, 2011). Next, because PISA and TIMSS student proficiency scores are estimated using IRT, the datasets include several plausible values (ten plausible values in PISA 2015, and five in TIMSS 2015 and PISA 2012) for each student proficiency score, in order to capture the measurement error associated with the estimation process. Accordingly, I re-estimate every regression for each plausible value, and then combine the coefficient and standard error estimates using Rubin's rules (Laukaityte & Wiberg, 2017b; OECD, 2009; Rubin, 1996).

All regressions are estimated using the iterated generalised least squares procedure, a form of full-information maximum likelihood estimation, in MLwiN 3.0.2. The number of levels in each model depends on the sampling design of the educational dataset in question: PISA models in which the outcome variable is student proficiency have three levels (pupil, school, country) and those in which the outcome is teacher motivation have two levels (school, country); TIMSS models have four levels for student outcomes (pupil, teacher, school, country) and three for teacher motivation (teacher, school, country); and TALIS models for teacher motivation have three levels (teacher, school, country). I preserve the levels of analyses at which the survey data were collected so that the calculated standard errors will account for any non-random similarity within sampled clusters at each level. All regressions use sandwich estimators for standard errors to mitigate the effects of potential heteroskedasticity.

Since I am interested in associations with teacher accountability instruments both between countries and within countries, the 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 weighted mean. This is sometimes called a within-between model (e.g. Bell & Jones, 2015), since country-level mean measures variation between countries, whereas the school-level differential measures variation within countries. As

Snijders and Bosker (2011) show, this within-between model is statistically equivalent to the Mundlak model, which includes the higher-level group mean alongside the original lower-level predictor (rather than a group-mean-centred differential for the lower-level predictor, as in the within-between model). Despite this statistical equivalence, these two models differ in how the parameter estimates for the group mean variable are interpreted. In the Mundlak model, the coefficient on the group mean variable gives the 'contextual' effect, i.e. the expected difference between two lower-level units with the same value of the lower-level predictor but belonging to higher-level groups that differ by 1 in their group means for the predictor (in this case, two schools with the same teacher accountability scale scores, in countries with different mean teacher accountability scores). In the within-between model, the coefficient on the group mean variable gives the 'between' effect, i.e. the expected difference between the mean responses of two groups that differ by 1 in their group means for the predictor (Bell & Jones, 2015). Since I am more interested in between-country differences—especially when considering the interaction between country-level accountability and sociocultural context—rather than in contextual effects, I estimate the within-between model rather than the Mundlak model. A further benefit of the within-between model is reduced multicollinearity due to group-mean-centring of the lower-level predictor.

Where possible, I estimate two sets of regressions for each dataset: (a) a regression that includes all six sociocultural constructs and their associated interaction terms; and (b) six separate regressions that each include just one sociocultural construct and its associated interaction term. While (a) is analytically preferable, because it accounts for the interplay between the sociocultural constructs, it is not always empirically feasible. Specifically, due to the small number of countries in the TIMSS and TALIS datasets, regressions following option (a) showed indications of multicollinearity. Since such indications did not appear in TIMSS and TALIS regressions following option (b), it is likely that the multicollinearity in (a) is due to overfitting, 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 overfitted. (Additionally, when (a) appears statistically sound, I treat the (b) regressions as sensitivity checks.) In both cases, I present results for a wide range of regressions, both from the main dataset and from sensitivity checks, to convey a clear picture of how robust the results may be across all six sociocultural constructs. For evidence of overfitting in (a) and a lack thereof in (b) for the TIMSS 2015 main dataset, see Appendix A, where I present results for both approaches.

For each research question, I focus on the educational dataset offering the most granular information on key variables, as shown above in Table 3.2. For RQ1, which looks at the relationship between teacher accountability, student outcomes, and sociocultural context, my main dataset is PISA 2015. This is because the PISA 2015 dataset covers not only student-level proficiency scores, but also school-level questionnaire items on teacher accountability. Thus, for model 1 with PISA 2015 data, I estimate:

$$\begin{aligned}
&\text{Proficiency}_{psc} = \beta_0 + \beta_1\text{ESCS}_{psc} + \beta_2\text{Autonomy}_{sc} + \beta_3\text{GDP}_c \\
&\quad + \beta_4\text{AccountabilityDiff}_{sc} + \beta_5\overline{\text{Accountability}}_c \\
&\quad + \beta_6\text{AccountabilityDiff}_{sc}\text{*ESCS}_{psc} + \beta_7\text{AccountabilityDiff}_{sc}\text{*Autonomy}_{sc} + \beta_8\text{AccountabilityDiff}_{sc}\text{*GDP}_c \\
&\quad + \beta_9\overline{\text{Accountability}}_c\text{*ESCS}_{psc} + \beta_{10}\overline{\text{Accountability}}_c\text{*Autonomy}_{sc} + \beta_{11}\overline{\text{Accountability}}_c\text{ *GDP}_c \\
&\quad + \beta_i\text{Sociocultural}_c + \beta_j\text{AccountabilityDiff}_{sc}\text{*Sociocultural}_c + \beta_k\overline{\text{Accountability}}_c\text{ *Sociocultural}_c \\
&\quad + v_c + u_{sc} + e_{psc} \qquad \textit{Equation (1)}
\end{aligned}$$

where Proficiency_{psc} is the science 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 $\overline{\text{Accountability}}_c$, the country-level weighted mean of teacher accountability, and $\text{AccountabilityDiff}_{sc}$, the school-level teacher accountability differential. 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, to determine the extent to which the effects of school- and country-level teacher accountability are moderated by sociocultural context as well as each control variable. 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.

For RQ2, which focuses on the relationship between teacher accountability instruments, teacher motivation, and student outcomes, the main dataset is TIMSS 2015. Since TIMSS 2015 included teacher questionnaires for all participating countries, this dataset has a measure of teacher motivation self-reported by teachers, which is likely to be a better proxy than the PISA 2012 principal-reported measures of school-level teacher morale. For model 2 with TIMSS 2015 data, I estimate:

$$\begin{aligned}\text{Proficiency}_{ptsc} = {} & \beta_0 + \beta_1\text{HomeResources}_{ptsc} + \beta_2\text{Experience}_{tsc} + \beta_3\text{SchoolResources}_{sc} + \beta_4\text{GDP}_c \\ & + \beta_5\overline{\text{Accountability}}_c + \beta_6\text{Motivation}_{tsc} \\ & + \beta_7\overline{\text{Accountability}}_c{*}\text{HomeResources}_{ptsc} + \beta_8\overline{\text{Accountability}}_c{*}\text{Experience}_{tsc} \\ & + \beta_9\overline{\text{Accountability}}_c{*}\text{SchoolResources}_{sc} + \beta_{10}\overline{\text{Accountability}}_c{*}\text{GDP}_c \\ & + \beta_i\text{Sociocultural}_c + \beta_j\overline{\text{Accountability}}_c{*}\text{Sociocultural}_c \\ & + \beta_k\text{Motivation}_{tsc}{*}\text{Sociocultural}_c \\ & + w_c + v_{sc} + u_{tsc} + e_{ptsc} \qquad \textit{Equation (2)}\end{aligned}$$

where $\text{Proficiency}_{ptsc}$ is the mathematics proficiency score of pupil *p* taught by teacher *t* in school *s* in country *c*. Control variables comprise pupil home educational resources ($\text{HomeResources}_{ptsc}$), teachers' years of experience (Experience_{tsc}), school resource levels ($\text{SchoolResources}_{sc}$), and national per-capita GDP (GDP_c). Equation (2) is similar to (1), except for (a) the addition of teacher-reported job satisfaction (Motivation_{tsc}) and its interaction with $\overline{\text{Accountability}}_c$, to test for mediation; and (b) the absence of a school-level teacher accountability differential, since TIMSS questionnaires lacked sufficient teacher accountability items, so the teacher accountability data here are taken from PISA and matched at the country level. Also, given the risk of overfitting described above, Sociocultural_c here represents each sociocultural construct entered singly, in turn (rather than a vector of all six constructs concurrently). Finally, w_c, v_{sc}, u_{tsc}, and e_{ptsc} are error terms at each level.

Testing for a mediation relationship in RQ2 also involves running model 3 on the TIMSS data to determine whether accountability instruments are associated with teacher motivation. However, for reasons I explain in Section 5.1, my statistical analysis presentation for RQ2 pays more attention to model 2. In turn, model 3 takes the spotlight in the RQ3 analysis. For RQ3, which examines the influence of sociocultural context on the relationship between teacher accountability instruments on teacher motivation, the main dataset is TALIS 2013. TALIS 2013 collected data on both of the educational variables in RQ3, at the lowest possible levels of analysis: teacher motivation as reported by teachers, and teacher accountability instruments as reported by principals. For model 3 with TALIS 2013 data, I estimate:

$$\begin{aligned}
&\text{Motivation}_{tsc} = \beta_0 + \beta_1\text{Experience}_{tsc} + \beta_2\text{Autonomy}_{sc} + \beta_3\text{GDP}_c \\
&\quad + \beta_4\text{AccountabilityDiff}_{sc} + \beta_5\overline{\text{Accountability}}_c \\
&\quad + \beta_6\text{AccountabilityDiff}_{sc}{*}\text{Experience}_{tsc} + \beta_7\text{AccountabilityDiff}_{sc}{*}\text{Autonomy}_{sc} + \beta_8\text{AccountabilityDiff}_{sc}{*}\text{GDP}_c \\
&\quad + \beta_9\overline{\text{Accountability}}_c{*}\text{Experience}_{tsc} + \beta_{10}\overline{\text{Accountability}}_c\,{*}\text{Autonomy}_{sc} + \beta_{11}\overline{\text{Accountability}}_c\,{*}\text{GDP}_c \\
&\quad + \beta_i\text{Sociocultural}_c + \beta_j\text{AccountabilityDiff}_{sc}{*}\text{Sociocultural}_c + \beta_k\overline{\text{Accountability}}_c{*}\text{Sociocultural}_c \\
&\quad + v_c + u_{sc} + e_{tsc} \qquad \textit{Equation (3)}
\end{aligned}$$

where Motivation_{tsc} is the job satisfaction of teacher *t* in school *s* in country *c*. Equation (3) is similar in form to equation (1), with the pupil proficiency score being replaced by teacher job satisfaction (Motivation_{tsc}), and pupil socioeconomic status being replaced by years of teaching experience (Experience_{tsc}). Also, given the relatively small sample of countries, Sociocultural_c represents each sociocultural construct entering the model singly, rather than a vector of all six constructs, as with the TIMSS regressions under equation (2). The other difference from equation (1) is that the lowest-level error term here is e_{tsc}, representing teachers nested within schools nested within countries, as compared to pupils in equation (1).

Sensitivity checks

For each research question, I run a number of sensitivity checks alongside regressions for the main dataset, as indicated above in Table 3.2. For RQ1, for which the main statistical analysis estimates model 1 using PISA 2015 data, I also estimate model 1 using data from PISA 2012 and TIMSS 2015. I use the PISA 2012 data in two ways: simply re-estimating model 1 using PISA 2012 data, to check the robustness of the model 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. It is important to check whether the results are robust with and without time lags between the measurement of teacher accountability and the measurement of student outcomes, because education policy implementation faces time lags in adoption and impact, with different factors operating at different phases of policy development (de Lancer Julnes & Holzer, 2001; Podgursky & Springer, 2007). For example, a principal's answers to the PISA 2015 school questionnaire may reflect recent changes in national policy or local initiatives, but these changes might not yet be affecting student performance in the accompanying science, mathematics, and reading tests. I also estimate model 1 for two separate TIMSS datasets, one matched with country-level accountability data from PISA 2015, and another with accountability data from PISA 2012.

Additionally, I analyse subsamples of the PISA 2015 and 2012 datasets containing observations only from OECD countries, as well as a subset of PISA 2015 data containing observations only from publicly funded schools.

For RQ2, to supplement the main analysis with models 2 and 3 using TIMSS 2015 data matched with country-level accountability data from PISA 2015, I run a sensitivity check with TIMSS 2015 data matched with accountability data from PISA 2012. Additionally, I tested the models using PISA 2012 data, both the full sample and an OECD-only subsample. PISA 2015 did not include a suitable proxy for teacher motivation, so it was not included in the sensitivity checks. For RQ3, in addition to the main TALIS 2013 dataset, I ran sensitivity checks using the two TIMSS 2015 datasets and the two PISA 2012 datasets used in RQ2.

In addition, as mentioned above, whenever the dataset included an adequate number of countries, I estimated models that included all six sociocultural constructs simultaneously, as well as models with each sociocultural construct singly, in turn. For the model 1 regressions with single sociocultural constructs, I also re-estimated the regressions using a separate cut of the data for each construct, such that countries that were excluded from the main dataset because they had participated in either WVS/EVS or Hofstede but not both (or because they had administered WVS/EVS items for some but not all of the social capital-related scales) would now be included in these separate cuts for the applicable sociocultural constructs, thus increasing the country sample sizes.

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 PISA 2015 student achievement and Mexico in TALIS 2013 teacher job satisfaction, 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.

3.4 Teacher interviews in Finland and Singapore

In addition to large-scale secondary datasets, I draw on semi-structured individual interviews with 12 teachers from secondary schools in Singapore (for pupils from ages 13 to 16/17), and 12 teachers from lower secondary schools in Finland (for ages 13 to 15). The level of schooling was chosen to match the level of students participating in the PISA (age 15) and TIMSS (Grade 8) datasets used in the statistical analysis. Additionally, Singapore's secondary schools and Finland's lower secondary schools are, in their respective countries, the stage of schooling immediately preceding specialisation into academic and vocational pre-tertiary institutions. I conducted the interviews in Singapore during July 2018 and in Finland during September 2018. (I also conducted an additional interview for the Singapore sample via video conference from Cambridge in August 2018.)

In this project, the teacher interviews serve two main functions. Firstly, the interviews add country-specific granularity to the standardised statistics of the cross-country educational and cultural surveys. For example, while PISA 2012 and TALIS 2013 include data on whether teacher appraisals can result in a financial bonus or a change in salary, this conceals vast differences in the magnitude of reward and the pervasiveness of the reward schemes, as observed in Section 5.2. Secondly, the interviews offer evidence that could support or challenge the causal pathway from accountability instruments to teacher motivation.[18] Given that teachers are the subjects—and, in some cases, implementors—of teacher accountability instruments, they have direct knowledge of the mechanisms underlying these instruments (Pawson & Tilley, 1997, Chapter 6). The usual limits of human awareness and memory apply, as well as the various inhibitions of interviewer-interviewee relationships, but I approach the interviews with the assumption that teachers are reflective practitioners who are capable of articulating meaningful observations about their experiences and contexts. This assumption is especially reasonable in the countries of interest, given that Finnish teachers are trained to conduct research and apply it to their practice (Sahlberg, 2015a), and Singaporean teachers participate in professional leaning programmes throughout their careers (Goh & Lee, 2008; Low & Tan, 2017).

Given the brand of realism that I adopt, the aim of these interviews was to refine my conceptualisation of teacher accountability instruments (Pawson & Tilley, 1997), as noted in Section 2.1. Accordingly, interview questions and procedures were framed around my conceptual framework and a working hypothesis—not in the sense of a falsifiable hypothesis to be rejected or temporarily retained based on the data; but rather an articulation of my current best guess about the relationship in question (Maxwell, 2013). Below, I explain how this theory-driven stance informed my fieldwork.

[18] As discussed above in Section 2.1, I use the term 'causal' in a realist sense, which is concerned with identifying mechanisms that actually generate a pattern of change—not a statistical or experimental sense, which is concerned with identifying and isolating all sources of variation that may bias the relationship between treatment, comparison, and outcomes.

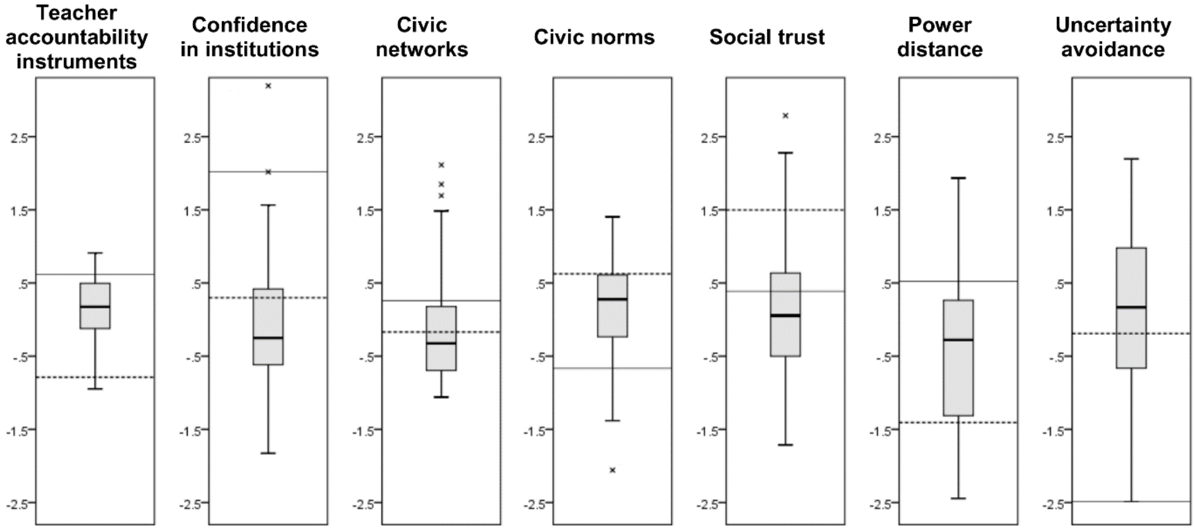
Country selection: Why Finland and Singapore?

Finland and Singapore are appropriate cases to compare because they (a) both have highly successful school systems, but with (b) contrasting approaches to teacher accountability, embedded within (c) different sociocultural contexts, despite (d) other similarities that facilitate comparison both analytically and logistically. I discuss each of these in turn.

In terms of educational efficacy, both countries have consistently shown outstanding performance in international student assessments. In recent international comparisons, Finland's scores have declined somewhat (alongside other concerns about declining educational attainment in Finland; Teivainen, 2019), whereas Singapore has been on an upward trajectory that has propelled it to the top of the tables across PISA and TIMSS subjects (Martin, Mullis, Foy, & Hooper, 2016b; Martin, Mullis, Foy, et al., 2016a; OECD, 2016d). Still, both education systems continue to receive international adulation, and justifiably so.

However, they have disparate approaches to teacher accountability. In Singapore, teachers' work is managed within the Enhanced Performance Management System (EPMS), a national system of tiered performance standards and formal appraisals, with a structured career ladder and sizeable bonuses (Kan, 2014; Sclafani & Lim, 2008). In contrast, Finland's teachers are not subject to formal appraisals. Teachers do not receive promotions (principalship is considered separate profession, requiring distinct qualifications), and formal rewards and punishments are minimal (Finnish National Board of Education, 2013; Sahlberg, 2015a). These differences are reflected in cross-country surveys. As shown in Figure 3.2, Singapore and Finland fall respectively in the highest and lowest quartiles of the teacher accountability scale that I derived from the PISA 2015 school questionnaires. (The same is true of these countries for the teacher accountability scales from PISA 2012 and TALIS 2013. See Section 4.1 for details of the accountability scale construction.)

Figure 3.2 *Box plots of country-level average scores for teacher accountability instruments and sociocultural context among PISA 2015 countries, including Finland (dotted line) and Singapore (solid line)*



Note. Sources: (a) teacher accountability instruments: PISA 2015, my calculations; (b) confidence in institutions, civic networks, civic norms, social trust: World Values Survey/European Values study, my calculations, (c) power distance, uncertainty avoidance: Hofstede, standardised.

Additionally, they differ considerably in sociocultural context. Again referring to Figure 3.2, Finland and Singapore are in opposite extreme quartiles for civic norms and power distance, meaning that Singaporeans are much more likely than Finns to accept hierarchical distributions of power, and to justify antisocial behaviour such as cheating on taxes or avoiding a fare on public transport. For the other four sociocultural constructs, both countries fall on the same side of the median, but never in the same quartile. Finns have unusually high levels of social trust, while Singapore is an outlier in its levels of confidence in institutions. The only sociocultural measure in which they do not differ substantially is civic networks, i.e. how much residents participate in local organisations (whether political, religious, professional, or leisure-related). For further illustrative data, Finland and Singapore fall at opposite ends of Oxfam's 2018 Commitment to Reducing Inequality Index, ranking 3rd and 149th, respectively, out of 157 countries (Lawson & Martin, 2018).

Despite these different accountability approaches and sociocultural contexts, Finland and Singapore are also similar in ways that facilitate comparison between them. Educationally, both countries had relatively low levels of educational attainment in the mid-20th century, but implemented pivotal education policy reforms in the 1970s and 1980s (Goh & Gopinathan, 2008; Sahlberg, 2015a). Moreover, both countries have long regarded education as pivotal to both economic development (S. K. Lee, Goh, Fredriksen, & Tan, 2008; Tirri, 2014) and the construction of national identity (Gopinathan, 2015; Heller-Sahlgren, 2015). Today, both countries are lionised as educational 'reference societies' (de Roock & Espeña, 2018; Takayama et al., 2013), and both governments have actively cultivated educational 'brands' to encourage profitable international consumption of their educational services (Candido & Eriksson, 2019; K. P. Tan, 2018). However, despite this global limelight, there has been relatively little research on the apparently effective processes of teaching and learning in Finland and Singapore (Dimmock & Tan, 2013; Simola, Kauko, Varjo, Kalalahti, & Sahlstrom, 2017). To illustrate, literature searches on 7 June 2019 in ERIC and Scopus for the terms 'Singapore' and 'Enhanced Performance Management System' yielded only two results (Kaur, 2010; Liew, 2012), despite this system being much lauded and used for as a template for teacher performance management elsewhere (e.g. in Scotland; Hepburn, 2017).

Further similarities between these countries also facilitate comparative study. Both countries have high levels of development and populations of approximately 5.5 million (United Nations, 2017), with ethnic diversity but a numerically dominant ethnolinguistic group. Logistically, although Finland has a far larger land area than Singapore, both countries have well-organised public transport systems. Moreover, most adults in both countries are fluent in English, so I could conduct interviews without translation. In Singapore, my research was further aided by the fact that I attended Singapore secondary schools for four years and still maintain friendships there; and my 2013 master's thesis was a comparative analysis of political developments in Singapore and Malaysia. Although I initially had far less familiarity with the Finnish context, I attended an education policy conference in Helsinki in May 2018, which allowed me to orient myself in the country prior to fieldwork that September. During my fieldwork there, I was affiliated with the University of Tampere's Faculty of Education as a visiting researcher, which allowed me to clarify my preliminary observations through discussions with local academics.

Sample design and recruitment

I set out to interview 10 to 15 teachers in each country, in order to represent a range of salient background characteristics among the interview participants, within a manageable amount of qualitative data for a PhD thesis that also analyses secondary quantitative datasets. My goal in seeking diverse interview participants was not to identify systematic differences in teachers' responses to accountability instruments—indeed, the samples were far too small to do so rigorously. Rather, I simply intended to test and refine my working theory of teacher accountability instruments with as wide a range of teachers as possible. In both countries, the interviews reached saturation (Bleich & Pekkanen, 2013), with additional participants adding specific details based on their respective experiences, but broadly echoing what prior participants has said about teacher accountability instruments, sociocultural context, and motivational processes.

In both countries, I aimed to speak with teachers across a range of subjects, years of teaching experience, management roles, and school characteristics; and of personal characteristics such as ethnolinguistic background and gender. Some of these categories differed between the countries. For example, the main ethnolinguistic classification in Finland is between the Finnish-speaking majority and the Swedish-speaking minority, whereas in Singapore the distinction is between the Chinese majority and the non-Chinese minorities.[19] Given the emphasis on national sociocultural context, I also aimed to interview at least one teacher in each country who had grown up in another sociocultural context, since their perspectives on the accountability system may differ from those who had been formatively socialised into the wider social system.

Educationally, in Singapore I wanted to interview teachers at different levels of the career ladder: 'regular' teachers (who are formally known as general education officers), as well as management-level teachers (known as senior education officers) who were responsible for appraising their colleagues. I also aimed for good representation across 'mainstream' schools and the more privileged autonomous and independent schools, which not only have more administrative leeway (which may affect teacher accountability) but also tend to have higher test scores and greater prestige (J. Tan & Gopinathan, 2000). In Finland, I aimed to interview teachers from both Finnish- and Swedish-medium schools, given the historic cultural differences between these schools (Heller-Sahlgren, 2015); as well teachers from urban, semi-urban, and rural schools, given the different working conditions across school localities (Kalaoja & Pietarinen, 2009). I also hoped to speak with teachers from different municipalities, because educational resources and governance can differ considerably across municipal administrations (Public School Insights, 2008; Simola et al., 2017). Additionally, I wanted to interview a few teachers from private schools that offered the Finnish national curriculum, as a parallel to Singapore's autonomous and

[19] Of course, these binary classifications efface a lot of diversity. For example, Finland has a third official language, Sami, which is spoken by a small population in the north. Finland also has a small but growing immigrant population. Singapore's dominant racial classification groups citizens into Chinese, Malay, Indian, and Other, alongside a diverse immigrant population. However, I use these larger binary categories both for convenience (in Finland's case, given that I was unable to speak with any interview participants from smaller minority groups, despite having circulated my participant information sheet among a network of migrant teachers) and for participant anonymity (in Singapore's case, given that naming the specific ethnolinguistic background of non-Chinese participants would considerably increase identifiability).

independent schools. (Finland's private school system is small, and most private schools are not sharply differentiated from their public counterparts. Such private schools are government-funded and non-fee-paying [Kauko, Corvalán, Simola, & Carrasco, 2015], but they retain some administrative autonomy. While most interview participants from public schools said there was no manifest difference between the two categories, a participant from a private school pointed out that their school principal had more control over the school budget than the typical public-school principal.)

Given the small targeted sample size, it was more efficient to recruit interview participants through personal and educational networks rather than seeking access to national administrative directories of teachers. Accordingly, I circulated my participant information sheets (which are available in Appendix B) in emails to personal contacts and via Facebook, posting it on my personal profile and in relevant Facebook groups (such as the Cambridge University Finnish Society and HundrED Suomi, a Finland-based education networking group). In Finland, the call for participants was also circulated by Jaakko Kauko, my faculty host at the University of Tampere. Given my smaller network of contacts in Finland, I supplemented this network-based recruitment with some cold contacting once I had arrived in Tampere. Specifically, I emailed and telephoned a number of Tampere schools, as well as schools in rural areas within reach of Tampere. I also emailed the national teachers' union (Opetusalan Ammattijärjestö, OAJ), but did not receive a reply.

In the final sample of interview participants, three participants in the Singapore sample were people whom I knew personally, while the remaining nine were contacts of people whom I knew personally. In the Finland sample, one participant was a contact of my University of Tampere host; four were contacts of people whom I knew personally; two were contacts of people within my extended networks whom I did not know personally; two were contacts of other interview participants, snowball-style; and three resulted from speculative emails to school principals.

Salient characteristics of interview participants are summarised in Table 3.5. The 12 Singaporean participants represent 11 different secondary schools, while the 12 Finnish participants represent 10 lower secondary schools. Broadly, both samples covered the range of characteristics I sought to represent. However, the balance of this coverage varied. The Singapore sample was skewed towards teachers with 10 or fewer years of teaching experience, due to the demographic skew of my personal networks. Still, the sample still had reasonably balanced coverage between (a) teachers who entered the profession prior to the introduction of the current performance management system in 2005, (b) those who entered the profession prior to a major revision of the system in 2014, and (c) those who had only experienced the revised system, as observed in Section 5.5. Among Singaporean participants, there was also better representation of subjects in the social sciences and humanities, although all STEM subjects were also represented. In Finland, only 3 out of my 12 interview participants were men; however, this corresponds loosely with the gender distribution of Finnish teachers (in 2016, 23% of basic education teachers were male; Paronen & Lappi, 2018). Although I spoke with participants from 8 municipalities across 4 out of Finland's 21 regions (i.e. Helsinki-Uusimaa, Pirkanmaa, Southwest Finland, and Central Finland), I only managed to interview one teacher from a rural school. This was due partly to logistical constraints, since I was based in urban Tampere and met participants at locations within a daytrip from Tampere, and partly to the limits of my networks.

That said, in both countries, the sample included a good mix of teachers across administrative responsibilities and school types. In each country, one participant originated from the United States but had been through local teacher training programmes and had taught in the system for several years. (The fact that both non-national participants were American men was purely coincidental.) Also, as shown in Table 3.5, the Singapore sample also included four teachers from Malaysia who had first moved to Singapore to attend secondary school or university. However, given the large (though far from total) similarity in sociocultural context between Malaysia and Singapore, this is a less significant source of sociocultural variation than the teachers of American origin.

Table 3.5 *Characteristics of interview participants, sorted by country and years of experience*

Pseudonym	Gender	Background	Teaching experience (years)	Position	Subjects	School type
Singapore: secondary school teachers (July 2018)						
Adeline	Female	Chinese, Malaysian origin	0–5	Teacher	Maths, Social Studies	Mainstream
Jeffrey	Male	Chinese	0–5	Teacher	Literature, English	Mainstream
Maggie	Female	Chinese, Malaysian origin	0–5	Teacher	Biology, Geography	Mainstream
Joseph	Male	Chinese	0–5	Teacher	History, English	Mainstream
Peter	Male	non-Chinese	0–5	Subject head	English, History	Autonomous/independent
Andy	Male	Chinese	5–10	Teacher	Literature, English	Mainstream
Sonia	Female	non-Chinese	5–10	Teacher, with admin portfolio	English, History	Mainstream
Timothy	Male	Chinese	5–10	Teacher	History, English	Mainstream
Mark	Male	non-Chinese, American	5–10	Subject head	History, Social Studies	Autonomous/independent
Eleanor	Female	Chinese, Malaysian origin	20–25	Senior teacher	Literature, English	Autonomous/independent
Geok Ling	Female	Chinese	25–30	Subject head	Chemistry, Maths	Autonomous/independent
Jane	Female	Chinese, Malaysian origin	30–35	Head of department	Maths, Physics	Mainstream
Finland: lower secondary school teachers (September 2018)						
Anneli	Female	Swedish-speaking	0–5	Teacher	Biology, Geography, Textiles	Swedish-speaking, urban, public
Liisa	Female	Finnish-speaking	10–15	Teacher, with admin portfolio	Maths, Physics, Chemistry	Finnish-speaking, urban, public
Emilia	Female	Bilingual Swedish & Finnish	10–15	Teacher	English	Swedish-speaking, urban, public
Kristiina	Female	Finnish-speaking	15–20	Teacher (formerly principal)	Music	Finnish-speaking, urban, public
Masa	Male	English-speaking, American	15–20	Teacher	Physics	Finnish- & English-speaking, urban, private
Satu	Female	Finnish-speaking	15–20	Teacher	Textiles, Geography	Finnish-speaking, rural, public
Maarit	Female	Finnish-speaking	15–20	Assistant principal	English, German	Finnish-speaking, urban, public
Hannele	Female	Finnish-speaking	20–25	Teacher	Swedish, Russian	Finnish-speaking, semi-urban, public
Helena	Female	Swedish-speaking	20–25	Teacher, with admin portfolio	Maths, Physics	Finnish- & English-speaking, urban, private
Antero	Male	Finnish-speaking	20–25	Teacher	Metalwork, Woodwork	Finnish-speaking, semi-urban, public
Päivi	Female	Finnish-speaking	25–30	Teacher (formerly principal)	Maths, Chemistry, Physics	Finnish-speaking, urban, public
Juhani	Male	Finnish-speaking	25–30	Teacher, union representative	Chemistry, Physics, Maths	Finnish-speaking, semi-urban, public

Some teacher characteristics are not included in Table 3.5, in order to minimise participants' identifiability. In both countries, some teachers had experience in teaching other school levels, whether primary or upper secondary. Furthermore, one participant in the Singapore sample had left the teaching profession one year prior to the interview, and another had been transferred to the Ministry headquarters (while retaining the same 'education officer' job title as any Singaporean teacher) two weeks prior to the interview. One Singaporean participant had been on medical leave for several months. In Finland, one participant had been working for a Finland-based international education organisation since 2015. However, there had not been any major changes in teacher accountability policy since these Singaporean and Finnish teachers left their respective classrooms, and there were no apparent differences between their views and other participants' views of teacher accountability processes. When I quote observations from these participants for which their background differences are relevant (e.g. when the Finnish participant working in international education discussing their experience in other countries), I will cite the pertinent distinguishing characteristic without giving the participant's pseudonym. Thus, I hope to avoid associating these distinguishing characteristics with the characteristics in Table 3.5, to further safeguard against identifiability.

Designing the interview guide

All 24 participants were interviewed using the same interview guide, with some country-specific variations. The guide is shown in Table 3.6, along with explanations of why I asked each question. In designing the interview guide, I followed three guiding principles. Firstly, in accordance with Pawson and Tilley's (1997) mode of realist interviewing (discussed above in Section 3.1), I explicitly told participants about some aspects of my conceptual framework, and asked for their input on both the entities within the framework and the soundness of the framework as a whole. For example, early in the interview I explained, with examples, how I defined teacher accountability instruments, and asked teachers about the instruments they had encountered. Toward the end of the interview, I described my working hypothesis and asked participants whether they agreed with it, and why. Unfortunately, at the point of conducting the interviews, I was in the process of developing the framework with accountability mechanisms and pathways described in Section 2.5, but I was not yet confident enough in this formulation to ask interview participants about it. Consequently, the most specific statement of my working theory that I gave in the interviews took the form of a platitude with which it would have been difficult to disagree: ‘effective education systems have teacher accountability instruments that are compatible with their sociocultural context’. Unsurprisingly, every participant did agree with this generalised hypothesis. (This agreement is even less surprising since my attempt to contextualise this hypothesis, as shown in #11 in Table 3.6, probably primed participants to agree for social desirability’s sake, since I positioned the hypothesis in opposition to superficial headlines and education rankings.) Fortunately, they did not only state their agreement, but also provided insightful elaborations that aided subsequent refinement of the conceptual framework.

Secondly, the progression of questions in the interview guide loosely followed the principle of ‘hierarchical focusing’ (Tomlinson, 1989)—that is, moving from more open to more focused questions. The goal here was to maximise coverage of participants’ views independent of researcher influence, while maintaining a focus on the research agenda. However, I do not follow all of the procedures that Tomlinson (1989) recommends, e.g. laying out interview questions in a strict visual hierarchy. This is partly because Tomlinson’s approach aims to minimise researcher framing and direction throughout the whole interview. In contrast, I sought to minimise my framing of the issues at the start of the interviews, but by the end of each interview I had explicitly communicated a significant portion of my conceptual framework, in line with Pawson and Tilley’s (1997) approach. For example, in discussing sociocultural context, I first asked participants to identify aspects of society and culture that affect their country’s education generally, then I asked about aspects of culture that affect teachers’ responses to accountability instruments. At the end of the interview, I asked whether they agreed with specific sociocultural descriptors drawn from the cross-country survey datasets that I used in the statistical analysis. (As shown in Table 3.6, the aggregate measures that I asked Singaporean participants about were high confidence in public institutions, high power distance, and weak adherence to civic norms. In turn, I asked Finnish participants about high social trust, low power distance, and strong adherence to civic norms. Even though Singapore also had the lowest uncertainty avoidance score in Hofstede’s sample, I did not ask participants about uncertainty avoidance, since I have reasons for doubting this measure, as noted in footnote 14 in Section 3.3.)

Table 3.6 *Questions asked in the interviews, and their respective purposes*

	Interview question	**Purpose**
1	Why did you decide to become a teacher?	To explore whether career decisions reflect changes in teacher accountability
2	What does it mean to be a good teacher? • Where did your idea of a good teacher come from? Who or what has influenced it?	To explore whether sociocultural context influences teachers' conceptions of their profession
3	What are you most proud of in your work as a teacher?	
4	[*After providing a definition of teacher accountability instruments*] Can you tell me about the main instruments for teacher accountability in [*Singapore/Finland*]? • *Setting goals, expectations, or standards* o What does the Ministry expect of you? o How are these expectations communicated to you? o How much does this overlap with your expectations of yourself? • *Communicating information* o How is the information obtained (e.g. reports, observations)? o How is feedback communicated to you? • *Rewarding or penalising teachers* o What consequences result from evaluations or judgements about teachers' work? o How accurate are the judgements? • Consider all stakeholders: government, teachers' union, principal, colleagues, parents, students • *For experienced teachers*: Has this changed over the course of your career? If so, how, and why? • *For non-national teachers*: How does this compare to your experience of teaching in other countries?	To elicit participants' experiences of teacher accountability instruments
5	How do these accountability instruments affect your work? • *Prompt for different areas*: how you teach and interact with students; which aspects of your work you prioritise; your workload; your autonomy; your motivation	To explore how accountability instruments affect teaching and learning, and whether there is evidence for the teacher motivation pathway
6	On the whole, do these accountability instruments make it easier or harder for you to be a good teacher?	
7	In what ways do these accountability instruments help [*Singapore's/Finland's*] education system be more effective in developing students' capabilities?	
8	In this research project, I am interested in how society and culture influence education. Can you tell me about aspects of [*Singaporean/Finnish*] culture that affect [*Singapore's/Finland's*] education system?	To elicit experiences of sociocultural contexts
9	In what ways does [*Singaporean/Finnish*] culture affect how teachers respond to accountability instruments? • *For non-national teachers:* How do you think teachers in your home culture would respond to [*Singapore's/Finland's*] teacher accountability instruments?	To determine whether sociocultural context affects experiences of accountability instruments

Table 3.6 (continued)

	Interview question	Purpose
10	[*Singapore version*] Imagine if Singapore were to adopt Finland's teacher accountability instruments. In Finland, teachers do not undergo formal evaluation. However, because there is no formal evaluation, and because the teaching career structure is flat, high-performing teachers are not rewarded with promotion. [*Finland version*] Imagine if Finland were to adopt Singapore's teacher accountability instruments. In Singapore, every teacher is formally evaluated twice a year by a supervising teacher, based on many different aspects of their professional work. Every teacher has a specific level of responsibility on the career ladder, and teachers are given grades based on how their performance compares to other teachers of the same level. These grades determine teachers' annual bonuses and affect their promotion through the teacher career ladder. • How would you react, and how would this affect your work? • How would most other [*Singaporean/Finnish*] teachers respond? • How do you think this would affect the effectiveness of the education system?	To elicit responses related to the efficacy and context-specificity (or lack thereof) of each country's teacher accountability instruments; and to share some of my preliminary findings with participants
11	Next, I would like to ask for your thoughts about some aspects of my research project. One thing in the backdrop of this research project is a set of discussions based on international rankings of education systems, where people say, 'Singapore does so well; let's copy their policies,' or, 'Finland has such good schools; let's copy their policies.' In this research project, I'm trying to argue that you can't successfully copy another country's policies wholesale. My current hypothesis is that effective education systems have teacher accountability instruments that are compatible with their sociocultural context. From your perspective, to what extent is this hypothesis plausible?	To seek participants' help in refining my working theory
12	[*Singapore version*] In my statistical analysis, besides using educational datasets like PISA and TIMSS, I also used sociocultural datasets like the World Values Survey. I would like to ask your opinion on how accurately these datasets describe Singaporeans. Would you agree that most Singaporeans whom you know: • are very confident in the quality of public institutions? • are quite accepting of authorities and hierarchies? • are a bit more likely than average to think that it's justifiable to go around the official system if you can, such as by avoiding taxes if you have the chance, or claiming government benefits that you're not actually entitled to? [*Finland version*] In my statistical analysis, besides using educational datasets like PISA and TIMSS, I also used sociocultural datasets like the World Values Survey. I would like to ask your opinion on how accurately these datasets describe Finnish people. Would you agree that most Finns whom you know: • trust other people to act fairly? • prefer equal distributions of power rather than unequal hierarchies? • think that you should follow the official system rather than trying to find loopholes or take shortcuts for your own benefit?	To triangulate the cross-cultural survey data with participants' first-hand experiences of their sociocultural contexts; and to share some of my preliminary findings with participants
13	Is there anything else that you would like to mention? In particular, are there any other important aspects of teacher accountability or sociocultural context in [*Singapore/Finland*]—or any connections between these things—that we have not discussed yet?	To identify areas that my working theory does not yet cover

Thirdly, in line with the realist emphasis on assessing the validity of an argument by considering of alternative explanations, I asked some questions that would allow me to weigh the evidence for or against some rival hypotheses. For example, one alternative hypothesis is that teacher accountability instruments influence people's decisions about whether to self-select into the teaching profession, rather than influencing teacher motivation in-service, as noted in Section 2.5. Accordingly, I asked each participant why they had decided to become a teacher (#1 in Table 3.6), to see whether there were systematic differences in the motivations of teachers who entered the profession during different incarnations of the teacher accountability system.

Besides these general principles, I referred to interview guides from previous accountability-related studies (e.g. Abelmann et al., 1999; Broadfoot & Osborn, 1993) when designing my interview guide. Additionally, I piloted the guide with a Singaporean teacher, a Singaporean Ministry official, and a Finnish principal, all of whom suggested ways of clarifying the phrasing of some questions.

Recognising the collective but not necessarily universal nature of sociocultural context, as discussed in Section 2.2, I asked participants not only about their own experiences and sentiments, but also their observations of the practices and values of other teachers and other Finns/Singaporeans (see, e.g. #10 and #12 in Table 3.6). Thus, I aim to gain insight into the group-level context, without effacing the heterogeneous individuals whose personal senses of responsibility is fundamental to meaningful accountability (Abelmann et al., 1999).

As my fieldwork progressed, I made three minor modifications to the interview guide. After conducting the first interview in Singapore and reflecting on realist-informed research methods, I began asking participants explicitly about whether they agreed with (a) my working hypothesis and (b) the summary descriptions of their country's sociocultural context derived from the secondary statistical datasets (#11 and #12, respectively, in Table 3.6). Secondly, I had initially asked participants about two hypothetical situations: what would happen if their country took away all teacher accountability instruments, and what would happen if if their country adopted the other country's teacher accountability approach (i.e. if Singapore adopted Finland's approach, or vice versa). However, after the second interview, I dropped the hypothetical about removing all teacher accountability instruments, because it seemed implausible given the breadth of my definition of accountability instruments, and because the second hypothetical appeared to be an adequate elicitation device. Finally, the first Finnish interview participant briefly struggled to understand my conceptualisation of teacher accountability instruments, partly because of second-language issues and partly because there is no convenient translation of 'accountability' in Finnish. In subsequent interviews, rather than reading out the definition, I gave participants a printout of my definition of teacher accountability instruments, in both English and Finnish, and suggested that they read the definition at their own pace. A copy of this printout is shown in Appendix B, together with copies of the participant information sheets, consent forms, and interview guide printouts.

Conducting the interviews

Each interview was conducted in a location of the participant's choosing, with the stipulation that the location had to be suitably quiet. Half the interviews—four for Singapore, eight for Finland—took place in the participants' schools. Among the other Singapore interviews, seven took place in cafés, and the final one took place via video conference a few weeks later, as the participant was not in Singapore at that time. As for the other Finland interviews, one was in a café, and the other three were in meeting rooms. The average interview duration was 61 minutes. (Coincidentally, this mean duration was the same for each country subsample as well as the full sample of 24 interviews. The Singapore interviews ranged from 31 minutes to 1 hour 27 minutes, and the Finland interviews ranged from 36 minutes to 1 hour 39 minutes.) All interviews were conducted in English, which is the dominant language in Singapore. In Finland, English is a widely spoken second language. Although some Finnish participants occasionally grasped for words, I asked follow-up questions to clarify ambiguities, and we consulted a dictionary when necessary. In a few instances, participants used Finnish terms, which I subsequently looked up.[20]

Procedurally, I began each interview by explaining my background and outlining the study. I then gave the participant a printout of the information sheet (which they would have received in advance electronically) and asked them to read it and to complete the consent form, if they agreed. All interviews were audio-recorded. In each interview, I referred to a copy of the interview guide to ensure consistency of coverage and of phrasing. I also used the guides to jot down clarifying notes. Where appropriate, I added follow-up questions, especially for those participants who were happy to spend more time on the interview. In a few cases, I skipped some questions because the participant faced time constraints.

[20] Chung (2009) conducted field interviews among education stakeholders in Finland using a mix of English, Swedish, and Finnish, and concluded that these interviews were enriched by participants' freedom to express themselves in their preferred languages. However, given the time constraints facing my PhD project and the complementary insight offered by the Singapore interviews and the statistical analysis, I believe that the English-medium interviews provided more than adequate insight from Finnish participants.

As soon as possible after completing each interview, I wrote down some field notes on the interview setting, our interactions, any participant remarks that might be ambiguous in the audio-recording (e.g. because of body language that the recording would not have captured), reflections on how to improve my interviewing technique, and my overall impressions from the interview. Subsequently, I transcribed all interviews verbatim, using the notation system in Table 30.4 of Poland (2001), which captures some conversation features (e.g. pauses) but not the minutiae of formal conversation analysis. I then played back each recording in full, in order to check transcription accuracy.

For confidentiality, personally identifying information has been redacted from the transcripts, and transcripts are identified using pseudonyms. In Singapore, I gave participants the option of choosing their own pseudonym, to strengthen participants' ownership over the interview process. I did not offer this option in Finland, as my University of Tampere host viewed this as unnecessary. Accordingly, for Finnish participants and for the Singaporean participants who did not want to choose their own pseudonym, I referred to lists of popular names by the participant's approximate birth cohort (e.g. from Finland's Population Register Centre, www.vrk.fi). I chose pseudonyms that matched the linguistic origin of participants' names (e.g., for Singaporean participants, Anglicised names or Chinese names).

For member validation, I emailed a copy of their respective transcript to the nine Singaporean and five Finnish participants who had requested it. All five participants who responded said that they were satisfied with the interview and transcription process.

Analytic approach

To facilitate analysis, I coded the interview transcripts in QDA Miner Lite. Since the aim of the interviews was to test and refine my conceptual framework, I began with a preliminary coding scheme that was based on the framework. The categories (and subcategories) in the preliminary scheme were: (a) teacher accountability (instruments, mechanisms, effects); (b) sociocultural context; (c) hypothesis, for text segments related to the interaction between accountability and sociocultural context; and (d) other (teacher role conceptions, change over time). As I coded each transcript, I added subcategories and codes where appropriate. In between coding each transcript, I reread the full coding scheme, conducted code retrievals to check the aptness of any code assignments that I was uncertain about, and made revisions to codes and code assignments.

After making any substantial change to the coding scheme, I revisited already-coded transcripts to reassess and update their code assignments.

Thus, this coding process involved a mixture of deductive and inductive analysis. I began the process with a rough coding scheme organised in broad categories, as noted above. Also, the overall causal pathway in the conceptual framework—where teacher accountability instruments use various mechanisms to influence teachers (or change stakeholders' decisions) in ways that improve student outcomes, with the whole pathway being shaped by context—did not change as a result of the interview analysis. These features of the process imply a deductive approach.

However, much of the coding and concept formation was guided by the data, reflecting an inductive approach. In terms of coding, I added and iteratively revised new subcategories and codes depending on what emerged from the data, such as a subcategory for trade-offs (e.g. monitoring costs vs. information availability) under the teacher accountability category, and codes for 'autonomy and conformity' and 'competitiveness and disinterestedness' under sociocultural context. As for concept formation, the interview analysis directly shaped the accountability mechanisms included in the framework. Prior to the interview analysis, the mechanisms had gone through several permutations, and the most recent version comprised four mechanisms. These were the standard-setting, informational, and consequential mechanisms that were eventually included in the final conceptual framework; as well as a reputational mechanism, i.e. accountability pressure resulting from teachers' desires to maintain their professional reputations in colleagues' or other stakeholders' eyes. However, interview participants' descriptions of such reputational pressure made it clear that this reputational mechanism was merely a special case of the informational mechanism, wherein accountability pressure results from teachers' desires to be regarded favourably when information about their practice is compared to implicit or explicit expectations (see Section 5.4 for specific examples). Accordingly, I took out the reputational mechanism, leaving the three other accountability mechanisms. In using the interview data to progressively refine a working theory, my analysis broadly followed Pawson and Tilley's (1997, Chapter 6) realist approach to interview research.

Although I coded each transcript in full, the primary function of this coding was to organise the interview corpus and to facilitate retrieval of related quotes. Given the small, non-systematic sample of interview participants, it could be misleading to present numerical counts or percentages of any of the codes or themes. Instead, I summarise some theoretically pertinent aspects of the interviews using matrices in which each column represents a participant, and the rows contain symbols (e.g. a fully shaded circle for 'agree', and a half-shaded circle for 'partially agree') based on my interpretations of each transcript as a whole. This provides a visual summary of the interview corpus, without a false impression of statistical precision. (For similar approaches to visualising qualitative data, see Abraham and Dohan [2015] and Wilkinson and Friendly [2009].) I also include illustrative quotes, which are edited for readability (e.g. removing 'umm's and repetitions). If the participant stressed a particular word or phrase, I indicate this emphasis using italics. In some instances, portions of the quotes have been excluded for conciseness. Where this occurs, excluded portions are indicated with '[…]', and care has been taken to ensure that the exclusions do not alter the meaning of the quote. Usually, quotes illustrate general themes in the interview corpus. Where I use a quote to illustrate an uncommon position, I state this explicitly (Bleich & Pekkanen, 2013).

To strengthen the reliability of the interview analysis, I reread each transcript in full after the initial round of constructing summary matrices and selecting representative quotes, in order to ensure that the matrices and quotes accurately represented the interview in question. (In this project, it would not have been particularly meaningful to approach reliability using the intra/inter-rater reliability statistics. The inter/intra-rater reliability approach focuses on the consistency of coding across two separate instances of coding the same segments of text. However, I conducted, transcribed, and analysed all the interviews myself, with iterative checks and revisions to the coding along the way, so it was more appropriate to check reliability in a more holistic, interpretive way. Moreover, since I used the codes instrumentally—i.e. to make text retrieval and analysis more efficient rather than to constitute final classifications—it makes more sense for the reliability checks to focus on the summary matrices and quotes rather than the coding.) I also regularly reread my field notes and listened to portions of the original audio throughout the process of analysis. Additionally, I recorded my analytic decisions in memos, which I referred to when appropriate (and which I had likewise done while working on the statistical analysis). Finally, I referred to other sources to corroborate my observations from the interviews, e.g. consulting available policy texts on teacher accountability instruments, and referring to prior descriptions of sociocultural context in these countries.

3.5 Ethical considerations

Throughout this research project, I adhered to BERA's (2011) Ethical Guidelines for Educational Research. I applied for, and received, ethical clearance from the Faculty of Education. Additionally, I submitted an application for fieldwork approval to the Corporate Research Office of Singapore's Ministry of Education. However, I was informed that I was 'not required to seek approval from MOE for data collection because [I was] recruiting participants through personal networks, and not using schools / MOE as a formal platform to recruit participants' (email from Ang Lee Cheng Stephanie, senior research analyst, 28 March 2018). For the Finland fieldwork, there was no requirement to apply for either ministerial or municipal approval because I was working with adult participants.

In the interviews, I ensured that all participants gave their voluntary informed consent, by providing them during initial contact with an electronic copy of the participant information sheet and consent form, which detailed the purpose of the research and what it would require of them. At the beginning of each interview, I discussed the contents of the consent form with the participant, and requested their signed endorsement of a copy of the form. I also ensured that participants were aware of their right to withdraw from the research at any point, and I sought verbal consent prior to turning on the audio recorder. As stated above, I also conducted member checks of interview transcripts, allowing participants to raise concerns or clarifications.

After the completion of the interviews, the protection of participants' personally identifying information is crucial, both to minimise any potential harm resulting from these interviews and to comply with the General Data Protection Regulation (GDPR). Accordingly, interview transcripts only include pseudonyms, rather than real names, and I have redacted any potentially identifying proper nouns mentioned during the interviews. (I also made sure that I did not write down participants' names in the notebook with my field notes.) Pseudonyms are matched to participants' names only in a spreadsheet with two layers of password protection. Audio recordings of interviews as well as the scanned copies of participants' consent forms and interview guides are likewise protected by two sets of passwords, one of which enables Microsoft's Bitlocker encryption.[21] In this thesis and in other research reports, personal details

[21] Besides the main copy of these files on my computer hard drive, I have a backup copy on a Bitlocker-encrypted USB stick, and another Bitlocker-encrypted copy uploaded to my University of Cambridge Microsoft OneDrive account, which is also password-protected. Microsoft states that its processors are GDPR-compliant (Microsoft Trust Center, n.d.).

are generalised into broad categories (e.g. I used five-year ranges for participants' teaching experience, rather than stating the exact year in which they started teaching), to ensure that participants are not identifiable. After the completion of this study and its associated reports and publications, I will destroy all personal data from which interview participants may be identified, including the audio recordings, interview guides, and the spreadsheet containing contact details and pseudonyms. As indicated on the blank consent forms in Appendix B, participants have given their consent for the anonymised transcripts to be retained for my own reference in password-protected storage after the completion of the study. I will also retain digitised copies of the consent forms, under separate password protection, for as long as I retain the transcripts.

Another set of ethical considerations surrounds the analysis and presentation of data from PISA and TIMSS. These international student assessments have, inadvertently or otherwise, been used in high-visibility global academic horse races that intensify performance pressures on students, teachers, and schools. Additionally, some scholars assert that data from these assessments have been used to justify harmful policy programmes (e.g. Komatsu & Rappleye, 2017a). Despite these adverse circumstances, I also believe that there is an ethical obligation to make education policy decisions as rigorous as possible, which includes using relevant, high-quality data to inform such decisions where appropriate. In this project, I attempt to mitigate the potential adverse effects of using ILSA data by balancing observations from the statistical analysis with teachers' voices from the interviews.

A thornier ethical issue in my statistical analyses surrounds the measurement of sociocultural context. Even setting aside the fact that sociocultural context is difficult to categorise, much less quantify (since it varies greatly even within countries, and its characteristics are neither clearly bounded nor distinctly separable), classifications of sociocultural context have a history of morally objectionable applications. For example, the factually inaccurate notion of 'Asian values' was influentially advanced by authoritarian leaders in Singapore and elsewhere in Southeast Asia to justify non-democratic rule (D. J. Kim, 1994; Sen, 1999). More generally, analysing sociocultural data at the national level may yield unfair generalisations that reinforce cultural prejudices.

To lessen the risk of such generalisations, I will analyse multiple sociocultural constructs, thus portraying sociocultural context as complex. The bulk of the sociocultural data will be drawn from the nationally representative WVS/EVS surveys. As far as possible, I will identify the sociocultural variables using language that is descriptively close to the questionnaire items, rather than layering on unnecessary interpretations or connotations (e.g. naming the scale derived from items about membership in various voluntary organisations as 'civic networks', rather than 'community strength' or 'enthusiasm for volunteering'; see Section 4.2). Also, I will not define sociocultural constructs as polarities (as in Inglehart and Welzel's, 2005, traditional vs. secular-rational and survival vs. self-expression values), because such polarities may facilitate pigeon-holing. For example, one of my sociocultural constructs is derived from questionnaire items asking respondents whether they would say that 'that most people can be trusted or that you need to be very careful in dealing with people' and to what degree 'most people would try to take advantage of you if they got a chance' (WVS Association, 2012). I am labelling this construct 'social trust', such that countries have more or less social trust; rather than 'trust vs. suspicion', which would favour an arbitrary designation of some countries as trusting and others as suspicious. Moreover, I complement these statistical scales with interview participants' firsthand descriptions. When asking participants about sociocultural context, I begin with questions that are as open as possible in order to elicit their independent descriptions of their contexts, as shown in Table 3.6. Subsequently, I ask them about the extent to which the statistical constructs reflect their own perceptions. Thus, I aim to construct sociocultural descriptions that are nuanced rather than caricatured.

3.6 Limitations

Besides these ethical considerations, this project faces a range of analytical limitations. To continue the discussion on sociocultural context, both of my sources of sociocultural data come from self-reports, whether from survey respondents or interview participants. Self-report data may incorporate substantial bias from participants' frames of reference and survey response styles (He, van de Vijver, Dominguez Espinosa, & Mui, 2014; Johansson, 2018), as well as social desirability bias. Such perceptional biases are far smaller in observational studies, such as Cohn, Maréchal, Tannenbaum, and Zünd's (2019) study of civic honesty using 17,000 'lost' wallets in 40 countries around the world. However, such large-scale comparative studies are rare. (Unfortunately, the Cohn et al. study was published in June 2019, which was too late for me to re-estimate the statistical models using their wallet return rates as a sociocultural proxy. Additionally, since their experiment was carried out between July 2013 and December 2016, there would be a slight problem with the time-ordering of the ILSA outcome data and their sociocultural data, even for TIMSS and PISA 2015. However, it would be fascinating to analyse

TALIS 2018 and, once released, PISA 2018 data alongside their wallet return rates.) Whatever the biases of the WVS/EVS data, they are far more extensive than extant observational studies of sociocultural context, whether in their coverage of countries or of sociocultural domains.

More generally, cross-country surveys face numerous issues with measurement quality (Survey Research Center, 2016). Scholars have raised questions about the limited forms of learning that PISA scores measure (Komatsu & Rappleye, 2017b), the comparability of PISA proficiency scores across paper- and computer-based testing (Jerrim, Micklewright, Heine, Salzer, & McKeown, 2018), the comparability of PISA attitudinal scales across countries (Liaw, Wu, Rutkowski, & Rutkowski, 2018),[22] the translation of WVS items (Kurzman, 2014; Mellon, 2011), and the replicability of Hofstede's indices (Merritt, 2000).

Further issues can arise in scale construction. The construction of any statistical scale entails choices about modelling and item inclusion, all of which can alter the meaning and validity of the scale. For example, Carrasco Ogaz (2016) proposes an alternative approach to constructing the TALIS 2013 teacher job satisfaction scale that yields better model fit than the OECD's official approach. Besides inheriting the biases inherent in any of the pre-existing OECD, IEA, and Hofstede scales that I use, additional biases arise from the scales that I construct. For example, although my accountability scales selectively included only the items that fit my definition of teacher accountability instruments, I was nevertheless limited to items that were administered in the questionnaires. Accordingly, these accountability scales include far more formal teacher accountability instruments than informal ones, as described in Section 4.1.

In addition, both the ILSA data and the teacher interviews have sample bias. With the ILSAs, there is selection bias in country participation.[23] Given that participation in such assessments incurs substantial resource costs (Engel & Rutkowski, 2018) and can place pressure on countries perceived as underperforming, the countries that choose to participate may differ systematically from those that do not, whether in wealth or in policy priorities. This would bias the results of my statistical analysis if participating countries tend to implement certain accountability instruments or tend to share certain sociocultural traits. (This is likely to be the case, especially

[22] For a recent overview of technical critiques of PISA data, see Hopfenbeck et al (2018).

[23] Another form of selection bias is that PISA and TIMSS only assess children who are enrolled in school, which, in some countries, excludes a large proportion of the relevant age group. This exclusion of out-of-school children may significantly skew country-level measures of educational achievement, thus affecting comparisons over time and space (Spaull, 2019). However, I do not expect that this will systematically bias my analysis, since I focus on in-school effects rather than aggregate learning levels countrywide.

since all OECD countries participate in PISA, and OECD countries tend to have higher levels of development, which is often associated with higher levels of social capital. In fact, t-tests indicate that countries participating and not participating in PISA and TIMSS do differ significantly in social trust and power distance, though not in the other sociocultural constructs. It is worth noting that social trust and power distance are also the constructs that are most strongly correlated with GDP, as shown in Table A.2, and that wealthier countries are more likely to participate in ILSAs. However, it is also possible that controlling for GDP may attenuate this particular bias.) Nonetheless, there is still a wide range of sociocultural contexts and approaches to teacher accountability among the sample of countries in these educational surveys. For example, according to principal-reported data in PISA 2015, only 3.4% of 15-year-olds in Germany attend schools where standardised tests are used to make judgements about teachers' effectiveness, while the corresponding figure for Russia is 89.1%. Additionally, the countries with the highest and lowest scores for power distance (Malaysia/the Slovak Republic and Austria, respectively) and for uncertainty avoidance (Greece and Singapore, respectively) all participated in both PISA 2012 and 2015.[24]

With the teacher interviews, one nontrivial source of bias is that all participating teachers were willing to take the time to speak to a researcher. This suggests that they all value education research, at least to some degree, and probably value educational improvement as well. While participants from both countries mentioned colleagues who were happy to do the bare minimum, and Singaporean participants mentioned colleagues who unapologetically focused on climbing the career ladder at the expense of classroom practice, none of the 24 interview participants appeared to fit into these categories. That said, given the demonstrated quality of teaching and learning in Finland and Singapore, my interview participants are likely to be more representative of their respective professions than the less principled colleagues they mentioned.

Finally, the nature of my data sources limits the scope of the causal arguments that I can make. From a statistical standpoint, the cross-sectional ILSA data do not permit causal inference. Beyond the ILSA data, the granular interview data paired with the realist approach of assessing alternative explanations do allow me to argue for some of the causal pathways in the conceptual framework. However, my field interviews, like the ILSA data, are temporally flat. As such, the

[24] As stated in the note on Table 3.3, Malaysian data is not included in the main PISA 2015 dataset because of data quality issues. Also, as stated in footnote 14 in Section 3.3, I doubt the accuracy of Singapore's uncertainty avoidance score. Nonetheless, the point still stands: there is considerable sociocultural variation within the datasets analysed in this thesis.

interviews cannot account for certain causal pathways that might yield endogeneities over time, such as cases where teacher accountability instruments gradually shape teachers' socioculturally embedded understandings of what constitutes good teaching and, by extension, how their motivation should be oriented (J. Holloway & Brass, 2018). Moreover, the interviews are not accompanied by any observational data on how self-reported teacher motivation might manifest in classroom practice and student learning, which limits my qualitative arguments to the first part of the teacher motivation causal pathway.

Still, PISA and TIMSS remain the most internationally representative datasets on teacher accountability and student outcomes that can be viably analysed for a PhD project. The same applies to WVS/EVS and Hofstede's scales for sociocultural context. Furthermore, although classroom observations alongside longitudinal student assessments may shed more light on the link between teacher motivation and student outcomes, it would not have been feasible to conduct such observations in addition secondary statistical analysis and field interviews within a three-year PhD project. Thus, given these limitations, my hope here is not to produce definitive answers to every possible question about teacher accountability and sociocultural context, but rather to make a case for the importance of sociocultural considerations in designing teacher accountability policies, and to build support for a framework that may facilitate such context-sensitive policy design.

Chapter 4: Teacher accountability, student outcomes, and sociocultural context

In this chapter, I begin by describing the process of constructing scales from the cross-country survey data on teacher accountability instruments (Section 4.1) and on four aspects of social capital (Section 4.2). I then present results from the multilevel moderation analysis (Section 4.3), thus addressing my first research question: *to what extent does the influence of teacher accountability instruments on student outcomes depend on sociocultural context?*

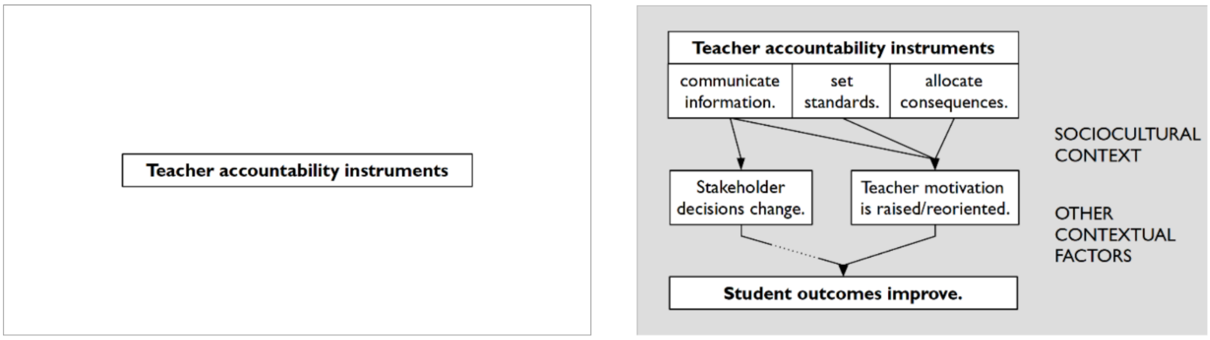
As noted above, the multilevel models in this chapter analyse cross-sectional datasets without an identification strategy. From a statistical standpoint, this analysis can establish, at best, an association rather than causation. Nonetheless, RQ1 refers to the *influence* of teacher accountability on student outcomes, because this phrasing reflects the realist ontological standpoint of this thesis. As discussed in Section 3.1, the realist conception of causality hinges on whether entity A (e.g. teacher accountability instruments) actually generates change in entity B (e.g. student outcomes)—in contrast to the statistical 'constant conjunction' understanding of causality. Accordingly, in investigating RQ1 using a realist approach, I end the chapter by complementing the essentially correlational statistical analysis with a consideration of the evidence for and against some alternative explanations for the statistical findings (Section 4.4).

4.1 Scale construction for teacher accountability instruments

As noted in Section 3.3, I operationalise teacher accountability instruments by constructing school-level scale variables for teacher accountability instruments in PISA and TALIS. In this section, I discuss the construction of these scale variables, as indicated in Figure 4.1. Descriptive statistics for both scales are available in Table 3.4 in the previous chapter.

To construct these scales, I used item-response theory (IRT) modelling in Stata 15.1. IRT is likewise used by the OECD and the IEA to construct the scales that are included in the public-release PISA and TIMSS datasets. One benefit of IRT modelling is that it can generate scores for any cases that have data on at least one of the items underlying the scale, thus yielding far less missingness in the generated scale variable. Specifically, 5.3% of schools in the pooled PISA 2012 and 2015 dataset did not have data on any of the included teacher accountability questionnaire items, and hence do not have an IRT score for accountability. In contrast, three times as many schools, 16.0%, did not have data on one or more items. If I had used a scale construction method that could only generate scores for cases that have observations for all items, e.g. constructing a simple sum of all included questionnaire items, then all 16.0% of these schools would have been excluded from the analysis. (There did not appear to be systematic differences between the cases that had data on at least one teacher accountability item—and, hence, were included in the multilevel regressions—and the 5.3% that did not have data on any of the accountability items. In the PISA 2015 dataset, the mean science proficiency scores for these groups differed by 3.20 points [SD=105.09 points] in favour of the group with at least some teacher accountability data. The difference for student socioeconomic status was 0.09 [SD=1.10] in favour of the group without any accountability data. The differences for PISA 2012 were similarly small.) Additionally, this capability of IRT modelling for constructing scores with partially missing data meant that I could pool school-level data across both PISA waves, even though the PISA 2015 school questionnaires lacked some accountability-related items that were included in PISA 2012. With this larger pooled dataset, the teacher accountability scale is more reliable than similar scales constructed from single-cycle datasets would have been.

Figure 4.1 *Relationship between Section 4.1 and the overall conceptual framework*


Teacher accountability instruments
Teacher accountability instruments
communicate information.
set standards.
allocate consequences.
Stakeholder decisions change.
Teacher motivation is raised/reoriented.
Student outcomes improve.
SOCIOCULTURAL CONTEXT
OTHER CONTEXTUAL FACTORS

This section

Overall conceptual framework

Accountability-related items from the school (i.e. principal-reported) questionnaires were identified for inclusion in the scales based on whether they fell within my definition of teacher accountability instruments, i.e. *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*. Thus, I included school-level accountability instruments that had clear implications for teacher practice (e.g. school self-evaluation), but excluded items that may not necessarily have direct impact on teachers' work (e.g. 'regular consultation aimed at school improvement with one or more experts'; OECD, 2016b, p. 137; where it is unclear whether the instrument only affects principal leadership or whether it also affects teachers' work).

Table 4.1 summarises the teacher accountability instruments from each survey cycle that are included in the scale measures. The table also gives examples of evaluation studies in which each instrument was found to have a significant relationship with student achievement or a related outcome in an accountability-related intervention. Most of these studies involved a comparison of treatment and control groups, whether in randomised trials or natural experiments; while others use identification strategies to analyse longitudinal datasets. This is not to imply that these teacher accountability instruments always improve educational outcomes in all contexts. For example, Glewwe, Ilias, and Kremer (2010) found that financial incentives for teachers raised student outcomes, as indicated in the table, whereas Fryer (2013) found no effect from such incentives. (see also Section 2.1 on the inconsistent effects of teacher accountability instruments). Two further caveats: firstly, some of these interventions involved not only accountability instruments, but also other treatments, such as providing resources for improvement (e.g. Speckesser et al., 2018, evaluated a programme that entailed peer observation and discussion, but it also included a series of professional development workshops). Based on the evaluation designs, it is difficult to disentangle the degree to which the positive effects derived from the accountability instruments as compared to the other components of the intervention, or from the interaction of both. Secondly, even though all the cited studies found at least one significant positive outcome, these effects did not always occur across the board (e.g. Garet et al., 2017, found a significant positive effect for mathematics achievement, but not for reading). Nonetheless, to use Cartwright and Hardie's (2012) terminology, the studies summarised in Table 4.1 show that these accountability instruments have the potential to play a causal role in education, at least with certain programme designs in certain settings. Hence the inclusion of these items in the scales measuring the extensiveness of teacher accountability instruments in schools.

Table 4.1 *Teacher accountability instruments included in the scale measures, and studies indicating their potential for positive influence on education-related outcomes*

Teacher accountability instrument	Study	Significant positive effect on	PISA 2015	PISA 2012	TALIS 2013
Monitoring and appraising teachers					
Monitoring teachers using student test results	Garet et al. (2017)	Student achievement	✓	✓	
Appraisal/lesson observation by peers	Speckesser et al. (2018)	Student achievement	✓	✓	✓
Appraisal/lesson observation by senior staff/school leaders	Steinberg & Sartain (2015)	Student achievement	✓	✓	✓
Appraisal/lesson observation by external persons	Briole & Maurin (2019)	Student achievement	✓	✓	✓
Consequences of teacher appraisals					
Financial rewards/penalties	Glewwe, Ilias, & Kremer (2010)	Student achievement		✓	✓
Change in the likelihood of career advancement/work responsibilities	Karachiwalla & Park (2017)	Teacher appraisal scores		✓	✓
Public recognition	Bradler et al. (2016) (*field setting: generic, non-educational employment*)	Employee performance		✓	
Professional development/coaching/etc.	Papay et al. (2016)	Student achievement		✓	✓
Dismissal/non-renewal of contract	Adnot et al. (2017)	Student achievement			✓
School-level instruments that affect teachers' work					
Publication of student results	Andrabi, Das, & Khwaja (2017)	Student achievement	✓	✓	
Tracking of student results by administrative authority	Lassibille, et al. (2010)	Grade repetition	✓	✓	
Providing student results directly to parents	Hastings & Weinstein (2008)	Student achievement	✓		
External school evaluation	Rockoff & Turner (2010)	Student achievement	✓	✓	
School self-evaluation	Demetriou & Kyriakides (2012)	Student achievement	✓	✓	
Seeking written feedback from students	Overall & Marsh (1979)	University student achievement	✓	✓	
Written specification of school educational goals and student performance standards	Khattri, Ling, & Jha (2012)	Student achievement	✓	✓	
Systematic recording of data (e.g. test results, attendance, teacher training)	Lassibille, et al. (2010)	Grade repetition	✓	✓	

Note. Teacher accountability instruments are grouped according to questionnaire item sets in the PISA and TALIS surveys. All studies cited are experimental, quasi-experimental, or longitudinal with an identification strategy.

That said, as discussed in Section 2.1, there is considerable disagreement about whether more or less extensive approaches to teacher accountability will lead to better student outcomes. While some scholars weigh the evidence and recommend more extensive forms of teacher accountability (e.g. Hanushek, 2019), others call for caution (e.g. Zhao, 2018). Based on the findings that I present in this thesis, I believe there is more evidence for tailoring teacher accountability instruments to particular settings rather than for applying similar approaches across the board. The latter argument is implicit in, for example, the theoretical justifications advanced for introducing more accountability instruments in educational settings, on the basis that more extensive measurement and monitoring will improve outcomes (e.g. Barber, Rodriguez, & Artis, 2016). As I argue in Chapter 7, effective teacher accountability is not merely a matter of extent. However, for the purposes of this multilevel cross-country statistical analysis, having unidimensional teacher accountability scales greatly facilitates the modelling process. Furthermore, if I can reduce the PISA and TALIS teacher accountability items to a single scale with some construct validity, and if I can then use this scale to show that the relationship between the extensiveness of teacher accountability and student outcomes depends on context, then this could be a useful piece of evidence against the 'more accountability instruments=better outcomes' argument, as outlined in the 'Operationalisation' subsection of Section 3.3.

To proceed with the IRT scale construction for the pooled PISA 2012 and 2015 dataset, most of the PISA teacher accountability questionnaire items collected binary 'yes'/'no' responses, as shown in Table 4.2. Since the response categories of the remaining non-binary variables also fell neatly into whether or not an instrument was implemented in the school, I recoded them into binary form for scaling. (As shown in Table 4.2, these non-binary response categories were that (a) the PISA 2015 items on school-level quality assurance and improvement separated 'yes' responses into whether each instrument was school-initiated or externally mandated; and (b) the PISA 2012 items on consequences of teacher appraisals collected ordinal responses on whether there was no change, or a small, moderate or large change in each area. For the PISA 2015 items on school quality assurance and improvement, it is unlikely that there would be consistent cross-school and cross-country patterns across the dataset in whether an externally mandated accountability instrument would affect teachers' work more or less intensively than a school-initiated one. Neither is it likely, for the PISA 2012 items on consequences of teacher appraisals, that principals' qualitative assessments of these degrees of change would be comparable across schools and countries. Hence, if I had retained these variables in non-binary form, this would have complicated the analysis without adding obvious analytic utility.)

With these binary items, I constructed the scale using a two-parameter logistic (2PL) IRT model. This type of model is also used for official PISA 2015 scales from binary data (OECD, 2017). Having confirmed the unidimensionality of this teacher accountability scale using exploratory factor analysis, I ran a 2PL IRT analysis. Questionnaire items, together with discrimination and difficulty parameters, are shown in Table 4.2. (For comparison, I also ran a 1PL model, which estimates difficulty parameters for each item but assumes equal discrimination across all items. However, a likelihood ratio test showed a better fit for the 2PL model than the 1PL model, so I retained the former.) From this analysis, I treat the resulting latent trait parameter estimate for each school as a measure of school-level teacher accountability instruments. For this measure, higher scores correspond to a greater incidence of teacher accountability instruments in the school. The mean score for teacher accountability instruments was 0.00, with a standard deviation of 0.839 (n=34,130 schools).[25]

As shown in Table 4.2, the difficulty parameters of the PISA teacher accountability items range from -3.48 for 'Achievement data are provided directly to parents' to 0.81 for 'Achievement data are posted publicly'. This indicates that, all else equal, providing achievement data to parents is likely to be practised even in schools with few teacher accountability instruments overall; whereas posting achievement data publicly tends to happen only in schools that also implement many other accountability instruments. However, these two items have lower discrimination parameters than any other items. These low discrimination parameters indicate that knowing whether or not a school provides achievement data directly to parents (or posts it publicly) does not convey very much information about how many teacher accountability instruments it probably implements overall. In contrast, items with larger discrimination parameters, such as teacher appraisals leading directly to a change in the likelihood of career advancement, tend to offer more predictive insight into a school's overall teacher accountability scale score.

Notably, the parameter estimates indicate some construct validity. In particular, the items with the greatest difficulty parameters 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, as shown in Table 4.2. That is, the schools that implement these teacher accountability instruments are very likely to implement many other teacher accountability instruments as well. This makes intuitive sense, because these teacher accountability instruments are intensive not only in their potential impact on teachers' work, but also in the resources required for implementing them (with the possible exception of public posting of student test results). Consequently, it would be reasonable to suppose that the schools that implement these intensive accountability instruments probably prioritise teacher

[25] These descriptive statistics differ slightly from those shown in Table 3.4 in the methodology chapter, because (a) Table 3.4 shows descriptive statistics for separate PISA 2012 and PISA 2015 subsamples rather than for the pooled teacher accountability IRT sample, and (b) the subsamples summarised in Table 3.4 exclude supplementary region-specific samples for Russia in PISA 2012 and Spain in PISA 2015 that were not part of the main PISA survey for Russia and Spain respectively, but which I included in the IRT scale construction to increase sample size and, thus, scale reliability.

accountability instruments in general, and hence would be likely to implement the lower-hanging fruit among accountability instruments as well.

Table 4.2 *Questionnaire items and IRT 2PL parameters for PISA 2012 and 2015 teacher accountability*

PISA 2012 identifier	PISA 2015 identifier	**Questionnaire item**	Discrimination	Difficulty
		During the last academic year, have any of the following methods been used to monitor the practice of teachers at your school? *1='Yes'; 0= 'No'.*		
SC30Q01	SC032Q01TA	Tests or assessments of student achievement	1.10	-1.74
SC30Q02	SC032Q02TA	Teacher peer review (of lesson plans, assessment instruments, lessons)	0.99	-1.05
SC30Q03	SC032Q03TA	Principal or senior staff observations of lessons	1.21	-1.26
SC30Q04	SC032Q04TA	Observation of classes by inspectors or other persons external to the school	0.77	0.48
		In your school, are achievement data used in any of the following accountability procedures? *1='Yes'; 0= 'No'.*		
SC19Q01	SC036Q01TA	Achievement data are posted publicly (e.g. in the media)	0.58	0.81
SC19Q02	SC036Q02TA	Achievement data are tracked over time by an administrative authority	0.77	-1.64
—	SC036Q03NA	Achievement data are provided directly to parents	0.54	-3.48
		Do the following arrangements aimed at quality assurance and improvements exist in your school and where do they come from? *1='Yes' (2012); 'Yes, this is mandatory' or 'Yes, based on school initiative' (2015); 0='No'.*		
SC39Q05	SC037Q01TA	Internal evaluation/Self-evaluation	1.35	-2.26
SC39Q06	SC037Q02TA	External evaluation	0.85	-1.27
SC39Q01	SC037Q03TA	Written specification of the school's curricular profile and educational goals	1.04	-2.60
SC39Q02	SC037Q04TA	Written specification of student performance standards	1.12	-1.60
—	SC037Q05NA	Systematic recording of data such as teacher or student attendance and professional development	1.94	-1.99
—	SC037Q06NA	Systematic recording of student test results and graduation rates	1.75	-2.17
SC39Q03	—	Systematic recording of data including teacher and student attendance and graduation rates, test results and professional development of teachers	1.07	-2.20
SC39Q07	SC037Q07TA	Seeking written feedback from students (e.g. regarding lessons, teachers, or resources)	0.91	-1.05
		To what extent have appraisals and/or feedback to teachers led directly to the following? *1='A small change', 'A moderate change', or 'A large change'; 0='No change'.*		
SC31Q01	—	A change in salary	1.79	0.44
SC31Q02	—	A financial bonus or another kind of monetary reward	1.56	0.36
SC31Q03	—	Opportunities for professional development activities	1.99	-1.01
SC31Q04	—	A change in the likelihood of career advancement	2.41	-0.48
SC31Q05	—	Public recognition from you	1.91	-1.25
SC31Q06	—	Changes in the work responsibilities that make the job more attractive	2.01	-0.93
SC31Q07	—	A role in school development initiatives (e.g. curriculum development group, development of school objectives)	2.17	-1.29

Note. IRT 2PL = item-response theory 2-parameter logistic model. *N* = 34,130, of which *n* (PISA 2012) = 17,691 and *n* (PISA 2015) = 16,439. Questionnaire item phrasing is from OECD (2013a, 2016b).

For TALIS 2013, the teacher accountability scale had fewer questionnaire items and fewer categories of accountability instruments than its PISA counterpart, as indicated in Table 4.1 above. All of these TALIS 2013 accountability items collected ordinal responses on the degree to which an instrument was used in the school, so I ran a graded-response IRT model (GRM) rather than the 2PL model used for the binary PISA data. (Although the OECD typically uses the generalised partial-credit model [GPCM] for polytomous IRT scales, I chose to use the GRM instead because the latter yielded more logically coherent parameter estimates. Specifically, in a GPCM analysis of the TALIS 2013 teacher accountability items, some of the difficulty parameters did not follow a logical sequence between the ordinal categories. For example, for some items, the threshold between implementing an accountability instrument 'never' and implementing it 'sometimes' had a higher difficulty estimate than the threshold between 'sometimes' and 'most of the time'. That said, the difference in latent trait parameter estimates was trivial in practice, with a correlation of 0.989 between the GRM and GPCM estimates.)

Again, an exploratory factor analysis prior to running the GRM analysis indicated scale unidimensionality. (The scree plot suggested a two-factor solution, but the first factor was clearly dominant, both in individual item loadings and in the proportion of variance explained.) Questionnaire items included in the GRM analysis, together with discrimination and difficulty parameters, are shown in Table 4.3. As with the PISA data, I used the latent trait parameter estimate as a measure of school-level teacher accountability. Higher scores correspond to more extensive teacher accountability instruments in the given school. The mean score for the TALIS 2013 teacher accountability instruments scale was 0.000, with a standard deviation of 0.905 (n=7,007 schools).[26]

With the TALIS 2013 GRM analysis shown in Table 4.3, it is less straightforward to compare the discrimination and difficultly parameters of different items than for the PISA 2PL analysis above. This is because each item with n ordinal categories has (n-1) difficulty parameters to mark the thresholds between each pair of categories, and the distribution of questionnaire responses across categories affects how informative an item may be. For example, the fact that 'Dismissal or non-renewal of contract' following a teacher appraisal has the smallest discrimination

[26] Again, these descriptive statistics differ slightly from those shown in Table 3.4 in the methodology chapter. The TALIS 2013 sample used for Table 3.4 excluded schools in Malaysia, Serbia, and the United States, for which there were survey administration issues that affect the representativeness of the overall survey dataset matched across schools and teachers (which are used in the main multilevel analysis), but not the quality of the principal-reported questionnaire data on their own (which were used in the IRT scale construction).

parameter despite having by far the largest difficulty parameters is likely due to the skewed distribution across its response categories. Only 3.6% of principals who gave valid responses to this item reported such dismissal or non-renewal happening most of the time, and 1.8% reported it happening always. This translates into large difficulty parameters, since only the schools with the most intensive teacher accountability approaches are likely to implement dismissal/non-renewal most of the time or always. It also contributes to the small discrimination parameter, since the relatively small number of schools responding 'most of the time' and 'always' means the model has less information to associate with these response categories.

Table 4.3 *Questionnaire items and IRT GRM parameters for TALIS 2013 teacher accountability*

Identifier	Questionnaire item	Discrim -ination	Difficulty			
	On average, how often is each teacher formally appraised in this school by the following people? *1='Never', 2='Less than once every year'; 3='Once every two years';* *4= 'Once per year'; 5='Twice or more per year'*		≥2	≥3	≥4	=5
TC2G27A	You, as principal	1.72	-1.56	-1.02	-0.69	0.79
TC2G27B	Other members of the school management team	2.37	-0.68	-0.45	-0.27	0.74
TC2G27C	Assigned mentors	1.55	-0.07	0.23	0.40	1.29
TC2G27D	Teachers (who are not part of the school management team)	1.51	-0.02	0.31	0.49	1.66
TC2G27E	External individuals or bodies	1.04	-0.62	0.63	1.04	2.41
	Please indicate the frequency that each of the following occurs in this school following a teacher appraisal. *1='Never', 2='Sometimes'; 3='Most of the time'; 4= 'Always'*		≥2	≥3	=4	
TC2G29A	Discussion of methods to remedy weaknesses in teaching	0.92	-5.01	-1.13	0.81	
TC2G29B	Development or training plan	1.16	-1.74	0.32	1.77	
TC2G29C	Material sanctions for poor performance	1.08	1.36	2.77	3.68	
TC2G29D	Appointment of a mentor to improve teaching	1.16	-0.98	0.95	2.14	
TC2G29E	Change in responsibilities (e.g. increase/decrease workload)	0.89	-0.95	2.51	4.63	
TC2G29F	Change in salary or payment of financial bonus	0.91	0.78	2.92	4.41	
TC2G29G	Change in likelihood of career advancement	0.78	-0.29	2.90	4.81	
TC2G29H	Dismissal or non-renewal of contract	0.40	-0.32	7.35	10.15	

Note. IRT GRM = item-response theory graded-response model. N = 7,007. Questionnaire item phrasing is from OECD (2014d).

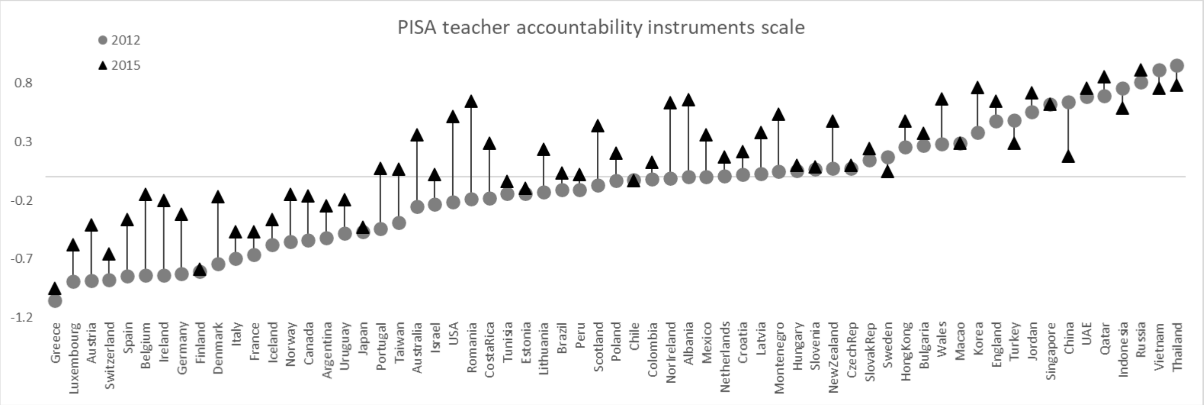
Nonetheless, the TALIS parameter estimates in Table 4.3 do align with the PISA estimates in Table 4.2 in that (a) the mode of teacher appraisal with the highest difficulty estimate is observations by inspectors or other external persons, and (b) the post-appraisal consequences with the highest difficulty estimates (i.e. dismissal or non-renewal of contract, a change in the likelihood of career advancement, a change in responsibilities, and financial or material consequences) are also the consequences that would be expected to have the largest effects on teachers' lives and livelihoods. Hence, this TALIS 2013 teacher accountability scale appears to have at least some construct validity, as with its PISA 2012 and 2015 counterpart.

These teacher accountability scale variables have some flaws, but they also have clear advantages. Weaknesses include the emphasis in the questionnaires on formal teacher accountability instruments, such that these scales leave out some forms of teacher accountability, such as teachers' interactions with parents. Furthermore, as with any statistical model for aggregating data, IRT incorporates assumptions that may not fully match the empirical data (e.g. IRT assumes that the latent trait is normally distributed, but the 2015 subset of the PISA teacher accountability latent trait estimates are somewhat left-skewed, though still within rules-of-thumb for normality). Moreover, these assumptions may not reflect lived reality (e.g. contrary to the notion of a latent trait for teacher accountability, implemented teacher accountability policy is not a manifestation of some shadowy, universal, bell-curved trait of 'accountability-ness').

Weaknesses notwithstanding, the parameter estimates for both the PISA and TALIS scales show some construct validity, as described above. Also, the scales appear to be reasonably consistent at the country level, with correlations of 0.86 between country-level weighted means for both (a) PISA 2012 and PISA 2015 (n=63 countries) and (b) PISA 2012 and TALIS 2013 (n=35 countries), and a correlation of 0.77 for (c) TALIS 2013 and PISA 2015 (n=34 countries).

That said, as shown in Figure 4.2, almost every country had a higher weighted mean teacher accountability score in PISA 2015 than in 2012. This rise in the extensiveness of teacher accountability instruments is evident not only in the scale scores, but also when comparing country-level weighted mean responses to the individual questionnaire items across the two survey cycles. Hence, the higher average teacher accountability scale scores in 2015 are not due to bias in the scale construction process. Instead, as far as can be determined from the principal-reported data, there were real increases in the average extent of teacher accountability instruments in these schools between 2012 and 2015. This growing intensity of accountability instruments in education has been observed in other studies, as noted in Section 1.1.

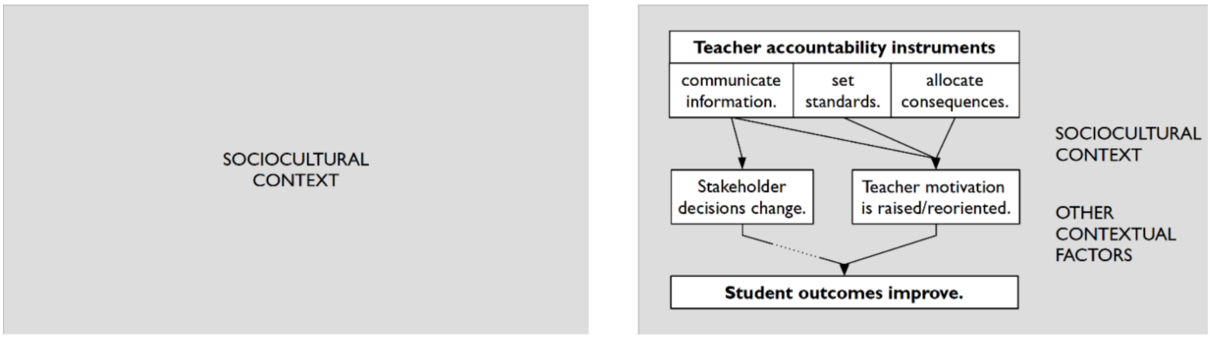
Figure 4.2 *PISA 2012 and 2015 teacher accountability scale scores, by country*



4.2 Measuring national sociocultural context: scale construction for social capital

Besides constructing scales for teacher accountability instruments, I also constructed scales for four elements of social capital. As detailed in Chapter 2, there are both theoretical and empirical reasons for expecting social capital and other sociocultural patterns to affect the efficacy of accountability instruments. To see whether this relationship holds for teacher accountability between countries, I draw on sociocultural data from two cross-country survey programmes. As noted in Section 3.3, I use two pre-existing scales from Hofstede's IBM dataset, as well as four scales related to social capital that I construct using data from the World Values Survey/European Values Study (WVS/EVS). In this section, I discuss the construction of these four WVS/EVS scales, as indicated in Figure 4.3. Descriptive statistics are available in Table 3.4 in the previous chapter.

Figure 4.3 *Relationship between Section 4.2 and the overall conceptual framework*


SOCIOCULTURAL CONTEXT
Teacher accountability instruments
communicate information.
set standards.
allocate consequences.
Stakeholder decisions change.
Teacher motivation is raised/reoriented.
Student outcomes improve.
SOCIOCULTURAL CONTEXT
OTHER CONTEXTUAL FACTORS

This section Overall conceptual framework

I derive the social capital scales from country-level averages for each questionnaire item, rather than using questionnaire data from individual respondents. This is partly because I am interested in sociocultural context at the national level, and also because the individual respondents in the WVS/EVS samples of these sociocultural surveys do not correspond to the respondents in the educational surveys. This approach to scale construction using country-level averages for each item aligns with Inglehart's (1997; Inglehart & Welzel, 2005) approach to aggregating WVS data.[27] For this process, I use factor analysis in SPSS 22.

As noted in Section 2.3, social capital has been defined in a number of different ways. Given the focus in this thesis on institutional relationships within teacher accountability, I am less interested in the forms of social capital that further an individual's opportunities for productivity (e.g. Bourdieu, 1986), and more so in the forms of social capital that concern groups. For example, Putnam (1995) defines social capital as 'features of social organisation such as networks, norms, and social trust that facilitate coordination and cooperation for mutual benefit' (p. 67).

Accordingly, I identified WVS/EVS questionnaire items that capture aspects of such group-based social capital. I found 25 such questionnaire items, which appeared in four clusters (which I describe below). Having identified these items, I ran an exploratory factor analysis of the country-level averages for each of these 25 items. As with the PISA accountability scale, I pooled the datasets across survey waves to increase scale reliability, such that each case was a country-by-wave value (e.g. Finland in EVS 4). This analysis indicated a four-factor solution in which the factor loadings aligned with each cluster of questionnaire items. (While the scree plot indicated a four-factor solution, Kaiser's criterion of retaining any factors with Eigenvalues greater than 1 suggested five factors. However, the fifth factor explained less than 3% of the variance. Moreover, it was not the primary factor for any of the variables, i.e. every variable loaded more heavily onto one of the first four factors.)

[27] For another approach to constructing scales for social trust and civic norms, Knack and Keefer (1997) construct (a) a scale for trust based on the proportion of respondents in each country who agree that 'most people can be trusted', and (b) a scale for civic norms by taking the sum of respondents' answers to whether it is never justifiable (=1) or always justifiable (=10) to breach each of five different norms. Knack and Keefer's approach has the virtue of being more straightforward than factor analysis. However, my factor analytic approach gives a better picture of the cross-country variation because each scale draws on more than one item (unlike their trust scale), and because factor analysis assigns greater weight to items better reflect the overall variability (unlike their uniformly weighted civic norms scale).

Given that these four aspects of social capital are somewhat correlated, I ran a separate analysis to extract each factor (rather than running a single analysis with all 25 items to obtain four orthogonal factors). Each set of variables had a high Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and rejected the null hypothesis in Bartlett's test of sphericity. An exception here is the two-item factor for social trust, which did reject the null in Bartlett's test, but had a KMO value of 0.500. Since this factor has only two component variables that load equally onto the factor, I could have constructed this scale simply from taking the sum of the standardised values of each country-level average. In fact, this sum correlates fully (r=1.000, for both Pearson's and Spearman's correlations) with the factor variable. However, for consistency with the other social capital variables, I retain the factor variable for social trust.

This resulted in four separate scales for different aspects of social capital. The first scale was derived from the set of questionnaire items that asked respondents how much confidence they have in a range of organisations, as shown in Table 4.4, along with factor loadings and model diagnostics. I call this the *confidence in institutions* scale. Next, the *civic networks* scale was derived from items asking respondents how many organisations they belong to (whether religious, political, professional, cultural, or leisure-related), as shown in Table 4.5. The *civic norms* scale was derived from items asking respondents about the degree to which certain self-interested but socially detrimental behaviours are justifiable, as shown in Table 4.6. Finally, the *social trust* scale draws on items about how trustworthy respondents believe most other people to be, as shown in Table 4.7. (For the 'confidence in institutions', 'civic networks', and 'civic norms' factors below, the variables included in the scale are not the full set of questionnaire items asked in each survey, but only the items that overlap across all three survey waves. For example, for the 'confidence in institutions' items, WVS respondents were asked about women's organisations, but EVS respondents were not; whereas EVS respondents were asked about the health care system, while WVS respondents were not; so these items were excluded.)

There are clear conceptual links between all of these elements of social capital—not least that higher levels of each element within a group could reasonably be expected to foster mutually beneficial cooperation, as in Putnam's definition. Conceptual links notwithstanding, high levels of one aspect of social capital may not correspond empirically to high levels of other such aspects (Newton, 2001). Hence the value of using separate scales to measure different aspects of social capital, rather than a single aggregate scale. For each of the factor variables, higher scores correspond to higher levels of social capital.

Table 4.4 *Questionnaire items and factor loadings for confidence in institutions*

WVS 6 identifier	Questionnaire item			Factor loadings
	How much confidence do you have in each of these organisations? *Country-level measure: weighted % of respondents answering 'A great deal' or 'Quite a lot'* *(Other options: 'Not very much', 'None at all')*			
V108	The churches/Religious organisations			0.181
V109	The armed forces			0.569
V110	The press			0.651
V112	Labour unions			0.869
V113	The police			0.614
V114	The courts			0.753
V115	The government (in your nation's capital)			0.855
V116	Political parties			0.840
V117	Parliament			0.919
V118	The civil service			0.879
V120	Major companies			0.735
V122	Environmental organisations			0.641
KMO value	.873	N (country-by-wave)	154	
Bartlett's χ^2 (df)	1591.615*** (66)	% of variance explained	54.0%	

Note. Questionnaire item phrasing is from EVS (2016b) and WVS Association (2012).
****p* < .001.

Table 4.5 *Questionnaire items and factor loadings for civic networks*

WVS 6 identifier	Questionnaire item			Factor loadings
	WVS: Could you tell me whether you are an active member, an inactive member, or not a member of each of these voluntary institutions? *Country-level measure: weighted % of respondents answering 'An active member' or 'An inactive member'* **EVS: Which of these voluntary organisations and activities do you belong to?** *Country-level measure: weighted % of respondents mentioning the organisation.*			
V25	Church or religious organisation			0.778
V26	Sport or recreational organisation			0.887
V27	Art, music or educational organisation			0.966
V28	Labour union			0.657
V29	Political party			0.787
V30	Environmental organisation			0.916
V31	Professional association			0.944
KMO value	0.869	N (country-by-wave)	158	
Bartlett's χ^2 (df)	1276.219*** (21)	% of variance explained	73.0%	

Note. Questionnaire item phrasing is from EVS (2016b) and WVS Association (2012).
****p* < .001.

Table 4.6 *Questionnaire items and factor loadings for civic norms*

WVS 6 identifier	Questionnaire item			Factor loadings
	Do you think each of the following actions can always be justified, never be justified, or something in between? *Country-level measure: weighted mean score (1–10 scale, where 1='Never justifiable' and 10='Always justifiable').*			
V198	Claiming government benefits to which you are not entitled			0.680
V199	Avoiding a fare on public transport			0.869
V201	Cheating on taxes if you have a chance			0.794
V202	Someone accepting a bribe in the course of their duties			0.831
KMO value	.802	N (country-by-wave)	158	
Bartlett's χ^2 (df)	319.164*** (6)	% of variance explained	63.5%	

Note. Questionnaire item phrasing is from EVS (2016b) and WVS Association (2012).
****p* < .001.

Table 4.7 *Questionnaire items and factor loadings for social trust*

WVS 6 identifier	Questionnaire item			Factor loadings
V24	Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people? *Country-level measure: weighted % of respondents answering 'Most people can be trusted.'*			0.784
V56	Do you think most people would try to take advantage of you if they got a chance, or would they try to be fair? *Country-level measure: weighted mean score (1–10 scale, where 1='People would try to take advantage of you' and 10='People would try to be fair').*			0.784
KMO value	0.500	N (country-by-wave)	157	
Bartlett's χ^2 (df)	73.776*** (1)	% of variance explained	61.5%	

Note. Questionnaire item phrasing is from EVS (2016b) and WVS Association (2012).
****p* < .001.

For each country, I match social capital scales from the most recent available WVS/EVS wave to each ILSA, as long as the WVS/EVS data were collected prior to the ILSA. This requires the strong assumption that the sociocultural scales are reasonably stable in between the earliest used WVS/EVS survey and the ILSA survey, such that different time lags for different countries do not compromise the analysis (see Section 3.3). As shown in Figure 4.4, some countries vary considerably on these social capital scales across survey waves. On the whole, however, there is fairly good agreement across survey waves. This is particularly visible for social trust. As shown in Table 4.8, there are moderate to strong correlations for most constructs across the survey cycles, with one exception that is likely affected by the small number of overlapping countries (i.e. confidence in institutions between EVS 4 and WVS 6).

Figure 4.4 *WVS and EVS social capital scale scores, by country*

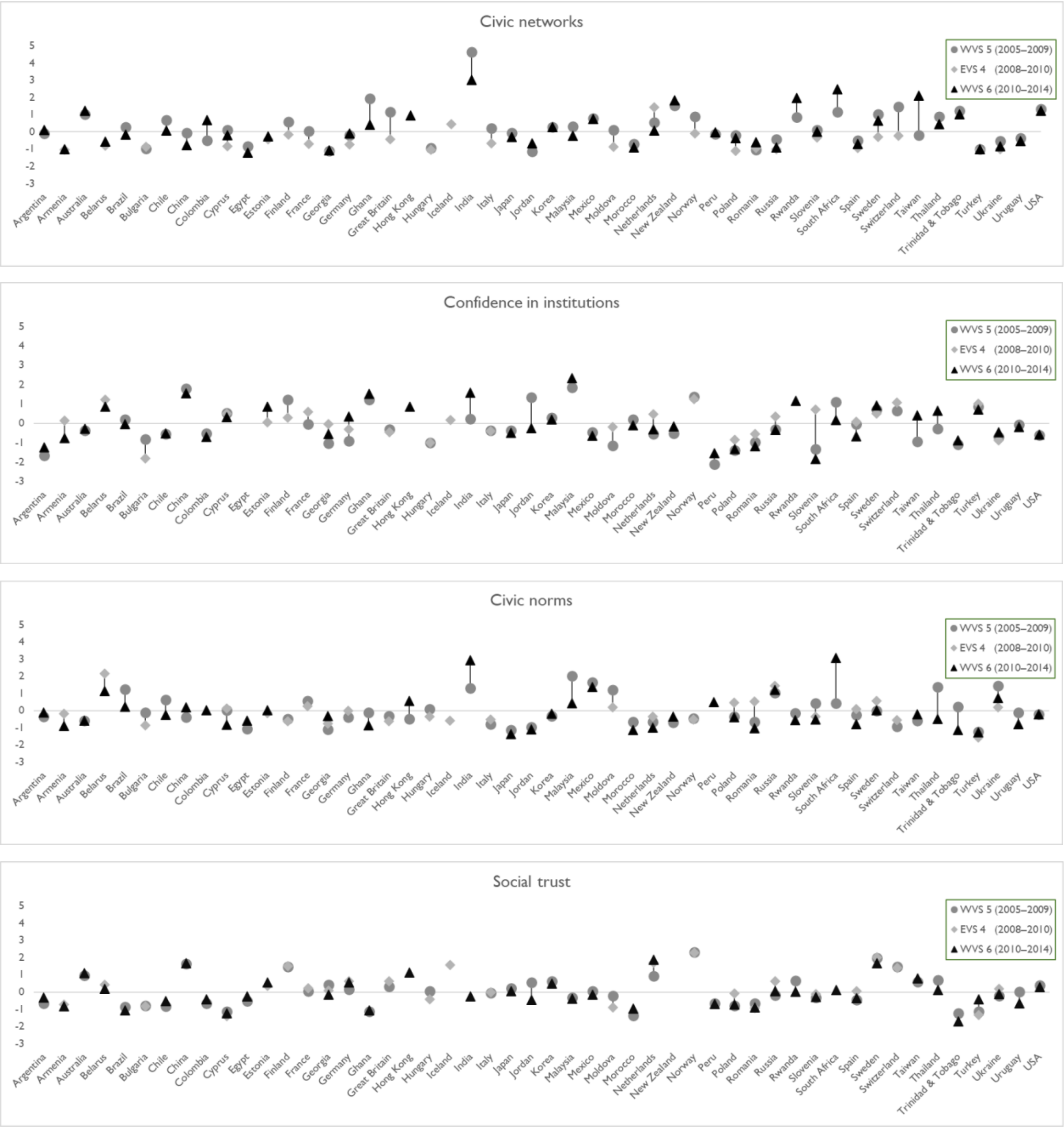


Table 4.8 *Pairwise correlations (and number of countries) between country-level scale scores for WVS/EVS social capital constructs*

	WVS 6 & WVS 5	WVS 6 & EVS 4	EVS 4 & WVS 5
Confidence in institutions	.804** (34)	.441 (15)	.655** (21)
Civic networks	.783** (36)	.608* (15)	.624** (21)
Civic norms	.591** (34)	.735** (14)	.616** (21)
Social trust	.883** (33)	.919** (15)	.923** (21)

**p* < 0.05. ***p* < 0.01. (two-tailed)

Although the correlations across survey cycles for each construct in Table 4.8 are far from perfect, I consider them adequate. Across sociocultural constructs, the average within-country difference between any pair of survey cycles (e.g. Argentina's social trust score in WVS 6 minus its social trust score in WVS 5) is -0.070, or approximately 7% of a standard deviation. Importantly, there were no systematic patterns in the magnitude of change across cycles. For example, if the accuracy of these sociocultural scales consistently worsened over time, we would expect the average within-country change in scale scores to increase in proportion to the time lag in between survey cycles, with the largest average changes appearing between WVS 5 and WVS 6. However, no such pattern appears. Accordingly, I do not anticipate any systematic bias from matching country-level sociocultural scores from different WVS/EVS cycles to the same ILSA cycle.[28]

4.3 PISA 2015 analysis: Teacher accountability, student outcomes, and moderating variables

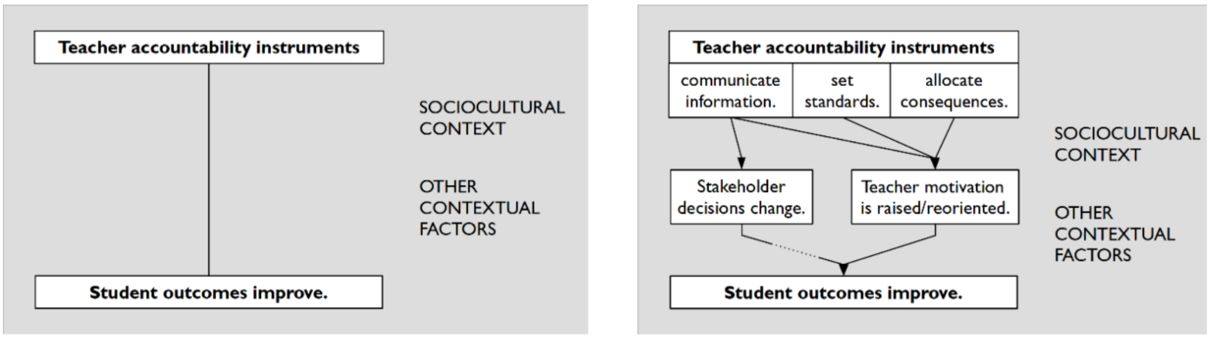
Having constructed scales for social capital and teacher accountability, I use the cross-country survey datasets to investigate RQ1, *To what extent does the influence of teacher accountability instruments on student outcomes depend on sociocultural context?* Thus, as indicated in Figure 4.5, in this section I investigate whether the ILSA and sociocultural survey data offer empirical evidence linking the starting point and endpoint of my conceptual framework (i.e. teacher accountability instruments and student outcomes, respectively), with sociocultural context.

As detailed in Section 3.3, I address this question using model 1, a multilevel moderation model. In model 1, the outcome variable is pupil proficiency scores. The main independent variables are (1) teacher accountability, both the country-level weighted mean and the school-level differential;

[28] I tested the possibility of such bias by re-estimating the model 1 civic norms regressions for PISA 2015 (which are presented in Section 4.3 below) using different cuts of the data. Specifically, I estimated the models using civic norms scores (a) only from WVS 6 (n=32 countries); (b) only from EVS 4 (n=39); (c) only from WVS 5 (n=40); (d) from WVS 6, but only for countries that participated in both WVS 5 and WVS 6 (n=25); and (e) from WVS 5, but only for countries that participated in both WVS 5 and WVS 6 (n=25). In all but one of these regressions, the coefficient on the interaction term of interest—i.e. the interaction between civic norms and country-level teacher accountability instruments—was similar in direction, magnitude, and significance to the main PISA 2015 regression that drew on all three WVS/EVS cycles for civic norms scores across countries (n=57). The sole exception was regression (c) with civic norms scores only from WVS 5, where the coefficient estimate was smaller and the standard error was larger than in the main regression, such that coefficient was insignificant. However, it seems likely that this may be a fluke due to the distribution of countries that happened to be included, rather than to an attenuation of WVS 5 accuracy over time. The reason for this claim is that regressions (d) and (e)—both of which included the same 25 countries, but differed in that their civic norms scores were either from WVS 6 and or from WVS 5, respectively—are comparable across all key variables in the size and significance of their coefficient estimates, as well as in the proportion of variance explained at the country, school, and student levels. For the main model 1 results and other sensitivity checks, see the next section.

and (2) sociocultural context, represented using six different sociocultural constructs. As described in Section 3.3, the model also includes control variables at each level of analysis, along with interaction terms to capture any moderation of the relationship between teacher accountability instruments and pupil proficiency by sociocultural context and by other contextual characteristics represented in the control variables. The main dataset for this analysis is PISA 2015 with pupil science proficiency as the outcome variable.

Figure 4.5 *Relationship between Section 4.3 and the overall conceptual framework*


Teacher accountability instruments
SOCIOCULTURAL CONTEXT
OTHER CONTEXTUAL FACTORS
Student outcomes improve.
Teacher accountability instruments
communicate information.
set standards.
allocate consequences.
Stakeholder decisions change.
Teacher motivation is raised/reoriented.
SOCIOCULTURAL CONTEXT
OTHER CONTEXTUAL FACTORS
Student outcomes improve.

This section

Overall conceptual framework

Table 4.9 shows results for five nested models for PISA 2015 science proficiency, culminating in the full model 1. Column (a) shows a model without any predictors, to indicate how the variance in pupil science scores is distributed across levels. Column (b) introduces both of the teacher accountability terms. Notably, neither the school- nor country-level teacher accountability terms have a significant unmoderated effect on science proficiency. These two accountability terms are, again, insignificant in column (c), which includes control variables for pupil economic, social, and cultural status (ESCS), school autonomy, and national GDP per capita as well as the interaction terms associated with each of these controls. Neither are they significant in column (d), which adds in the six country-level sociocultural constructs. In column (e), which introduces the interactions between these sociocultural constructs and the accountability terms, the country-level accountability term remains insignificant, whereas the school-level accountability differential is significant at the 5-percent level, though small in magnitude. Overall, the full model in column (e) accounted for 75% of all the between-country variation. That is, moving from the null model in column (a) to the full model in column (e) reduced the unexplained country-level variance in the dataset from 1980.19 to 500.19.

Table 4.9 *Model 1: results for nested models for PISA 2015 science proficiency*

Y_{psc} = science proficiency	(a)	(b)	(c)	(d)	(e)
Pupil level					
Constant_{psc}	465.42 ** (5.92)	466.37 ** (5.73)	441.48 ** (8.37)	437.72 ** (8.33)	437.23 ** (9.64)
ESCS_{psc}			17.65 ** (1.11)	17.65 ** (1.11)	17.65 ** (1.11)
School level					
$\text{School autonomy}_{sc}$			39.00 ** (7.61)	39.00 ** (7.59)	39.05 ** (7.58)
$(\text{Accountability}–\overline{\text{Accountability}})_{sc}$		2.28 (2.08)	4.22 (2.97)	4.22 (2.97)	6.18 * (3.11)
Country level					
$\overline{\text{Accountability}}_{c}$		-10.10 (13.26)	7.08 (14.52)	6.30 (13.86)	0.03 (15.87)
GDP_{c}			12.35 ** (2.62)	8.32 ** (2.63)	10.96 ** (3.29)
$\text{Confidence in institutions}_{c}$				8.51 (5.72)	0.05 (7.44)
$\text{Civic networks}_{c}$				-7.16 (5.38)	-5.41 (6.88)
Civic norms_{c}				4.51 (5.22)	11.21 * (5.18)
Social trust_{c}				19.18 ** (5.68)	17.72 ** (6.20)
$\text{Power distance}_{c}$				1.25 (5.19)	2.29 (6.40)
$\text{Uncertainty avoidance}_{c}$				7.26 (4.76)	3.98 (6.08)
Interactions with $(\text{Accountability}–\overline{\text{Accountability}})_{sc}$					
$^*\text{ESCS}_{psc}$			0.12 (0.41)	0.12 (0.41)	0.15 (0.42)
$^*\text{School Autonomy}_{sc}$			-2.53 (3.84)	-2.53 (3.85)	-3.30 (4.24)
$^*\text{GDP}_{c}$			-0.74 (0.74)	-0.74 (0.74)	-0.16 (0.88)
$^*\text{Confidence in institutions}_{c}$					1.07 (1.64)
$^*\text{Civic networks}_{c}$					-0.65 (1.50)
$^*\text{Civic norms}_{c}$					-1.72 (2.27)
$^*\text{Social trust}_{c}$					-1.01 (2.55)
$^*\text{Power distance}_{c}$					1.49 (2.02)
$^*\text{Uncertainty avoidance}_{c}$					-2.03 (2.00)
Interactions with $\overline{\text{Accountability}}_{c}$					
$^*\text{ESCS}_{psc}$			-5.75 * (2.50)	-5.73 * (2.50)	-5.74 * (2.50)
$^*\text{School Autonomy}_{sc}$			-5.05 (12.18)	-4.63 (12.15)	-4.56 (12.11)
$^*\text{GDP}_{c}$			9.20 (5.91)	17.23 ** (3.13)	9.85 * (4.59)
$^*\text{Confidence in institutions}_{c}$					19.99 (14.14)
$^*\text{Civic networks}_{c}$					-18.32 (13.28)
$^*\text{Civic norms}_{c}$					-28.74 ** (10.92)
$^*\text{Social trust}_{c}$					20.04 (17.69)
$^*\text{Power distance}_{c}$					-4.24 (10.39)
$^*\text{Uncertainty avoidance}_{c}$					5.43 (10.78)
Variance parameters					
Pupil	5 687.48	5 687.49	5 539.71	5 539.78	5 539.70
School	2 786.22	2 780.64	2 080.03	2 079.23	2 077.56
Country	1 980.19	1 958.61	918.34	626.43	500.19
-2*loglikelihood	4 148 699.83	4 148 676.47	4 136 436.90	4 136 415.49	4 136 388.82
Likelihood ratio test (df)	—	23.36** (2)	12 239.57** (11)	21.41** (6)	26.67** (12)
VPC (country-level)	0.19	0.19	0.11	0.08	0.06

Note. ESCS=economic, social, and cultural status. VPC=variance partition coefficient. N(pupils)=346 726; N(schools)=12 764; N(countries)=57.
$^*p<0.05$; $^{**}p<0.01$.

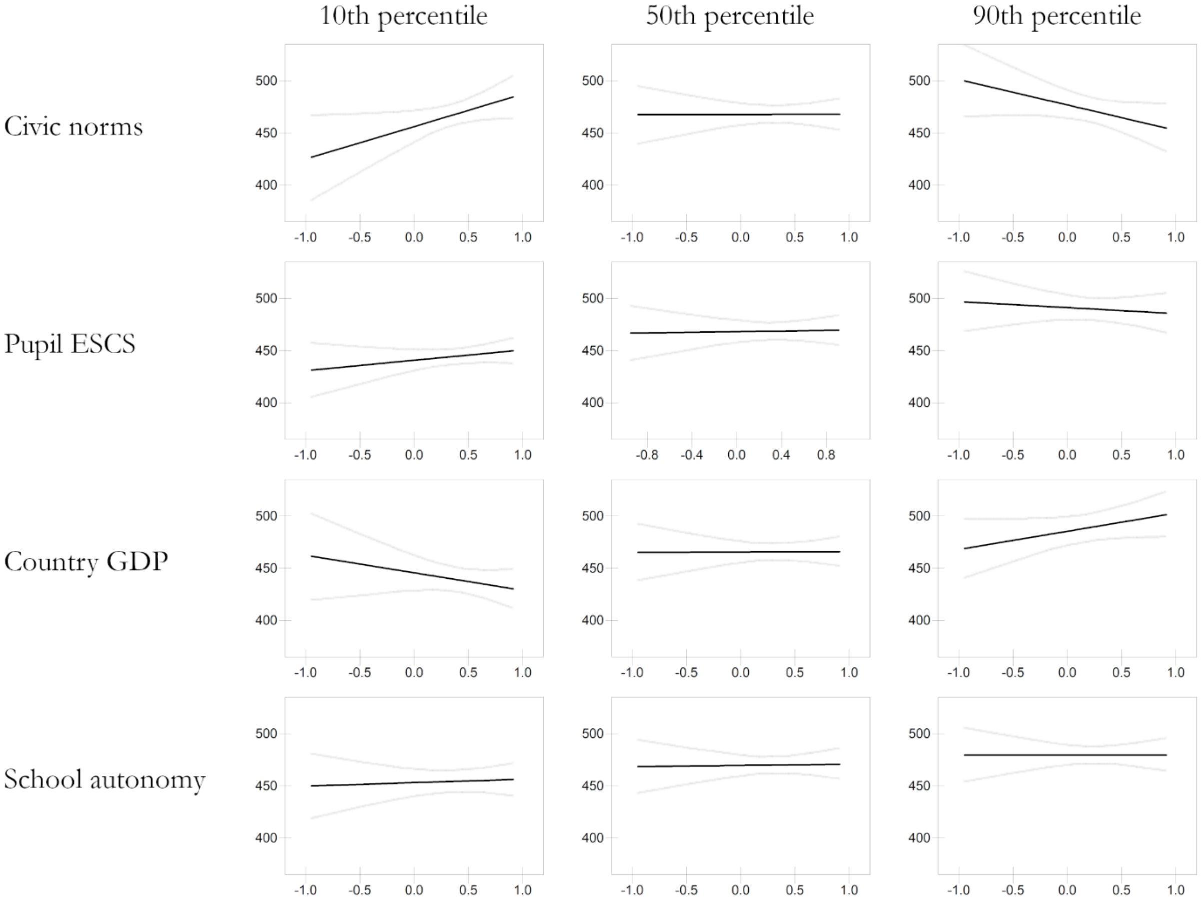
As for the moderation effects, none of the interaction terms for the school-level accountability differential are significant. However, the relationship between science proficiency and the country-level weighted mean of teacher accountability is dependent on context, as seen in column (e) in the significant coefficients of the interactions between $\overline{\text{Accountability}}$ and some contextual moderators. This indicates that between-country differences in teacher accountability matter more to student outcomes than between-school differences, at least in the PISA 2015 data.

As for the interaction terms, three were significant: $\overline{\text{Accountability}}$*ESCS, $\overline{\text{Accountability}}$*GDP, and $\overline{\text{Accountability}}$*civic norms. Among these significant estimates, the most pertinent to the research question is that civic norms significantly and substantially moderates the relationship between teacher accountability and science proficiency. Crucially, these moderation effects are present despite controlling for moderation from pupil socioeconomic status, school autonomy, and national per-capita GDP. Besides the large and significant $\overline{\text{Accountability}}$*civic norms interaction, parameter estimates for the country-level interaction with the three other social capital scales—confidence in institutions, civic networks, and social trust—are fairly large, though insignificant. (Social trust also has a significant unmoderated relationship with pupil science proficiency, as seen in columns (d) and (e). However, given space constraints and the focus in this project on teacher accountability policy, I am less interested in this unmoderated effect of social trust and more so in whether social trust moderates the relationship between teacher accountability instruments and student outcomes.)

To determine the relative importance of the predictors, I re-estimated the full model in column (e) with standardised predictors. In this model, parameter estimates for four of the non-interacted predictors were sizable: 19.5 for GDP, 19.1 for ESCS, 17.5 for social trust, and 9.4 for school autonomy. Civic norms had the fifth largest parameter estimate among non-interacted predictors, at 5.8. For comparison, PISA scores are scaled to have a standard deviation of approximately 100 points (OECD, 2016d, p. 58). Five of the interaction terms were noticeably larger than the other interactions. These were the interactions between $\overline{\text{Accountability}}$ and, respectively, civic norms (–7.9), confidence in institutions (7.6), social trust (7.3), GDP (6.6), and civic networks (–6.1), in order of decreasing magnitude. Among the three variables that significantly moderated the effects of teacher accountability on student outcomes, civic norms had a smaller unmoderated effect on science proficiency than ESCS or GDP. However, the civic norms interaction term not only had the largest magnitude (-7.9), but also the smallest p-value (0.004).

Figure 4.6 illustrates these moderation relationships by showing predicted science proficiency scores against $\overline{\text{Accountability}}$ 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 (1) all of the parameter estimates in column (e) of Table 4.9, including both the non-interacted and interacted terms; as well as (2) the empirical values of each explanatory variable, represented by the range of $\overline{\text{Accountability}}$, the various percentiles of the contextual moderator of interest in each row, and the mean values of all the other variables. The figure shows predictions for different levels of the three contextual variables that significantly moderated the relationship between $\overline{\text{Accountability}}$ 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 interact with $\overline{\text{Accountability}}$ significantly.

Figure 4.6 *Predicted PISA 2015 science proficiency scores (and 95% confidence intervals) against $\overline{\text{Accountability}}$, for the 10th, 50th, and 90th percentiles of each contextual moderator*



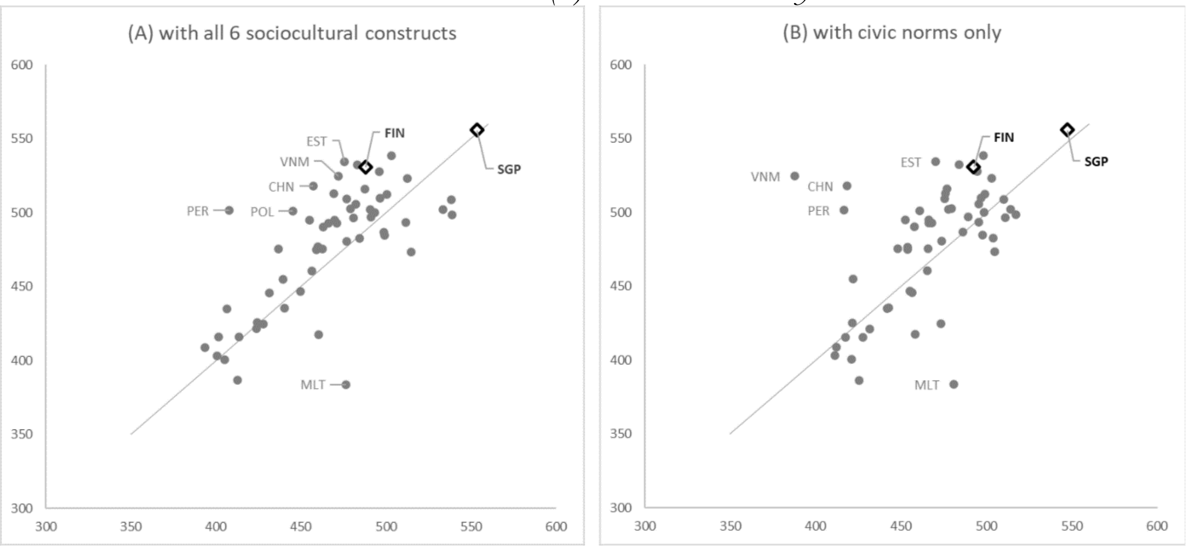
Note. All predictions are based on the regression in column (e) of Table 4.9. Each row shows predicted science proficiency scores against $\overline{\text{Accountability}}$ for the 10th, 50th, and 90th percentile of the named contextual predictor. All other variables are held constant at their means. Rows are sorted by decreasing magnitude of the p-value on the coefficient of corresponding interaction term.

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 $\overline{\text{Accountability}}$ 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 $\overline{\text{Accountability}}$ have no effect of predicted science proficiency. At the 90th percentile of civic norms, pupil science proficiency is expected to decrease as $\overline{\text{Accountability}}$ increases. In contrast, the relationship between $\overline{\text{Accountability}}$ and science proficiency is not affected by school autonomy, since the predicted score plots are similarly flat for all three levels of school autonomy. (However, school autonomy does have an independent effect on pupil science proficiency, as indicated by the different intercepts in the three plots and by the significant coefficient on the unmoderated school autonomy variable in Table 4.9.) Interestingly, for an average student in an average school in a hypothetical average country—that is, when all independent variables except $\overline{\text{Accountability}}$ are held constant at their means (or medians, for the focal control variable in each row)—the level of teacher accountability instruments makes little difference to student outcomes, as shown in the flat slopes in all of the plots in the 50th percentile column.

The civic norms plots also indicate the degree to which sociocultural context can affect the relationship between teacher accountability and student outcomes. Comparing the rightmost ends of each plot—that is, at the 100th percentile of $\overline{\text{Accountability}}$—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 $\overline{\text{Accountability}}$ corresponds to Russia, where the typical school had 13 out of the 14 the teacher accountability instruments in the PISA 2015 questionnaires.) Conversely, at the 0th percentile of $\overline{\text{Accountability}}$ shown at the leftmost ends of each plot (which corresponds to Greece, where the typical school had approximately half 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. Given that PISA proficiency scores are scaled to have a standard deviation of approximately 100 points, as noted above, these differences are considerable.

To examine how well model 1 predicts actual PISA 2015 science scores, Figure 4.7 plots the official OECD-calculated country-level science scores against predicted scores from (A) model 1, which corresponds to column (e) of Table 4.9, and (B) a version of model 1 that drops all sociocultural variables and all interaction terms involving sociocultural variables except for civic norms and its interactions. Both models perform fairly well, with predictions for all but six countries falling within 50 points (≈0.5 SD) of their actual scores in (A). The same is true for all but five countries in (B). Unsurprisingly, the deviations for some of these outlying countries were considerably larger for (B), since it was the more parsimonious model. Still, the average country-level deviations for both countries were similar, at 24.61 points for (A) and 26.87 points for (B).[29] For the countries where I conducted fieldwork, both models performed very well for Singapore, with deviations for 1.80 points for (A) and 7.99 points for (B); but worse than average for Finland, with deviations of 42.43 points for (A) and 37.99 points for (B).

Figure 4.7 *Actual vs. predicted PISA 2015 science proficiency scores from model 1 (A) with all 6 sociocultural constructs and (B) with civic norms only*



Note. Predicted values in (A) are drawn from the regression model shown in column (e) of Table 4.9. Predicted values in (B) are from a similar model that drops all other sociocultural variables and interaction terms involving sociocultural variables other than civic norms. Data labels indicate Finland (FIN) and Singapore (SIN), as well as countries that over- or underperformed by more than 50 points: China (CHN), Spain (ESP), Estonia (EST), Malta (MLT), Peru (PER), Poland (POL), and Vietnam (VNM). Source for the actual PISA 2015 scores: OECD (2016d).

[29] Note that deviations from the y=x line in these plots are *not* country-level regression residuals, since the y-values here are official country weighted averages from the full PISA dataset, rather than the country weighted averages from the sample used in this analysis (which only includes cases that had complete observations for all predictors in the model). As for the actual residual plots for the full model (i.e. the regression shown in column (e) of Table 4.9), country-level Q-Q plots and standardised residual plots did not indicate any obvious departures from either normality or homoscedasticity. However, they did indicate (1) that China was an outlier, and (2) that there was a cluster of high-GDP countries that were underperforming relative to the model. Adding a dummy variable for China and removing all GDP-related terms from the model eliminated these irregularities from the diagnostic plots, while almost doubling the magnitude of the coefficient of the Accountability*civic norms interaction term, with a slight reduction in its standard error. Hence, it appears that any deviations in model 1 caused by (1) and (2) attenuate rather than falsely inflating the moderation effect of civic norms.

It is worth noting that different aspects of social capital moderate the relationship between teacher accountability and pupil science scores in different directions. While the coefficients of the interaction terms for civic norms and civic networks are negative (i.e. more extensive teacher accountability is associated with better student outcomes in countries with weaker adherence to civic norms and less extensive civic networks), the coefficients of the confidence in institutions and social trust interactions are positive (i.e. more extensive teacher accountability is associated with better student outcomes in countries with greater confidence in institutions and higher social trust). While these results should be treated with caution, given that only the civic norms interaction term is significant, they suggest that national sociocultural context may affect the implementation of teacher accountability instruments in multiple cross-cutting ways. I discuss some aspects of this complexity in Chapter 6, using interview data from Finland and Singapore.

Sensitivity checks

To test the robustness of these results, I re-estimated model 1 using different cuts of the PISA 2015 data, and also using PISA 2012 and TIMSS 2015 data. The PISA 2015 and 2012 regressions use the same functional form as in column (e) of Table 4.9, with all six sociocultural constructs included concurrently. However, the TIMSS 2015 model 1 regressions with all six sociocultural constructs showed evidence of multicollinearity due to too many country-level variables relative to the country sample size, as described in the 'Modelling' subsection of Section 3.3, and as demonstrated in Appendix A. Accordingly, I ran six separate models for each TIMSS 2015 sample, where each model tested one of the six sociocultural constructs, in turn. Table 4.10 summarises the significance and direction of parameter estimates for the interaction between country-level teacher accountability and each of the six sociocultural constructs.

As shown in Table 4.10, the interaction between $\overline{\text{Accountability}}$ and civic norms is consistently negative and significant for all of the PISA 2015, PISA 2012, and TIMSS 2015 models tested. These include the PISA 2015 and TIMSS 2015 datasets matched with teacher accountability data from PISA 2012 in order to test for potential time lags in the effects of accountability instruments; as well as the OECD-only subsamples for both PISA cycles and the PISA 2015 subsample with only publicly funded schools. None of the other sociocultural constructs consistently moderated the relationship between teacher accountability instruments and student outcomes.

Table 4.10 *Summary of sociocultural constructs that significantly moderate the relationship between Accountability 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

Note. 1 symbol (+/–) indicates p<.05; 2 symbols (+ +/– –) indicate p<.01. Full results can be provided upon request.

For a more detailed look, Table 4.11 shows parameter estimates from the sensitivity checks for the main effect of Accountability as well as its interactions with civic norms and the non-sociocultural moderators. As with the main PISA 2015 regressions, the unmoderated effect of Accountability is insignificant in all the models. However, many of the interactions between Accountability and the contextual moderators are significant, most consistently with civic norms. The consistency of the significant interaction between Accountability and civic norms across different test cycles and cuts of the data suggests that this aspect of national sociocultural context does, in fact, moderate the relationship between student outcomes and teacher accountability instruments.

For further sensitivity checks, I also ran models with each sociocultural construct entered singly (rather than all six constructs concurrently) for each cut of the PISA 2015 and 2012 data. These results, which are not shown in the tables, were broadly consistent with those shown in Table 4.10, with four exceptions, one of which related to civic norms. Specifically, for the PISA 2012 full sample, the interaction between Accountability and civic norms was insignificant, unlike in the model with all six sociocultural constructs, although it was similarly negative.[30]

[30] The other three differences between the models with all six sociocultural constructs shown in Table 4.10 and the models with each sociocultural constructs entered singly were: (1) for the PISA 2015 data matched with Accountability from PISA 2012, the interaction between Accountability and confidence in institutions was significant and positive, in contrast to the insignificant coefficient in the model with all six sociocultural constructs; and for the PISA 2012 OECD-only sample, the interactions between Accountability and (2) civic networks and (3) social trust were insignificant, unlike the significant interactions in the model with all six sociocultural constructs.

Table 4.11 *Model 1: sensitivity checks (showing selected variables only)*

PISA 2015	*With all six sociocultural constructs in the model*							
Y_{psc} = science proficiency	Full sample		OECD only		Public schools only		With *Accountability* from PISA 2012	
$\overline{\text{Accountability}}_c$	0.03	(15.87)	-22.02	(18.45)	2.80	(17.13)	-5.50	(10.19)
$\overline{\text{Accountability}}_c$ *ESCS$_{psc}$	-5.74*	(2.50)	-2.29	(3.03)	-3.76	(2.29)	-7.16**	(1.81)
$\overline{\text{Accountability}}_c$ *School autonomy$_{sc}$	-4.56	(12.11)	20.79	(16.59)	2.75	(10.57)	-6.99	(7.43)
$\overline{\text{Accountability}}_c$ *GDP$_c$	9.85*	(4.59)	36.48**	(12.99)	5.24	(5.19)	8.65*	(3.66)
$\overline{\text{Accountability}}_c$ *Civic norms$_c$	-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	
PISA 2012	*With all six sociocultural constructs in the model*							
Y_{psc} = mathematics proficiency	Full sample		OECD only					
$\overline{\text{Accountability}}_c$	-24.98	(16.59)	-21.12	(19.43)				
$\overline{\text{Accountability}}_c$ *ESCS$_{psc}$	-6.57**	(2.22)	-5.50	(4.41)				
$\overline{\text{Accountability}}_c$ *School autonomy$_{sc}$	-5.51*	(2.29)	-5.66	(4.21)				
$\overline{\text{Accountability}}_c$ *GDP$_c$	12.69	(7.79)	21.56**	(8.04)				
$\overline{\text{Accountability}}_c$ *Civic norms$_c$	-16.52*	(8.10)	-29.39*	(12.02)				
N Pupils	375 207		275 715					
Schools	14 840		11 169					
Countries	52		36					
TIMSS 2015	*With only civic norms in the model*							
Y_{ptsc} = mathematics proficiency	With *Accountability* from PISA 2015		With *Accountability* from PISA 2012					
$\overline{\text{Accountability}}_c$	26.59	(15.98)	-3.17	(13.93)				
$\overline{\text{Accountability}}_c$ *Home resources$_{ptsc}$	-12.59**	(3.67)	-11.99**	(2.75)				
$\overline{\text{Accountability}}_c$ *Teaching experience$_{tsc}$	0.13	(0.11)	0.06	(0.11)				
$\overline{\text{Accountability}}_c$ *School resources$_{sc}$	-2.20	(4.23)	-6.83	(3.91)				
$\overline{\text{Accountability}}_c$ *GDP$_c$	21.57*	(10.55)	23.59**	(8.58)				
$\overline{\text{Accountability}}_c$ *Civic norms$_c$	-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					

Note. ESCS = economic, social, and cultural status. Full results can be provided upon request.
*$p < 0.05$. **$p < 0.01$.

To increase the country sample size, I also ran regressions using separate cuts of the data for each of the six sociocultural constructs; such that a country that had WVS/EVS data but lacked Hofstede data would appear in the regression for, say, civic norms, but not those for power distance. This raised the maximum sample size from 57 to 64 countries (for PISA 2015 with WVS/EVS sociocultural data), and the minimum sample size from 22 to 24 countries (for TIMSS 2015 with WVS/EVS sociocultural data and accountability data from PISA 2012). Results for civic norms were broadly consistent with those shown in Table 4.11, again with the exception that the Accountability*civic norms interaction was insignificant (though also negative) for the single-sociocultural-construct model with the full sample of PISA 2012 (n=56 countries).

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 found evidence for moderation by only one of these six constructs, i.e. civic norms. Still, the three other constructs related to social capital—i.e. confidence in institutions, social trust, and civic networks—also had sizable though insignificant interactions with country-level teacher accountability in the main dataset. Moreover, the civic norms interaction terms were robust across different assessment cycles and subsamples of the data, with very minor exceptions.

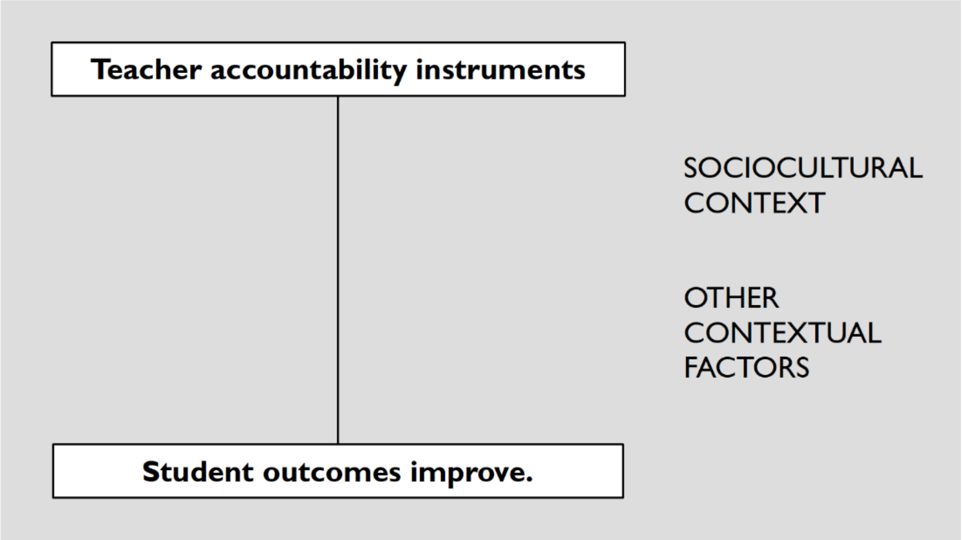
4.4 Discussion

To the extent that we want to use cross-sectional analyses of international large-scale assessments as sources of policy evidence, the results in the previous section suggest that we cannot overlook the influence of national sociocultural context on the efficacy of teacher accountability instruments. Although the interaction between country-level teacher accountability and national sociocultural context was only significant for one of the six sociocultural constructs in the main model 1 regression, these results were consistent across different subsamples of PISA 2015, PISA 2012, and TIMSS 2015 data. Moreover, the magnitude of these effects was large, with a divergence of 73 points in PISA 2015 science scores between the 10th and 90th percentiles of civic norms at the lowest levels of teacher accountability. For comparison, the PISA science score was scaled to have a standard deviation of approximately 100 points, and the PISA 2015 results report interprets 30 score points as being approximately equal to one year of schooling (OECD, 2016d, pp. 58, 65).

While this result shares the weaknesses of other cross-sectional ILSA analyses—such as non-causal conclusions and measurement errors—it also shares their strengths. A key advantage of these analyses is scale. My analysis shows the existence of a relationship across hundreds of thousands of students in dozens of countries. I am not denying the numerous measurement and modelling assumptions required for comparisons at this scale, nor the fact that this probabilistic relationship does not uniformly affect all students in all countries. Nonetheless, the scope of these datasets and the magnitude of the interaction effects are a persuasive piece of suggestive evidence for the influence of sociocultural context on teacher accountability. This evidence is especially noteworthy because it is based on the same ILSA datasets that fuel the preoccupation with acontextual best practices in education, as noted in Section 1.1.

Thus, as shown in Figure 4.8, the statistical analysis in this chapter gives suggestive evidence that (a) there is a relationship between teacher accountability instruments and student outcomes; but (b) this relationship varies with sociocultural context. These findings cohere with the systematic literature search presented in Section 2.3. Studies from this search found that sociocultural context can influence various teacher accountability processes, from policy formulation (e.g. Hopmann, 2008; Mattei, 2012; Osborn, 2006) to implementation (e.g. Broekman, 2016; Mizel, 2009; Narwana, 2015). Furthermore, the specific finding that teacher accountability instruments are associated with lower student outcomes in countries with strong adherence to civic norms accords with arguments from other disciplines that some constructive and altruistic behaviours may be undermined by certain forms of accountability (e.g. Deming, 1993 in management; O'Neill, 2002 in philosophy; Stout, 2010 in legal studies).

Figure 4.8 *Evidence for the conceptual framework from the RQ1 analysis*



Still, as noted above, this analysis uses cross-sectional data, so these relationships are correlational rather than causal. This is reflected in Figure 4.8 with a line—rather than an arrow—connecting accountability instruments and student outcomes. Other features of these results also indicate a need for caution. For one, the significant interaction terms had large standard errors. Also, given that the analysis did not include a causal identification strategy, it is possible that all these relationships may be spurious.

As noted in Section 3.1, one important way of strengthening the validity of an argument, especially in realist research, is by weighing it against alternative explanations. One alternative explanation for the results presented in this chapter is that the association between teacher accountability instruments, civic norms, and student outcomes is simply a matter of happenstance—that there is no causal pathway linking (context-compatible) teacher accountability instruments to student outcomes. This all the more possible given that the largest country sample had just 64 countries. Not only is this a relatively small regression sample, but it also constitutes less than one-third of all sovereign states globally. Another possible explanation is that there are confounding variables. Rather than sociocultural context, the differential effects of teacher accountability instruments may be due to capacity constraints, or to the internal coherence (see Pritchett, 2015, for a theoretical argument) or implementation quality (see Dee & Dizon-Ross, 2019, for an empirical case) of any given set of instruments.

However, based on the field interviews I conducted in Finland and Singapore, as well as prior studies reviewed thus far, I believe the most plausible explanation is that teacher accountability instruments can affect student outcomes, and that these effects are influenced by sociocultural context (among other factors). To address the first alternative explanation, i.e. that the association between teacher accountability instruments and student outcomes is coincidental rather than reflecting a real causal pathway, recall the evaluation studies listed in Table 4.1. These studies found that certain teacher accountability instruments can play positive causal role in educational outcomes, given the right combinations of instrument and context. Some of these effects were small, and all of them were limited to particular intervention designs and contexts, but they offer evidence that such a causal pathway is possible. Besides this evidence of statistical causality, my field interviews offer evidence of realist causality along one step of the causal pathway, i.e. that teacher accountability instruments can generate real change in teacher motivation. I discuss this evidence in the next chapter.

The second alternative explanation does not deny the existence of this causal pathway. Rather, it argues that the key contextual moderators come from some non-sociocultural aspect of context that is correlated with civic norms. However, it is worth noting that the model 1 regressions already control for some variables that would be expected to show associations with both sociocultural context and teacher accountability, i.e. student socioeconomic background, school autonomy (or school resources for TIMSS), and national GDP. As for the possibility that the significant sociocultural moderators are inadvertently proxying for some aspect of policy design or implementation quality, the matrix in Table 4.12 offers some suggestive evidence to the contrary. In this matrix, I summarise what interview participants said in response to a hypothetical situation: what would happen if their country adopted the other country's teacher accountability instruments? That is, what would happen if Finland instituted Singapore's system of performance standards, formal appraisal, competitive bonuses, and tiered promotions; or if Singapore replaced all of those instruments with a largely autonomous teaching profession?

Table 4.12 *Summary of interview participants' responses when asked how they would respond if, hypothetically, their country adopted the other country's teacher accountability instruments*

● agree ◑ partially agree ○ disagree [ ] not mentioned during the interview

	FINLAND												SINGAPORE											
	Anneli	Liisa	Emilia	Kristiina	Masa	Satu	Maarit	Hannele	Helena	Antero	Päivi	Juhani	Adeline	Jeffrey	Maggie	Joseph	Peter	Andy	Sonia	Timothy	Mark	Eleanor	Geok Ling	Jane
I would prefer working under the other country's teacher accountability instruments.	○	◑	○	○	○	◑	○	○	◑	○	○	◑	○	●	●		●	●	◑	●			●	
Most teachers would prefer the other country's teacher accountability instruments.	○	○	○		○	○	○	○	○	○		○		◑	○	◑			○	○	○			○
The other country's teacher accountability instruments would improve education in my country.				◑			◑	○			○		◑			○	◑	◑				○	○	

Note. This hypothetical question was posed in an open-ended manner. Thus, the statements in the table represent my summaries of participants' responses; I did not present these statements to participants for their agreement or disagreement.

As shown in the summary matrix, no interview participants unreservedly believed that the other country's teacher accountability system would be more effective or more popular than the existing system. This is especially striking in the case of the Singaporean participants who said that they would personally prefer Finland's system, but nonetheless doubted that it would effectively fulfil all the functions of Singapore's performance management structures. Maggie, a Singaporean participant, said that Finland's approach would give her the autonomy to pursue better long-term achievements, but added that:

> I think if you suddenly changed the system, most of the teachers will be very stressed because they have no idea what is being observed. Singaporeans find comfort, I think, in knowing exactly what is expected, and they like to do it to the letter.

In turn, Finnish participant Juhani said that he appreciated Singapore's structured approach to identifying teachers' strengths and weaknesses, but added that:

> We are so independent here. And we like that independence in our classrooms so much, that even the bonuses would not make this system a good thing. […] And we are so equal, among teachers. […] We do *not* want to give others the possibility of rushing higher.

Whatever their personal preferences, many participants had similarly strong reactions against the other country's accountability approach (e.g. Masa, describing Finnish teachers: 'They would quit. They would go on strike'; Sonia: 'The average Singaporean teacher will probably be up in arms').

Considering that the hypothetical accountability approaches were both internally coherent, and that each approach have been implemented successfully in an effective education system, these interviews suggest that some approaches to teacher accountability, however well-designed, may fail to be effective in uncongenial settings. Internal coherence within a set of teacher accountability instruments may be necessary for efficacy, but the interview participants' responses suggest that it is not sufficient. In the next chapter, I explore teachers' responses to accountability instruments in greater depth. For now, it is sufficient to observe that the interviews weigh against the likelihood that policy design can wholly account for the sociocultural moderation relationships observed in the regressions. Similarly, the regressions themselves weighed against the possibility of confounding effects from student background, school autonomy, and national wealth. Hence, my argument—that sociocultural context affects the relationship between teacher accountability and student outcomes—fits the empirical data better than the second alternative argument outlined above.

Chapter 5: Teacher accountability and teacher motivation

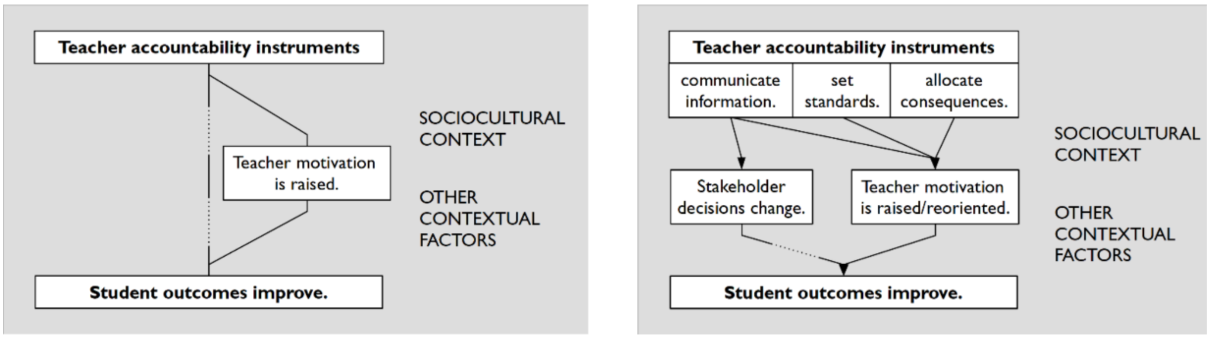
Having demonstrated the possibility that at least one aspect of sociocultural context is associated with the efficacy of some teacher accountability instruments, I turn to a possible causal pathway through which accountability instruments may affect student outcomes. Accordingly, this chapter explores my second research question: *to what extent, and how, does teacher motivation mediate the influence of teacher accountability instruments on student outcomes?*

First, I investigate the extent to which such mediation may occur, using cross-country data from TIMSS and PISA (Section 5.1). Next, I turn to the 'how' aspect, using field interviews that I conducted with teachers in Finland and Singapore to look at one step in this posited pathway. I begin by describing the teacher accountability instruments that the interview participants encounter in their work (Section 5.2). I then discuss their experiences of how accountability instruments influence teacher motivation (Section 5.3). I also explore the mechanisms underlying this influence: setting standards for teacher practice, communicating information on teacher practice, and allocating consequences based on stakeholders' judgements of teacher practice (Section 5.4). Finally, as in the previous chapter, I discuss these findings and consider an alternative explanation (Section 5.5).

5.1 TIMSS 2015 analysis: Teacher motivation and the relationship between teacher accountability and student outcomes

In this section, I explore the relationship between teacher motivation, teacher accountability, and student outcomes, as shown in Figure 5.1. This exploration uses ILSA data and sociocultural survey data. I begin by describing the pre-existing scales that I use to proxy for teacher motivation in this thesis, before presenting results from the mediation analysis as well as some sensitivity checks.

Figure 5.1 *Relationship between Section 5.1 and the overall conceptual framework*


Teacher accountability instruments
Teacher motivation is raised.
Student outcomes improve.
SOCIOCULTURAL CONTEXT
OTHER CONTEXTUAL FACTORS
Teacher accountability instruments
communicate information.
set standards.
allocate consequences.
Stakeholder decisions change.
Teacher motivation is raised/reoriented.
Student outcomes improve.
SOCIOCULTURAL CONTEXT
OTHER CONTEXTUAL FACTORS

This section

Overall conceptual framework

Using proxy scales for teacher motivation

As noted in Section 2.4, I adopt Schunk et al.'s (2010) definition of motivation as 'the process whereby goal-directed activity is instigated and sustained' (p. 4). Clear definition notwithstanding, motivation is difficult to measure. Schunk et al. (2010) further observe that motivation is a cognitive process that cannot be observed directly, but instead is inferred from observations of actions and verbalisations, such as an actor's choice of tasks, level of effort, or degree of persistence. Motivation can also be measured using self-reported indicators of actors' thoughts, choices, and feelings (Schunk et al., 2010), as with the self-report questionnaire items used to construct the proxy scales that I analyse. Such standardised questionnaires generated the teacher motivation data in most of the 130 studies analysed in a recent review of teacher motivation research (J. Han & Yin, 2016). Besides the problem of indirect observation, another challenge in measuring teacher motivation is that motivation is a multidimensional construct, such that any given scale will necessarily leave out some of its components. Among the instruments for assessing teacher motivation that Han and Yin (2016) reviewed, some instruments aim to capture various dimensions of motivation, such as different goal orientations (Butler, 2007, 2012; Nitsche et al., 2011; Roth et al., 2007), or the factors influencing teachers' professional engagement (de Jesus & Lens, 2005). Overall, as discussed in Section 2.4, there is no consensus across empirically validated theories of motivation about the motivational factors that matter most.

As noted in Section 3.3, I operationalise teacher motivation using three proxy measures: teacher-reported job satisfaction from TALIS 2013, teacher-reported job satisfaction from TIMSS 2015, and principal-reported teacher morale from PISA 2012. In this chapter, the outcome of interest is student achievement. Consequently, the regression analysis below does not use the TALIS

2013 scale, since TALIS does not measure student outcomes. However, TALIS 2013 is the main dataset in Chapter 6, so I discuss the TALIS job satisfaction scale alongside the TIMSS and PISA scales here because it is important to consider the similarities and differences between all of the teacher motivation proxies in this thesis.

One challenge in linking these proxy scales to teacher motivation in the conceptual framework is that the theoretical foundations of these scales are either weak or, at least, inadequately reported in the database documentation. The PISA 2012 teacher morale scale was included in the school questionnaires within a cluster of school climate variables, but I found no theoretical or empirical references justifying the scale design, whether in the PISA 2012 assessment framework, technical report, or results report on school factors (OECD, 2013a, 2013b, 2014b). Neither did I find any relevant scholarly citations in the documentation for PISA 2000 and 2003, which had also included this scale. The TIMSS 2015 assessment framework does cite an empirical study that found associations between teacher self-efficacy, teacher job satisfaction, and student achievement (i.e. Caprara et al., 2006), but I did not find any discussions of the theoretical basis for the questionnaire items (Mullis & Martin, 2013, p. 71). The TALIS 2013 results report mentions several studies that found associations between job satisfaction and job performance as well as other teacher-related variables, and it cites Locke's (1969) conceptual overview of job satisfaction, but only in passing (OECD, 2014c, p. 182).

Stepping back to examine job satisfaction and morale more generally, there are conceptual and empirical grounds for linking these constructs with motivation. Establishing this link conceptually requires a few steps for job satisfaction, but is straightforward for teacher morale. The APA Dictionary of Psychology defines morale as 'the level of enthusiasm, sense of purpose, or confidence in the worthiness of a goal that can affect a person's or a group's overall performance in working toward that goal, especially when under pressure' (American Psychological Association, n.d.)—which is a close cousin of motivation defined as 'the process whereby goal-directed activity is instigated and sustained' (Schunk et al., 2010, p. 4). In contrast, there is no widely agreed-upon definition of job satisfaction (Evans, 1997), perhaps because any job encompasses a range of needs—material and affective, primal and aspirational—that may or may not be satisfied. Nonetheless, some theories of motivation posit that motivation is founded on the pursuit of needs satisfaction (Herzberg, 1966; Maslow, 1954; Murray, 1938). Also, if motivation relates to not only the instigation but also the sustainment of goal-directed activity (Schunk et al., 2010), then job satisfaction may be considered a factor of motivation in that a more satisfied actor is more likely to continue pursuing a given activity. As noted above, Schunk et al. (2010) identify persistence as one behavioural indicator of motivation; and job satisfaction can indicate some elements of the likelihood that an employee will persist in a particular job.

Empirically, the reverse situation holds, in that it is much easier to establish an empirical link between motivation and job satisfaction than between motivation and teacher morale. To illustrate, a search in ERIC on 15 April 2020 for peer-reviewed publications containing both the terms ‘morale’ and ‘motivation’ in the title yielded only four results, of which the only quantitative empirical study was based on the same PISA 2012 teacher morale scale that I use (Abazaoglu & Aztekin, 2016). In contrast, a similar search for ‘job satisfaction’ and ‘motivation’ yielded 35 results. This is partly because there is an established, extensive strand of research on job satisfaction (Evans, 1997). In a special issue of *Learning and Instruction* focused on teacher motivation, one commentator observed that few studies of motivation have examined teacher motivation, ‘with the exception of writings about teachers’ sense of efficacy or teachers’ job satisfaction’ (Hoy, 2008, p. 492). Studies of teacher job satisfaction have found significant correlations between self-reported motivation and self-reported job satisfaction (variously operationalised) among teachers in China, the United States, Germany, and Indonesia (Chen, 2007; Davis & Wilson, 2000; Kunter et al., 2008; Murtedjo & Suharningsih, 2016), as well as in related professions (see Munyengabe et al., 2016, on university lecturers in Rwanda; and L. Li et al., 2014, community health workers in China). Skaalvik and Skaalvik (2009) also found significant negative correlations between burnout and job satisfaction among Norwegian teachers.

Turning to the proxy scales that I use in this study, it important to note that these three scales are not directly comparable with each other. As shown in Table 5.1, two items from the TALIS 2013 job satisfaction scale are roughly comparable to two other items in the TIMSS 2015 scale, although they are not identical (‘I enjoy working at this school’ in TALIS vs. ‘I am satisfied with being a teacher at this school’ in TIMSS; and ‘All in all, I am satisfied with my job’ in TALIS vs. ‘I am content with my profession as a teacher’ in TIMSS). However, the other scale items only relate loosely, at best, to items in the other scales. Given the interpretive, affective nature of these questionnaire items, it is difficult to determine how much the items align conceptually, and even more challenging to gauge the degree to which they overlap across survey respondents’ interpretations in practice. This applies particularly to the PISA 2012 principal-reported scale for teacher morale, which may be distorted by any number of conscious or unconscious biases on the part of the principal. Statistically, the TALIS and TIMSS teacher job satisfaction scales are moderately correlated (r=0.640), as are the TIMSS teacher job satisfaction and PISA teacher morale scales (r=0.524), as shown in Table 5.2. However, there is no noteworthy correlation between TALIS job satisfaction and PISA teacher morale (r=0.185). Since these scales measure different motivation-related constructs, the sensitivity checks across different datasets are less likely to show similar patterns. Still, it is worth conducting analyses using these different constructs, because, as discussed above, motivation is multidimensional, and different constructs may proxy for different elements of it.

Besides the fact that these three scales are not directly comparable, and that each scale only captures a limited, blurry view of teacher motivation, there is another possible source of statistical noise in the analyses that use these scales, reported in this section and Section 6.1. Namely, we would only expect these teacher motivation proxies to mediate the relationship between accountability instruments and student outcomes if the goals of the teachers and principals who responded to the questionnaires overlap at least to some degree with the forms of student learning measured by these ILSAs. It is entirely possible that a teacher may have a high levels of job satisfaction, while channelling their effort only toward personal financial goals, or bureaucratic compliance goals, or student development goals that are not captured in ILSA proficiency scales. This potential lack of convergence is less of a problem with the PISA 2012 principal-reported teacher morale scale, where the four items include one that explicitly pertains to students' cognitive outcomes ('Teachers value academic achievement') and one that broadly relates to school performance ('Teachers take pride in this school'), as shown in Table 5.1. However, the lack of convergence is a real possibility with the TALIS 2013 teacher-reported job satisfaction scale, for which the eight items discuss 'my job', 'working at this school', and 'being a teacher' in general terms, with no orientation toward student wellbeing and development. The TIMSS 2015 teacher-reported job satisfaction scale falls between the two, with four of the seven items referring to teaching, work, and satisfaction in general terms; but with three items referring to enthusiasm, inspiration, and pride, which are more likely to be interpreted with a student-oriented perspective.

Notwithstanding these limitations, as well as the other limitations pertaining to self-report cross-sectional survey data (see Section 3.6), the teacher job satisfaction scales in TALIS 2013 and TIMSS 2015, as well as the teacher morale scale in PISA 2012, are the best available proxies for teacher motivation in these cross-country educational surveys. As discussed above, there is some conceptual and empirical basis for regarding these scales as proxies for motivation.

Table 5.1 *Items in the TALIS 2013, TIMSS 2015, and PISA 2012 teacher motivation scales*

Questionnaire item	Comparable (and loosely related) items in		
	TALIS 2013	TIMSS 2015	PISA 2012
TALIS 2013 teacher job satisfaction (teacher-reported) *Response options: strongly disagree, disagree, agree, strongly agree*			
1. I would like to change to another school if that were possible.*		(#2)	(#3)
2. I enjoy working at this school.		#2	(#3)
3. I would recommend my school as a good place to work.		(#2)	(#3)
4. All in all, I am satisfied with my job.		#1	(#1)
5. The advantages of being a teacher clearly outweigh the disadvantages.		(#1)	(#1)
6. If I could decide again, I would still choose to work as a teacher.		(#7)	(#1)
7. I regret that I decided to become a teacher*		(#7)	(#1)
8. I wonder whether it would have been better to choose another profession.*		(#7)	(#1)
TIMSS 2015 teacher job satisfaction (teacher-reported) *Response options: never or almost never, sometimes, often, very often*			
1. I am content with my profession as a teacher.	#4		(#1)
2. I am satisfied with being a teacher at this school.	#2		(#3)
3. I find my work full of meaning and purpose.			(#1)
4. I am enthusiastic about my job.			(#2)
5. My work inspires me.			(#1)
6. I am proud of the work I do.			(#1)
7. I am going to continue teaching for as long as I can.	(#8)		(#1)
PISA 2012 teacher morale (principal-reported) *Response options: strongly disagree, disagree, agree, strongly agree*			
1. The morale of teachers in this school is high.	(#4)	(#6)	
2. Teachers work with enthusiasm.		(#4)	
3. Teachers take pride in this school.	(#1)	(#2)	
4. Teachers value academic achievement.			

Note. Parentheses indicate items that are loosely related but not directly comparable. Sources: Martin, Mullis, & Hooper (2016); OECD (2014b, 2014d).
*Item was reverse-coded in the scale.

Table 5.2 *Pairwise correlations (and number of countries) for country-level weighted means of the teacher motivation scales*

	TALIS 2013		TIMSS 2015		PISA 2012	
TALIS 2013 teacher job satisfaction	1	(35)				
TIMSS 2015 teacher job satisfaction	.640*	(15)	1	(39)		
PISA 2012 teacher morale	.185	(31)	.524**	(26)	1	(64)

*p < 0.05. **p < 0.01. (two-tailed)

The PISA 2012 teacher morale scale was scaled by survey administrators such that the OECD subsample of the data would have a mean of zero and a standard deviation of one (OECD, 2014b, p. 312). While the TIMSS 2015 and TALIS 2013 teacher job satisfaction scales were generated using different scaling procedures (Martin, Mullis, & Hooper, 2016, p. 15.4; OECD, 2014d, p. 213), I standardised them for this analysis such that their respective means were zero and standard deviations were one in the overall sample.

Mediation analysis with civic norms

To investigate the extent to which teacher motivation mediates the relationship between teacher accountability instruments and student outcomes, I draw on results from both model 2 and model 3. As detailed in Section 3.3, model 2 is similar to the main moderation model presented in the previous chapter, but with the addition of a teacher motivation variable as well as an interaction between teacher motivation and sociocultural context. Model 3 is also a moderation model, but it explores the joint association of teacher accountability and sociocultural context with teacher motivation (rather than with student outcomes, as in model 1). To preview the analysis, there is no clear evidence that teacher motivation mediates the relationship between teacher accountability instruments and student outcomes.

The main dataset for this analysis is TIMSS 2015, because TIMSS includes student outcome data matched with self-reported teacher job satisfaction data. I present full results from the regressions with civic norms. Given the relatively small number of TIMSS 2015 countries for which sociocultural data were available, using all six sociocultural constructs concurrently would result in multicollinearity from overfitting (see Appendix A). Moreover, sociocultural context is not of primary interest in this research question, so for the sake of conciseness I present full results only for one set of regressions with a single sociocultural construct. Among the six sociocultural constructs, I focus on civic norms since the model 1 analysis in the previous chapter found that the relationship between teacher accountability and student outcomes is moderated by civic norms. Hence, it is worth looking at whether civic norms similarly moderates the relationship between teacher accountability and teacher motivation. I discuss the other five sociocultural constructs as sensitivity checks subsequently. (The sensitivity checks also look at PISA 2012, as described below.)

First, I use model 2 to ascertain the degree to which the relationship between teacher accountability instruments and student outcomes changes when teacher motivation and its interaction with sociocultural context are added to the model. If (1) at least one of the teacher motivation terms is significant, and if (2) their addition to the model significantly reduces the coefficient on teacher accountability instruments, then this would indicate that teacher motivation significantly alters the relationship between teacher accountability and student outcomes. However, to constitute a mediation effect, it must also be true that (3) teacher accountability is a significant predictor of teacher motivation—which I investigate using model 3.

Results from these analyses are shown in Table 5.3. Columns (a) to (d) show a series of nested regressions culminating in model 2, where the outcome variable is pupil mathematics proficiency. Columns (e) to (h) show nested regressions for model 3, where the outcome variable is teacher job satisfaction.

Focusing first on model 2, the TIMSS 2015 data do not offer any clear evidence that teacher job satisfaction affects the relationship between teacher accountability and student outcomes. The coefficient of teacher job satisfaction is significant, as shown in column (c), thus fulfilling condition (1). However, this coefficient is relatively small, and its addition to the model—that is, moving from column (b) to (c)—only changes the teacher accountability term and its associated interaction terms marginally, thus failing to fulfil condition (2). Likewise, adding the interaction between teacher job satisfaction and civic norms does not meaningfully reduce any of the accountability-related coefficients, as shown in column (d). Moreover, the coefficient of this interaction is insignificant and even smaller than that of the main teacher job satisfaction term.

Turning to model 3, these data offer some limited evidence that teacher accountability significantly predicts teacher job satisfaction, conditional on sociocultural context. Although the main teacher country-level teacher accountability term is small and insignificant, as shown in column (g), the $\overline{\text{Accountability}}$*civic norms interaction term has a significant and relatively large effect on teacher job satisfaction, as shown in column (h). (For reference, the standard deviations of teacher job satisfaction and civic norms are approximately 1, and the standard deviation of the country-level weighted mean of teacher accountability instruments is 0.42, as shown in Table 3.4.) However, in this research question, I am not primarily interested in the determinants of teacher motivation, which I explore in detail the next chapter.

Table 5.3 *Models 2 and 3: results for multilevel regressions for TIMSS 2015, with and without teacher job satisfaction*

		Model 2 Y_{ptsc} = mathematics proficiency				Model 3 Y_{tsc} = job satisfaction			
		(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
Constant		505.62** (12.04)	483.63** (7.70)	484.03** (7.92)	483.98** (7.92)	-0.171** (0.049)	-0.133* (0.055)	-0.123 (0.077)	-0.103 (0.081)
Pupil level									
Home educational resources$_{ptsc}$			23.48** (1.75)	23.47** (1.75)	23.47** (1.75)				
Teacher level									
Teaching experience$_{tsc}$			0.21** (0.07)	0.21** (0.07)	0.21** (0.07)		0.001 (0.002)	0.001 (0.002)	-0.001 (0.002)
Teacher job satisfaction$_{tsc}$				3.84** (1.09)	4.03** (1.19)				
School level									
School resources$_{sc}$			5.53** (1.59)	5.35** (1.57)	5.39** (1.57)		0.048* (0.020)	0.048* (0.020)	0.049 (0.027)
Country level									
$\overline{\text{Accountability}}_c$			26.59 (15.98)	26.94 (16.12)	27.16 (16.03)			-0.035 (0.135)	-0.085 (0.113)
GDP$_c$			12.74* (6.13)	12.83* (6.24)	12.88* (6.25)		-0.075* (0.030)	-0.077* (0.03)	-0.026 (0.039)
Civic norms$_c$			14.06 (8.22)	14.46 (8.33)	14.32 (8.36)		-0.018 (0.068)	-0.023 (0.073)	-0.117 (0.073)
Interactions									
$\overline{\text{Accountability}}_c$ *Home ed resources$_{ptsc}$			-12.59** (3.67)	-12.58** (3.66)	-12.58** (3.66)				
$\overline{\text{Accountability}}_c$ *Teaching experience$_{tsc}$			0.13 (0.11)	0.11 (0.11)	0.11 (0.12)				0.007 (0.006)
$\overline{\text{Accountability}}_c$ *School resources$_{sc}$			-2.20 (4.23)	-2.18 (4.26)	-2.26 (4.22)				0.002 (0.053)
$\overline{\text{Accountability}}_c$ *GDP$_c$			21.57* (10.55)	21.99* (10.71)	21.96* (10.71)				-0.105 (0.064)
$\overline{\text{Accountability}}_c$ *Civic norms$_c$			-71.20** (16.48)	-72.32** (16.60)	-72.32** (16.56)				0.316** (0.094)
Teacher job satisfaction$_{tsc}$*Civic norms$_c$					-0.87 (1.44)				
Variance parameters	Pupil	4 169.71	3 968.19	3 968.26	39 68.26	—	—	—	—
	Teacher	2 256.89	1 898.60	1 911.60	1 911.55	0.802	0.800	0.800	0.801
	School	701.60	416.99	386.57	386.18	0.175	0.175	0.175	0.174
	Country	3 313.15	821.18	841.31	840.47	0.050	0.043	0.043	0.030
N	Pupils	118 363				—			
	Teachers	6 147				6 147			
	Schools	3 761				3 761			
	Countries	23				23			
-2*loglikelihood:		1374531.12	1367617.49	1367584.67	1367583.76	19282.57	19270.06	19269.98	19258.64
Likelihood ratio test (df)		—	6913.63** (11)	32.82** (1)	0.91 (1)	—	12.51* (4)	0.08 (1)	11.34* (4)

*$p < 0.05$. **$p < 0.01$.

Instead, the focus here is on the mediation relationship. The model 3 results in column (h) indicate (context-dependent) fulfilment of condition (3), i.e. that teacher accountability significantly predicts teacher job satisfaction. Nonetheless, models 2 and 3 together fail to show that teacher job satisfaction mediates the relationship between teacher accountability and sociocultural context, because the addition of the teacher motivation terms to model 2 does not alter the relationship between teacher accountability and pupil mathematics scores, as observed above.

Sensitivity checks

I conducted sensitivity checks to test whether this lack of evidence for conditions (1) and (2) of the mediation relationship may be a by-product of how the TIMSS data were constructed—perhaps the TIMSS teacher job satisfaction scale did not capture the aspects of teacher motivation that matter for accountability, or perhaps there were too few countries in the sample. Table 5.4 presents sensitivity checks for model 2, showing selected variables only, and continuing to focus on the civic norms regressions. The first two columns replicate columns (b) and (d) of Table 5.3. The other columns present analogous models firstly for TIMSS 2015 data matched to accountability data from PISA 2012 (rather than 2015), and then for full and OECD-only samples of PISA 2012.

In all of these regressions, the teacher motivation term is significant but small, as shown in each respective column (d), consistent with the main analysis above. The teacher motivation*civic norms interaction term is also small, although it is significant in the PISA regressions. More importantly, in comparing each pair of columns, the addition of these teacher motivation terms does not meaningfully affect any of the country-level teacher accountability terms, whether the unmoderated terms or the interactions (including the other A$\overline{\text{ccountabili}}$ty*context interaction terms that are included in each model but not shown in Table 5.4). In the PISA models, moving from the regressions in columns (b3) to (d3) and from (b4) to (d4) had similarly little effect on the school-level accountability terms, which are not shown in the table.

Table 5.4 *Model 2: sensitivity checks for maths proficiency in TIMSS 2015 and PISA 2012, with and without teacher motivation terms (showing selected variables only)*

Predictors (selected)	**TIMSS 2015** with Accountability from PISA 2015		**TIMSS 2015** with Accountability from PISA 2012		**PISA 2012** Full sample		**PISA 2012** OECD only	
	(b1)	(d1)	(b2)	(d2)	(b3)	(d3)	(b4)	(d4)
$\overline{\text{Accountability}}$	26.59 (15.98)	27.16 (16.03)	-3.17 (13.93)	-3.62 (14.20)	8.98 (8.97)	8.69 (9.12)	4.63 (10.48)	5.17 (10.62)
Teacher motivation		4.03** (1.19)		3.99** (1.17)		6.53** (0.92)		5.90** (1.07)
Civic norms	14.06 (8.22)	14.32 (8.36)	-1.42 (9.08)	-1.59 (9.38)	7.83 (6.16)	8.17 (6.32)	6.41 (5.86)	7.25 (6.14)
$\overline{\text{Accountability}}$ *Civic norms	-71.20** (16.48)	-72.32** (16.56)	-40.39* (16.96)	-40.59* (17.19)	-8.03 (9.19)	-8.49 (9.55)	-21.83* (10.40)	-22.58* (10.79)
Teacher motivation*Civic norms		-0.87 (1.44)		-0.74 (1.77)		1.80* (0.92)		2.74** (0.95)
N Pupils	118 363		120 117		375 207		275 715	
Teachers	6 147		6 062		—		—	
Schools	3 761		3 779		14 840		11 169	
Countries	23		22		52		36	

Note. The outcome variable is pupil mathematics proficiency. The teacher motivation variable for TIMSS 2015 is teacher job satisfaction (teacher-reported scale); and for PISA 2012, teacher morale (principal-reported scale). Columns (b1) and (d1) correspond to columns (b) and (d) in Table 5.3. Full results can be provided upon request. **p* < 0.05. ***p* < 0.01.

The fact that adding the teacher motivation terms to model 2 does not affect the parameter estimates of the accountability terms holds true not only for the civic norms models, but also for the other sociocultural constructs. For all four of the datasets shown in Table 5.4 above, swapping civic norms for any of the other five sociocultural constructs did not affect whether or not the model 2 regressions fulfilled the mediation conditions. As summarised in Table 5.5, regardless of which sociocultural construct was included in the model, these estimations resulted in significant but small teacher motivation coefficients, thus fulfilling condition (1) of the mediation relationship; with little effect on the teacher accountability coefficients, thus failing to fulfil condition (2). The same is true for PISA models with all six sociocultural constructs entered simultaneously, which are not summarised in the table.

For the model 3 estimations across sociocultural constructs, results are less robust. As shown in Table 5.5, the country-level teacher accountability weighted mean is consistently insignificant across all four datasets for all six sociocultural constructs. Among the interactions between sociocultural context and teacher accountability, the only significant parameters are $\overline{\text{Accountability}}$*civic norms in the main TIMSS 2015 dataset, $\overline{\text{Accountability}}$*confidence in institutions in the TIMSS 2015 dataset matched with PISA 2012 accountability data, and $\overline{\text{Accountability}}$*power distance in both PISA 2012 samples. Accordingly, when analysing country-level teacher accountability, only the TIMSS 2015 civic norms model for the main dataset, the TIMSS 2015 confidence in institutions model for the dataset with PISA 2012 accountability data, and the PISA 2012 power distance models for both the full and OECD samples meet condition (3) of the mediation relationship, i.e. that the accountability term is significant in model 3. Given space constraints, I only present parameter estimates for model 3 sensitivity checks in the next chapter, where the analysis focuses teacher motivation as an outcome variable.[31]

Table 5.5 *Summary of whether mediation conditions are fulfilled in TIMSS 2015 and PISA 2012, for all sociocultural constructs*

TIMSS 2015	with Accountability from PISA 2015 *N (countries) = 23*						with Accountability from PISA 2012 *N (countries) =22*					
	Conf	Netw	Norm	Trust	PDI	UAI	Conf	Netw	Norm	Trust	PDI	UAI
(1) Motivation is significant in model 2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
(2) Motivation reduces accountability estimate												
(3) Accountability is significant in model 3			✓				✓					
PISA 2012	Full sample *N (countries) = 52*						OECD countries only *N (countries) = 36*					
	Conf	Netw	Norm	Trust	PDI	UAI	Conf	Netw	Norm	Trust	PDI	UAI
(1) Motivation is significant in model 2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
(2) Motivation reduces accountability estimate												
(3) Accountability is significant in model 3					✓						✓	

Note. Each column represents a separate mediation analysis. The outcome variable in all the models is pupil mathematics proficiency. 'Motivation' refers to the proxy for teacher motivation and its associated interaction terms. 'Accountability' refers to the country-level weighted mean of the teacher accountability scale and its associated interactions (and, in the PISA 2012 data, includes both the country-level weighted mean and the school-level differentials). Full results can be provided upon request.

For the purposes of RQ2, however, the sensitivity checks for model 2 clearly establish the lack of a mediation effect from teacher motivation—at least, for these combinations of TIMSS 2015 and PISA 2012 data alongside WVS/EVS and Hofstede data. There are several possible reasons why these regressions failed to find a mediation effect. Theoretically, as discussed in Section 2.5, the pathway from teacher motivation to student outcomes is contingent on numerous teacher, student, and classroom characteristics. Hence, it is possible that the straightforward correlational analysis in this section may obscure crucial steps along the causal pathway. Moreover, the three

[31] The foregoing discussion focused on the country-level teacher accountability weighted mean because the analysis as a whole focuses on country-level patterns in teacher accountability, rather than school-level variation. That said, the PISA 2012 model 3 regressions also include a school-level teacher accountability differential term. In these regressions, the school-level teacher accountability differential term is consistently significant across all the PISA 2012 regressions, thus fulfilling condition (3) of the mediation relationship for the school-level accountability term. However, as with its country-level counterpart, the school-level accountability term is not affected by the addition of motivation to model 2, thus failing to fulfil condition (2) of the mediation relationship. In short, the PISA 2012 data do not provide any evidence that sociocultural context mediates the relationship between school-level teacher accountability instruments and student outcomes.

key constructs in these models—teacher motivation, teacher accountability, and sociocultural context—are all measured using scales derived from self-report cross-country surveys. Such scales may incorporate substantial measurement error, which may attenuate possible mediation effects.

Given the complexity of both the causal pathway and the constructs in question, it may be more appropriate to examine empirical data that offers a more nuanced view. Accordingly, in the rest of this chapter, I look at evidence for the relationship between teacher accountability instruments and teacher motivation in my other empirical data source: interviews with 24 teachers in Finland and Singapore.

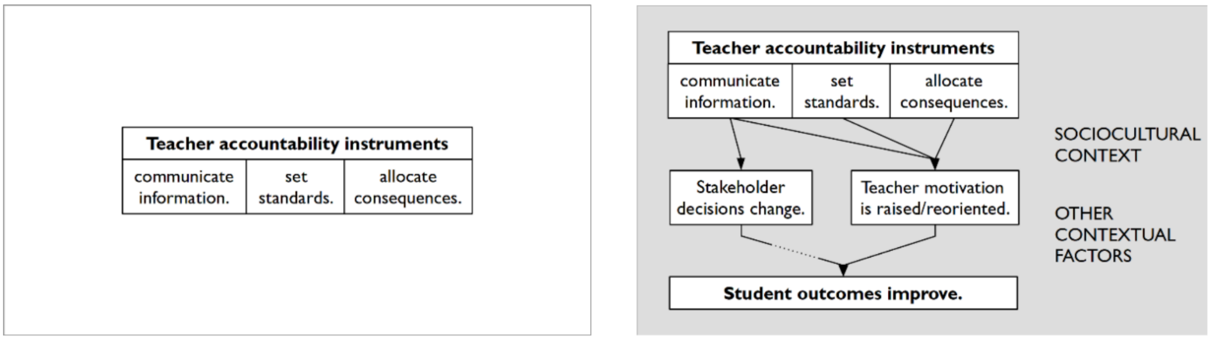
5.2 Teacher accountability instruments in Finland and Singapore[32]

Although the statistical analysis in the previous section did not find a consistent association between teacher accountability instruments and teacher motivation, the interviews that I conducted with teachers in Finland and Singapore showed clear evidence of such a relationship. As observed in the introduction and methodology chapters, Finland and Singapore have strikingly different—but comparably effective—approaches to teacher accountability. As noted in Section 3.4, Finland and Singapore fall respectively in the lowest and highest quartiles of $\overline{\text{Accountability}}$ in PISA 2012, PISA 2015, and TALIS 2013. Despite these differences, interview participants in both countries regarded teacher accountability instruments as playing a positive causal role in their respective systems, to use Cartwright and Hardie's (2012) terminology. Throughout the rest of this chapter, I make the case that this positive causal role takes place (at least partially) via the influence of accountability instruments teacher motivation.

To lay the foundation for this argument, in this section I describe each country's approach to teacher accountability, as indicated in Figure 5.2. This description focuses on teacher accountability instruments as experienced by my interview participants, although I do cite other sources for corroboration and context.

[32] Some of the material in this section is modified from a chapter titled 'Contrasting approaches, comparable efficacy? How macro-level trust influences teacher accountability in Finland and Singapore', which I wrote for a volume that is currently in preparation for publication, titled *Trust, Accountability, and Capacity in Education System Reform: Global Perspectives*, edited by Melanie Ehren and Jacqueline Baxter. In particular, material under the subsections titled 'Setting standards', 'Collecting information', and 'Allocating consequences' was taken (and, in some places, expanded or amended) from the book chapter manuscript. I am the sole author of the book chapter, and all data and analysis in the book chapter were carried out as part of this PhD project. All other parts of this section were written specifically for this thesis.

Figure 5.2 *Relationship between Section 5.2 and the overall conceptual framework*



This section

Overall conceptual framework

In the interviews, participants from Finland and Singapore painted very different pictures of the accountability instruments they encounter in their work. The matrix in Table 5.6 gives one snapshot of the instruments that participants mentioned.

In interpreting this matrix, note that the absence of a symbol does not necessarily mean that the participant did not experience the instrument in question. Rather, it simply means that the participant did not mention the instrument during the interview, for any number of reasons (e.g. Hannele and Geok Ling had tight schedules on the days when we met, so their interviews covered less ground). Note also that I specifically asked participants about certain instruments (e.g. I asked Finnish participants about discussions with the headmaster and about lesson observations; and I asked most participants in both countries about parental feedback), whereas my questions on some areas were more general (e.g. 'What rewards or penalties are there?'), and some other areas were only mentioned if the participant brought them up independently (e.g. standards for entry into teacher training). Even where the matrix shows that a teacher accountability instrument is present in both countries, there may be vast qualitative differences between the prevailing versions of the instrument, as I describe below for financial rewards.

One immediate observation from Table 5.6 is that there is more within-country variation in teacher accountability instruments in Finland than in Singapore. While Singaporean participants differed on the prevalence of a few instruments (performance-oriented interactions with colleagues, ad hoc parent feedback, and student feedback), Finnish participants reported heterogeneous experiences across most instruments. This is partly because Singapore's education system is highly centralised (Dimmock & Tan, 2013), whereas Finnish municipalities and school leaders can determine many aspects of teacher accountability, within the limits established by central government agencies and trade unions (Simola, Rinne, Varjo, Pitkänen, & Kauko, 2009).

Table 5.6 *Teacher accountability instruments mentioned by interview participants in Singapore and Finland*

● implemented ○ not implemented
◑ implemented partially or only occurring in particular circumstances [ ] not mentioned during the interview

	FINLAND												SINGAPORE											
	Anneli	Liisa	Emilia	Kristiina	Masa	Satu	Maarit	Hannele	Helena	Antero	Päivi	Juhani	Adeline	Jeffrey	Maggie	Joseph	Peter	Andy	Sonia	Timothy	Mark	Eleanor	Geok Ling	Jane
PRE-SERVICE																								
Setting standards																								
Selection criteria for teacher training				●		●				●		●								●				
Initial teacher training	●		●	●	●	●	●					●						●						
IN SERVICE																								
Setting standards																								
National curriculum	●	●	●	●	●	●	●		●	●	●	●	●		●	●	●	●	●		●	●		
Government goals or guidelines for teachers					●	●		○		○	●	●	●	●	●	●	●	●	●	●	●	●	●	●
School-level goals or guidelines for teachers		●	●					●		●			●	●	●	●			●		●	●	●	●
Communicating information																								
Reporting requirements	●	●	◑		○	●	●	○	○	●		◑	●	●	●	●	●	●	●	●	●	●	●	●
Lesson observations	○	○	○	○	◑	○	◑	○	●	○	○	◑	●	●	●	●	●			●	●	●		●
Performance-oriented discussions with managers	●	◑	●	◑	◑	◑	◑	◑	◑	●	◑	●	●	●	●	●	●	●	●	●	●	●	●	●
Performance-oriented interactions with colleagues	●	◑	●	◑	◑	○	◑		●	●	◑	◑		◑	◑		●	●		●	●	●	●	◑
Use of student test results in teacher accountability				○	◑				○			◑	●	●	●	●	○	●	●	●	●	●		●
Routine parent-teacher discussions		●								◑			●	●			●			●				●
Ad hoc parent feedback	●	◑	◑	◑	◑	◑	◑	●	◑	◑	◑	●	●	◑	●	◑	●			◑	●	●		●
Channels for student feedback	◑				◑	●	●		●	◑	◑	●	◑	●			●		◑	◑	◑	◑		◑
External collaborations			◑		◑	◑	◑		◑	◑				◑						◑		●		
Mass/social media					◑		○						◑				◑			◑	◑	◑		
Formal appraisal ratings													●	●	●	●	●	●	●	●	●	●	●	●
Allocating consequences																								
Promotions		◑									◑		●	●	●	●	●	●	●	●	●	●	●	●
Financial rewards	○	◑	◑	○	○	●	●	○	○	●	●	●	●	●	●		●	●	●	●	●	●	●	●
Public recognition for good practice	◑			◑											●							●	●	●
Discussions, warnings, or monitoring for improvement			◑	◑	◑			◑		◑		◑		●		●					●	●		●
Penalties	◑		○	◑	◑	◑	◑	◑		◑	◑		◑		●				●	●	●	○		

Another immediate observation is that there appears to be far more extensive use of teacher accountability instruments in Singapore than in Finland. The greater extent of accountability instruments was likewise evident in the PISA data on Finland and Singapore, as shown in Figure 5.3. From this figure, it is worth noting that there is reasonably good agreement between these principal-reported PISA surveys and the teacher interviews, despite the time lags between the 2012 and 2015 PISA surveys and the 2018 interviews. For example, based on PISA 2012 data, 95.7% of school-going Singaporean 15-year-olds were then attending schools where teacher appraisals could lead to changes in the likelihood of career advancement (OECD, 2013b, Table

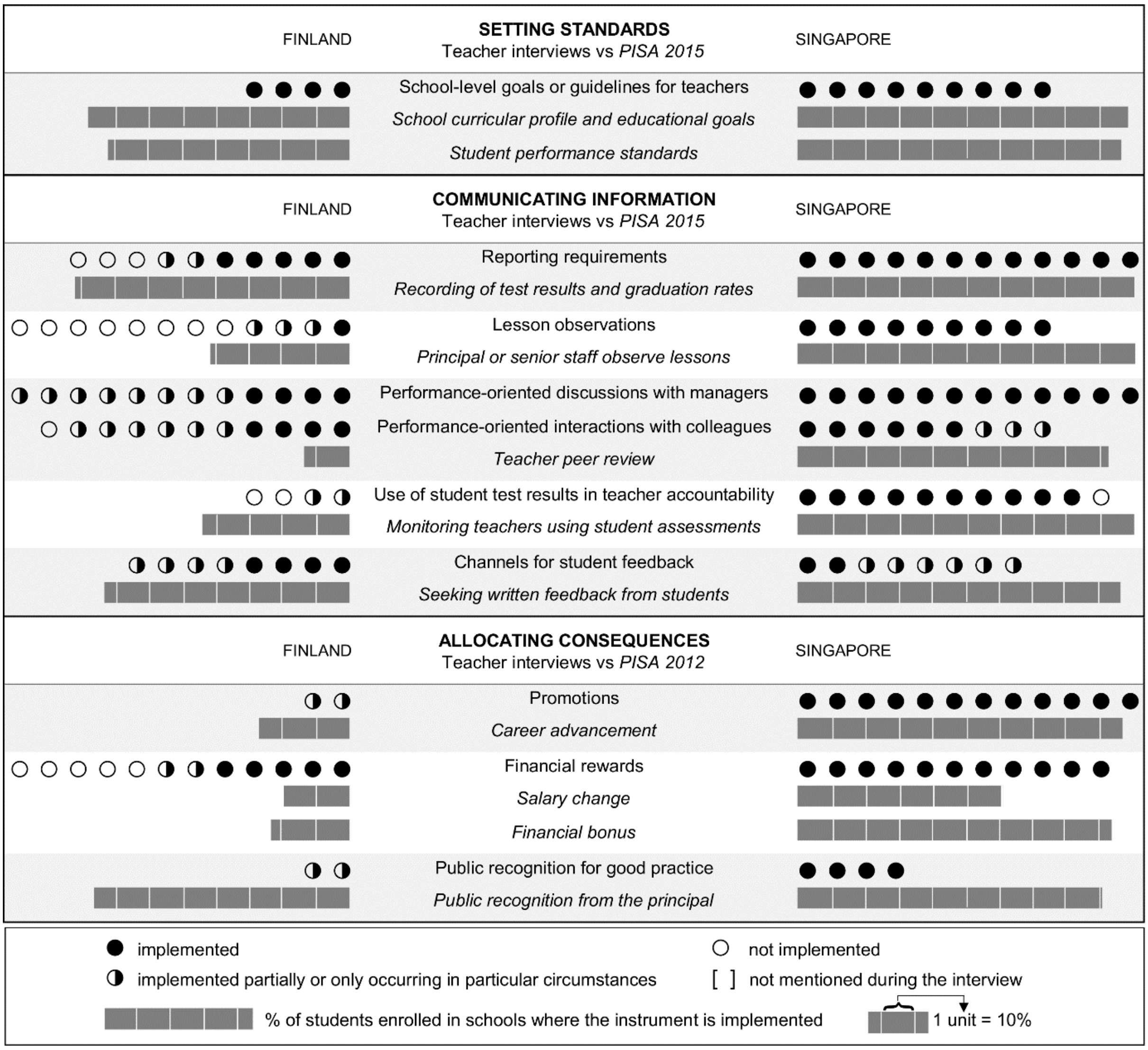
IV.4.35); and all 12 Singaporean interview participants similarly said that annual performance grades affect promotion speed.

Where there are notable divergences, these are likely due to either (a) information sources, i.e. PISA questionnaire answers from principals versus interviews responses from teachers, and (b) the design of the data collection instruments. For an example of a divergence due to different information sources, 74.9% of Finnish school-going 15-year-olds and 89.8% of their Singaporean counterparts were attending schools where teacher appraisal could lead to public recognition from the principal, according to PISA 2012 (ibid). However, none of the interview participants mentioned such recognition specifically from the principal—although some did mention school-level or national-level awards for good teaching practice, with four Singaporean participants identifying such awards as a regular occurrence and two Finnish participants saying that such recognition may be awarded under certain circumstances. This divergence is likely due to teachers and principals having different notions of what constitutes public recognition from the principal, or to principals' ideas of such recognition being too informal or commonplace for the interview participants to recall them during the interviews as rewards of good performance. As for a divergence due to instrument design, 'teacher peer review' in the PISA 2015 school questionnaires is much narrower than my interview categories of performance-oriented discussions or interactions with managers or colleagues,[33] so it is unsurprising that Finnish principals in PISA 2015 reported the former far less than my Finnish interview participants reported the latter. Overall, though, there is broad agreement between the two sources, which strengthens the credibility of the field interviews.

Below, I discuss in greater detail the accountability instruments experienced by interview participants. As in Table 5.6 and Figure 5.3, this discussion is organised according to the three accountability mechanisms in my conceptual framework: setting standards for teacher practice, communicating information about teacher practice, and allocating consequences based on judgements of teacher practice. In each subsection, I first discuss the teacher accountability instruments that operate via this mechanism in Finland, before turning to those in Singapore.

[33] In my interview coding, "performance-oriented discussions with managers" and "performance-oriented interactions with colleagues", as shown in Table 5.6, interpret "performance-oriented" in the broad sense of helping a teacher to be a better teacher. While "performance-oriented" includes quantified or otherwise standardised performance evaluations, in this coding it goes far beyond those terns; also including, for example, an interview participant saying that they informally approach their headteacher for advice, or a participant saying that they collaborated with colleagues to develop formative assessments for a particular unit.

Figure 5.3 *Teacher accountability instruments in Singapore and Finland, as reported in the interviews versus PISA 2012 and 2015*



Note. For the full phrasing of the PISA 2012 and 2015 questionnaire items, see Table 4.2 in the previous chapter. Sources: OECD (2013b, Table IV.4.35; 2016a, Tables II.4.33 and II.4.39).

Setting standards

As discussed in Section 2.5, one of the mechanisms through which teacher accountability instruments can influence teacher motivation is by setting standards for teacher practice, thus changing the goals or practices to which teachers' motivation is oriented. To begin with the field interviews from Finland, setting high standards for teacher quality is fundamental to Finland's teacher accountability approach. Almost every Finnish interview participant named the national curriculum as an accountability instrument. While the curriculum is set at the national level, every municipality publishes a localised version of the national curriculum, 'which, of course, cannot be different from the national version, but it's more specific', as noted by interview participant

Anneli (see also Finnish National Board of Education, 2014). Additionally, subject-specific branches of the teachers' union sometimes publish guidelines or exemplar assessments as benchmarks for interpreting curricular standards (as noted by Maarit and Masa). A further source of curricular interpretation is the textbooks produced by educational publishers, which many teachers rely on for planning their lessons (as noted by Masa and Emilia; see also Crehan, 2016, Chapter 4). Even if some teachers may not comply with all the minutiae in the curriculum, they still value its standard-setting role. For example, while calling the current curriculum 'very complicated to understand' and 'sometimes unclear', and saying that there were probably some curricular expectations that she was not fulfilling, Finnish interview participant Liisa also said:

> Well, we need to have a national curriculum. It's definitely a 'must'. If we didn't have it, then we would not have a common ground for the students to continue on to upper secondary school.

Similar observations were made by Antero, Satu, and Päivi. In addition to agreeing that the curriculum was important, participants also agreed that it left them substantial freedom in their practice. In Anneli's words, 'The national curriculum gives us guidelines, but I can still do my work the way that I feel is the best way for me, and for my students.' Some of this freedom stems from policy changes in the 1990s that, among other things, abolished the school inspection system (Nikki, 2000; Webb, Vulliamy, Häkkinen, & Hämäläinen, 1998). However, much of this teacher autonomy is of even longer standing. As early as 1981, the director general of Finland's National Board of General Education spoke to OECD officials about 'the importance of traditional freedom for teachers in Finland to decide about the curriculum', such that '[it was] common … for the teacher to decide what issues shall be stressed within the syllabus or the textbook' (as reported in OECD, 1982, p. 81). Thus, Finland's curriculum plays a fundamental role in maintaining common standards alongside teachers' pedagogical autonomy.

Arguably more important than the curriculum are the standards set by Finland's famously stringent admissions processes for teacher training (Malinen, Väisänen, & Savolainen, 2012; Muhonen, 2017), as well as its rigorous, master's-level teacher training programmes (Sahlberg, 2015a; Tirri, 2014; Wiksten, 2018). There is an established literature on accountability instruments that affect the overall composition of the teaching profession—and, thus, their collective levels of motivation—via performance-based consequences such as merit pay, as outlined in Section 2.5. Although Finland does not use such consequence-based instruments to influence teachers' self-selection into and out of the profession, it makes significant use of standard-setting in teacher training admissions and preservice training to influence the level and direction of teacher motivation collectively. The admissions process includes not only academic requirements, but also aptitude assessments that evaluate 'applicants' suitability, motivation and commitment to teacher education and the teachers' work' (Malinen et al., 2012, p. 572; see also Sahlberg, 2015b). In turn, preservice training ingrains the expertise that orients future teachers' motivation toward effective principles and practices. Although a few Finnish participants mentioned neither the admissions criteria nor the preservice training as accountability instruments, as shown in Table 5.6, this is probably because such point-of-entry instruments may be taken for granted in teachers' day-to-day work. Nevertheless, Antero called the stringent admissions processes 'the most important thing' that Finland's government can do for the quality of teaching. Likewise, Masa said that preservice teacher training was 'the most influential thing' in teacher-related policy, far more important 'incentives and disincentives and that kind of stuff'.[34]

In Singapore, teacher training is similarly selective, with one report stating that it admits roughly 12.5% of applicants (Butrymowicz, 2014), comparable to Finland's 11% admission rate for primary school teacher education programmes in 2016 (Paronen & Lappi, 2018). However, as shown in Table 5.6, Singaporean interview participants did not identify teacher training as an instrument for teacher accountability. The closest they came was Timothy mentioning that he had to apply twice before successfully getting a teacher training scholarship, and Andy saying that one of his lecturers had told their class of trainees that 'sometimes you do need to know when to blow your own trumpet' in order to get a good performance ranking.

Such ranking takes places through Singapore's Enhanced Performance Management System (EPMS). The EPMS is the education ministry's version of the national civil service appraisal system, which was itself based on the performance management system used by the Shell Petroleum Company in the 1980s (J.-M. Ho & Koh, 2018; Neo & Chen, 2007; Quah, 2010). Central to the EMPS is its teaching career ladder, which has three tracks: teaching, leadership, and senior specialist (C. H. Teo, 2001). Every career track has distinct performance standards in several key results areas under three outcome categories, which vary according to the teacher's position on the career ladder (Kan, 2014). To illustrate, Joseph said that:

> Under the teaching track, there are three areas that they will look at. The first, nurturing the child, will encompass your subject and your form teacher responsibilities, as well as aspects of your CCA [i.e. co-curricular activity]. Then you have professional

[34] I am grateful to Lucy Crehan for the observation that teacher accountability systems can concentrate their quality controls on different junctures, whether the inputs, processes, or outputs of the teaching profession. Similarly, Oates (2015) observes that the teaching profession in Finland emphasises 'front-end restrictions' rather than 'back-end' ones (p. 4).

> development—so, what are the training plans that you have? Another one is organisational contribution—so, what are the portfolios and school programmes that you actually contribute to?

Besides EMPS structures, many interview participants also mentioned standard-setting instruments within their subject departments. Department-level instruments include timelines for syllabus coverage and assessment frequency—as well as targets for student achievement in the national standardised exams that loom large over the school system.

Alongside these standard-setting instruments, Singapore's teacher management system includes some developmental structures. Some of these structures may not fall strictly within my definition of teacher accountability instruments, but they aim to help teachers to meet the accountability standards. For example, every school has a teacher in middle management who is designated as the school staff developer. As Eleanor explained, the staff developer considers each teacher's career track and career goals, and 'will work with the teacher to see what professional development this teacher requires in order to get to the next level that he or she desires to get to'. Besides the staff developer, every teacher is assigned a reporting officer who is responsible for monitoring their performance via the EPMS. Reporting officers, who are teachers holding management-level responsibilities in the same school, sometimes offer a great deal of support toward meeting standards. For example, Mark described a positive experience with a former reporting officer:

> Because I had an RO [reporting officer] who took EPMS very seriously from a developmental perspective, I did feel like I was developed well, based on what he and I had decided were my goals for the year. And I felt like if I was not on track, the system and my school leaders were very willing to say that, 'Well, it seems like maybe to take that next step, you would need to attend a certain training course. Or you might need to meet with this senior teacher who's done this before.'

Mark's and Eleanor's accounts suggest that Singapore's education system pays a great deal of attention to the standards set for each teacher, and invests resources in supporting them to reach these goals. However, some participants said that these developmental processes are not implemented effectively in their schools, partly because reporting officers are overworked. According to Jeffrey, 'If you get feedback about how to do it better next time, it's usually in the form of scolding. If they even tell you. (laughing) Or if they can be bothered to tell you.' I describe some other aspects of teachers' interactions with their reporting officers in the next subsection.

To summarise, both the Finnish and Singaporean interview participants experience substantial teacher accountability via standard-setting. However, the structure and scope of these instruments differs considerably. Singapore's teacher accountability approach includes a comprehensive set of nationally mandated standards for individual teachers' performance in-service, alongside school- and department-level standards guiding teacher practice collectively. These standards can be exacting, but teachers are (or, at least, are meant to be) given developmental support to reach these standards. In contrast, Finland sets high standards for teachers' collective motivation via extensive preservice training and stringent selection processes to enter the training in the first place. However, in-service teacher practice is circumscribed primarily by the national curriculum, which leaves them considerable freedom.

Communicating information

Besides setting standards, teacher accountability instruments can also attempt to influence teacher motivation by collecting and communicating information on teacher practice. In many Finnish schools, one source of information on teacher practice is an annual developmental discussion between each teacher and their principal (Kumpulainen & Lankinen, 2016; NCEE, 2016, pp. 5–6). While every Finnish participant mentioned such discussions, as shown in Table 5.6, some said that these discussions did not take place every year, or that they took place in groups of subject teachers rather than as targeted individual reviews. For comparison, in a 2017 survey, 65% of Finnish teachers said they had had one developmental discussion during the prior 12 months, 27% said they had had two or more discussions, while 8% said they had not had any (OAJ, 2018, p. 31). Most interview participants said that the developmental discussions were informal in tone, involved two-way feedback, and did not lead to any follow-up (see also Webb et al., 2004, p. 100, for similar observations on a lack of follow-up to self-evaluation processes in Finnish schools). In Maarit's account,

> I think that the sitting down and talking are sometimes forgotten after that. I'll write down something fine, and then we talk, and then the year, every day continues, and probably now I don't even remember what I answered.

Apart from the developmental discussions, sources of information about teacher practice that were identified by Finnish participants include self-initiated collaborations between colleagues (e.g. to develop content for a particular unit, or to remedy a classroom problem), and discussions or lesson observations triggered by parental complaints. Such parent-triggered instruments appear to be rare, with participants mentioning only a few specific incidents over their careers. (That said, a few participants did mention that parents have become more demanding in recent years, especially in more socioeconomically privileged areas. Similar observations were also reported by Singaporean participants.)

Besides these relatively sparse instruments that communicate information in order to directly influence teacher practice, Finnish participants also mentioned some instruments that collect information on teacher practice in order to facilitate administrative decisions. For example, even though pupils in Finland do not take any national standardised tests until they reach the matriculation exam for university entry, the Finnish Education Evaluation Centre administers a system of sample-based tests to monitor national educational quality (Andere, 2014, pp. 88–89; Vainikainen et al., 2017). However, these sample-based tests do not play an accountability function in individual teacher practice. In Kristiina's words, 'I have *never* known anybody, or any school, or any teacher who has taken part in them. So they are not related to the everyday work of a teacher.' Below the national level, each of Finland's municipalities—of which, in 2018, there were 311 (Statistics Finland, 2018)—has considerable decision-making power over local education (European Commission/EACEA/Eurydice, 2018; Simola et al., 2009). As a result, some municipalities frequently collect information from teachers (Emilia: 'a massive amount of different questionnaires'), whereas others do not (Hannele: 'not at all').

Contrastingly, under Singapore's centralised EPMS structures, all teachers regularly report on their work through formal channels, which inform annual performance grades. These performance grades, which range from A to E (Singapore Teachers' Union, 2014), are awarded based on several sources of information. At the beginning, middle, and end of each school year, every teacher is required to document their targets and achievements on an EPMS form, and then to discuss their performance toward these targets in a work review session with their reporting officer. Teachers are observed in the classroom once a year by their respective reporting officers, who also examine a sample of students' work. These sources of information are then discussed at appraisal panels, where reporting officers triangulate each other's observations and compare teachers across each level of the career ladder, before allocating performance grades for the year (Singapore MOE, 2018b). This can lead to the sense that teachers are constantly under observation by all reporting officers in the school. As Sonia said:

> Every time someone walks by, you know you are being judged. Say you turn up five minutes late to the parade ground for assembly, you know that someone out there is eyeballing you and marking you down and saying, 'Okay, this is the person with the so-called punctuality problem.'

Besides heightening self-consciousness among some teachers, the informational requirements of the EPMS generate substantial administrative work for reporting officers, who may have to appraise numerous colleagues (Eleanor: 'sometimes you have got a good ten staff to oversee, plus all the other admin work').

Although EPMS criteria do not formally include student test results, such results may nonetheless be used as EPMS performance targets or impact measures (Loh & Hu, 2014, p. 16). As shown in Table 5.6, almost all Singaporean participants mentioned accountability pressures from tests and exams, whether national exam results at the end of secondary school, or school-level assessments throughout the year. Another frequently cited teacher accountability instrument was parental feedback. Participants gave varying accounts of the frequency and intensity of parental feedback, but it was clear that this feedback could sometimes be onerous. For instance, Mark said that 'parents expect the teachers to be on call 24/7 for student needs', and Maggie mentioned that some teachers use phone number masking services when contacting parents in order to forestall such 24/7 access.

Interestingly, Finnish and Singaporean participants had similarly mixed views on the extent to which professional collaborations with other teachers functioned as an accountability instrument. Among Finnish participants, some spoke of regular collaborations, especially with colleagues who teach the same subjects. However, others said that collaboration was infrequent or only happened in specific situations, such as adapting new curricular requirements to their school's needs. A few noted that levels of collaboration vary from school to school. Partly because of these differing levels of collaboration, Finnish participants mentioned different degrees of accountability to colleagues, ranging from clear accountability relationships (e.g. Emilia: 'At least in this school, because we collaborate so much, it's more difficult to do things your own way, or to cut corners, or things like that') to tenuous ones (e.g. Satu: 'Nobody knows what I do in my classroom, [...] only myself and the students, but not my colleagues or the headmaster, nobody'; see also Y. Li & Dervin, 2018.).

Likewise, some Singaporean participants did not mention peer collaboration at all; whereas some said that they collaborated regularly with colleagues, whether through formal departmental structures and professional learning communities, or via informal information-sharing and self-initiated partnerships. One participant, Maggie, said that most Singaporean teachers 'hoard their materials and information quite a lot'—but she also described an intensive, yearlong collaboration with a colleague as her proudest achievement in teaching.

These similarly mixed pictures of professional collaboration among teachers contradict both the image of Finland as a haven of teacher professional collaboration (e.g. Strauss & Sahlberg, 2015), and the image of Singaporeans as relentlessly competitive, selfish people (e.g. Pierson, 2019). However, beyond professional collaborations, Singaporean participants experienced far more informational accountability instruments than their Finnish counterparts, partly because the former were subject to the monitoring demands of the EPMS.

Allocating consequences

Finally, the third teacher accountability mechanism in my framework entails allocating consequences based on judgements about teacher practice. Finland's teacher accountability instruments for allocating rewards and penalties are similar to its instruments for collecting information on teacher practice: they are usually unobtrusive, and they vary across municipalities. Rewards come in the form of small salary supplements for teachers who take on extra tasks that are otherwise uncompensated—rather than bonuses based on how well teachers have performed their tasks. Some of these supplements are part of the union-negotiated salary structure, while others come from the municipality's discretionary budget. Supplements from the discretionary budget are allocated in different ways (e.g. Antero: based on the principal's decision; Liisa: based on an application to the municipality) and for different tasks, whether routine (e.g. Kristiina: 'if you take care of the annual choir performances at the school') or ad hoc (e.g. Emilia: 'a project which has touched the whole school'; Juhani: 'when there's been renovations in some schools, and you had to figure out new ways of teaching and change places a lot').

Although there may be some element of merit in awarding the supplements for ad hoc tasks or in how the routine tasks are allocated, most Finnish participants did not regard these supplements as merit based. In Masa's words:

> I see that as, 'If you want to do this crap job, then we'll give you money for it.' And some crazy person is going to be okay with that, whereas the rest are like, 'Phew, thank goodness I don't have to do it.'

Different municipalities distribute these salary supplements differently, and they can be so inconspicuous that two participants from the same school disagreed about whether or not these supplements existed. Where they do exist, interview participants regarded them as token sums (e.g. Liisa: 'around a hundred euros a month'; Emilia: 'a gift card somewhere for fifty euros or something'; Satu: 'the [salary] difference is very low, maybe you can buy one movie ticket'). As for penalties, besides the developmental actions that can be triggered by parental complaints, the only penalty participants identified is being fired for egregious misconduct, such as drunkenness or physical violence in school.

In contrast, every teacher in Singapore is eligible for an annual performance bonus—or subject to career progression penalties—based on their EPMS performance grade. According to Mark, top-performing teachers can receive bonuses of up to 3.5 months' salary. Performance grades also affect the speed at which teachers are promoted along the career ladder. Additionally, good performance grades render teachers eligible for funded study leave (as noted by Maggie), whereas unsatisfactory grades lead to developmental coaching and extra monitoring (as noted by Jane and Joseph) and, eventually, firing (as noted by Maggie and Mark; see also Singapore Teachers' Union, 2014, 2015).

A noteworthy feature of Singapore's EPMS is that performance grades do not derive solely from EPMS standards. Rather, as Geok Ling observed, 'It's criterion-referenced, and then it's also norm-referenced.' Within each school, performance grades are awarded on a forced curve that benchmarks teachers against colleagues of the same level of the career ladder. Official guidelines about the EPMS grading system are not publicly available, but one non-Ministry source says that approximately 30% of teachers receive A or B grades, 65% receive C grades, and 5% receive D or E grades (McMillan, 2017; see also Bruns & Luque, 2014). (In addition to their substantive grade, i.e. their rung of the career ladder, and their annual performance grade, every Singaporean teacher is also rated on their 'currently estimated potential' [CEP], which reflects managers' expectations of the highest substantive grade the teacher will attain prior to retirement. A teacher's CEP is never disclosed to them, but it shapes their professional development opportunities and the speed at which they are promoted [Neo & Chen, 2007; Quah, 2010; C. H. Teo, 2018]. Interview participants confirmed that they did not know their own CEPs.) However, despite its inbuilt competitiveness, and despite participants' acknowledgement that the fairness of the grading system could vary vastly across schools and reporting officers, the EPMS did not appear to undermine collegiality among teachers. In Andy's words, 'We do recognise those who are deserving of credit because [...] something about them enables them to go above and beyond for the students, and we don't begrudge them if they are rewarded accordingly.'

Summary

Finland's approach to teacher accountability, as experienced by the interview participants, depends heavily on standards. These standards are set at the point of entry into the teaching profession, with selection processes for teacher training ensuring that those who enter the profession are capable and highly motivated, and the subsequent training ensuring that teachers know how to do their jobs well. Strikingly, Finnish participant Päivi suggested that these standards are not only sustained, but also strengthened, as teachers continue in their service. An experienced teacher who had once spent a stint in school leadership before choosing to return to classroom teaching, Päivi said:

> When I was the headmaster, I had the idea that when young teachers come to the school, then we could learn so many new things from them. […] But then I saw that, after a year or two, new teachers are just the same as old teachers. […] But I wouldn't change it. Teachers are kind and they are caring and they are hoping for the children's best. And that's the way it is.

These selected, socialised, and trained teachers are then accorded considerable autonomy in their classroom work, with few informational and consequential accountability instruments. In Helena's words, 'because the accountability system is so light, it actually works through the negative'. That is, rather than actively monitoring teachers' performance, the system hinges on the probability that student dissatisfaction with subpar teaching would trigger complaints to the principal, which would in turn trigger remedial action. Helena also observed that:

> You don't need as many control tools when the assumption is that we teachers are professionals who know their job and have the skills, and that we are all interested in the same goals and in delivering the curriculum.

This reflected a general view among interview participants that Finland's teacher accountability instruments are effective in supporting desirable educational outcomes.

Singapore's teacher accountability approach is far more extensive, depending on an interlocking system of standards, monitoring, and consequences within the annual EPMS cycle. Nonetheless, Singaporean participants likewise endorsed the efficacy of the accountability instruments they experienced. (See also Section 4.4, where I discuss participants' negative responses when asked what would happen if Singapore adopted Finland's teacher accountability approach, or vice versa.) In Mark's view:

> I really do think that the EPMS works well for the identification of teachers who are in need of support, and the identification of teachers who are in need of greater opportunity in order to stretch them for greater impact. I also think that, on a simple day-to-day level, […] it helps to guide the work of every general education officer and senior education officer in the system. […] Now, from an incentives perspective, it

> certainly incentivises hard work, if it's done fairly.

Like Mark, many other participants from Singapore agreed that the EPMS was effective on the whole, while indicating some degree of reservation about whether the system was uniformly fair. For example, Peter observed that the EPMS

> has to be taken with a pinch of salt as well, especially since certain things just cannot be captured by these measurements. But certainly it ensures that what you're doing is visible to others, to your reporting officer, and that it matters—it counts toward something. And it also incentivises people to really work hard, because just being seen to be skiving would ultimately be reflected in a lower performance grade.

Thus, despite the acknowledged limitations of the standardised EPMS in encompassing teachers' multifaceted work, this combination of monitoring and rewards does influence teacher motivation and practice considerably.

Nevertheless, the extent of this influence varies across teachers. Peter continued his observations by remarking,

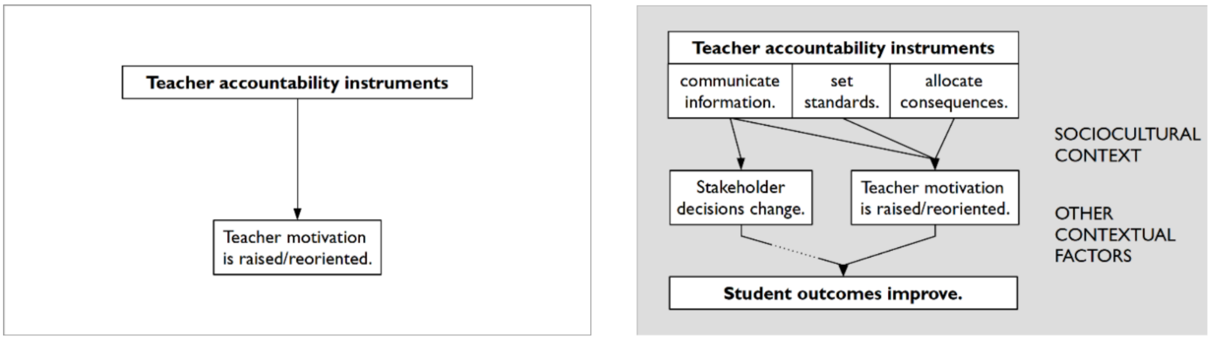
> Having said that, I think if someone is just set on cruising, she'll probably cruise regardless, you know? For teachers who give tuition [i.e. private tutoring] outside of school, they honestly may be fine with just getting a D for a performance grade, or a C. Doing the bare minimum that's expected in school, because they don't see that as a priority. So it does provide an incentive, but it incentivises people who want to do well anyway.

Later in this chapter, I explore how the effects of teacher accountability instruments depend on teachers' individual priorities and perspectives.

5.3 The effects of accountability instruments on teacher motivation in Finland and Singapore

Having described the accountability instruments that interview participants encounter, I turn to the effects that these instruments can have on teacher motivation. As indicated in Figure 5.4, in this section I focus on the effects of teacher accountability instruments jointly, rather than distinguishing between instruments that use different accountability mechanisms.

Figure 5.4 *Relationship between Section 5.3 and the overall conceptual framework*


Teacher accountability instruments
Teacher motivation is raised/reoriented.
Teacher accountability instruments
communicate information.
set standards.
allocate consequences.
Stakeholder decisions change.
Teacher motivation is raised/reoriented.
Student outcomes improve.
SOCIOCULTURAL CONTEXT
OTHER CONTEXTUAL FACTORS

This section

Overall conceptual framework

All 24 of the interview participants described multiple ways in which teacher accountability instruments influence their motivation or the motivation of the teachers around them. Their responses are summarised in Table 5.7. This table shows not only the intended effects of teacher accountability instruments, but also their undesirable side effects (Zhao, 2017, 2018). I classify intended effects as either raising teacher motivation (i.e. prompting teacher to work harder) or reorienting teacher motivation (i.e. prompting teachers to work differently, in ways that serve system-level goals). Side effects are: (a) demotivating teachers and generating stress; (b) introducing conflicting priorities, when accountability instruments prompt teachers to focus on tasks that they do not regard as central to their teaching practice; and (c) time-consuming paperwork, as a special case of conflicting priorities. Where there was ambiguity in whether an outcome was an intended effect or side effect, I relied on participants' judgements of whether a given effect was desirable or undesirable. I recognise that participants' evaluations of the desirability or undesirability of an effect may not match the intentions of those who initiated the accountability instrument. For the most part, however, there was little ambiguity about whether an effect was educationally and professionally positive or negative.

Note that this table does not include the effects of one-off or spontaneous accountability-related interactions among colleagues, because such interactions lack the regularity to qualify as 'tools, practices, or structures', as in my definition of teacher accountability instruments. However, the table does include self-initiated practices that take place among colleagues with some regularity, such as informally but regularly troubleshooting with colleagues teaching the same subject. Neither does it include the effects of point-of-entry standards (e.g. teacher training admissions criteria) that can influence the collective level of teacher motivation, because I did not ask participants about such standards directly. Also, the table makes a distinction between accountability instruments having an effect on participants themselves and having an effect on other teachers whom they know. However, in the discussion below, I do not draw a sharp distinction between these two categories, since this difference derives partly from participants' phrasing (e.g. 'we teachers feel …' would be categorised as 'self', whereas 'teachers tend to feel …' would be categorised as 'others'), and it is not central to this analysis.

Table 5.7 *Effects of teacher accountability instruments on teacher motivation in Singapore and Finland*

		FINLAND												SINGAPORE											
		Anneli	Liisa	Emilia	Kristiina	Masa	Satu	Maarit	Hannele	Helena	Antero	Päivi	Juhani	Adeline	Jeffrey	Maggie	Joseph	Peter	Andy	Sonia	Timothy	Mark	Eleanor	Geok Ling	Jane
Intended effects																									
Raise motivation	self						●	●		●	●	●		●		●		●		●	●		●		
	others		●		●	●	●									●	●	●	●	●	●	●	●	●	●
Redirect motivation	self	●	●	●		●	●	●	●		●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
	others				●	●	●	●		●		●	●	●	●	●	●		●	●	●	●	●	●	●
Side effects																									
Demotivate/generate stress	self		●					●	●							●	●			●	●		●		
	others		●		●	●	●	●			●	●		●	●	●	●	●		●	●	●	●	●	●
Conflicting priorities	self							●						●					●		●			●	
	others												●						●	●	●	●			
Time-consuming paperwork				●				●							●				●			●	●		●

Note. This table summarises interview participants' descriptions of the effects of teacher accountability instruments collectively, without distinguishing between particular instruments or mechanisms. The table does not include (a) one-off interactions that were autonomously initiated by teachers (e.g. spontaneous, informal encouragement); and (b) instruments that focus on the point of entry into the teaching profession (i.e. teacher training and selection).

As shown in Table 5.7, every interview participant mentioned instances of accountability instruments reorienting teachers' motivation toward particular tasks or goals. Also, every Singaporean participant and almost every Finnish participant reported instances where accountability instruments changed the level of teacher motivation, whether this desirably raised or undesirably lowered motivation levels. Most Singaporean teachers also observed that accountability instruments could introduce conflicting priorities and/or burdensome paperwork, whereas only a few Finnish participants reported such effects.

In Finland, motivation-raising instruments often took the form of informal feedback. For example, Antero said that 'the main motivator' was:

> when my principal says something good to me, that 'I am proud of you. You have done a very good job. Thank you for helping me.' The words, they are very effective.

A few Finnish participants also said that another aspect of teacher accountability that contributed to their motivation was the autonomy enshrined in Finland's approach to standard-setting. In Helena's words:

> [The accountability instruments] affect my work in that I have the freedom to do what I think is necessary for each group, for each subject, for each topic. […] If I don't have to tell anyone what my lesson objectives are, I can change them halfway, because it's Friday afternoon, and things just aren't working out. […] So it gives me the mental freedom to be flexible, and to change things, and to try new things.

While Helena did not explicitly say that this principle of autonomy raises her motivation, the inclination to 'try new things' does indicate healthy levels of motivation. As I discuss in Section 6.3, Helena's observation aligns with Deci and Ryan's (2000) argument that a sense of autonomy is crucial to intrinsic motivation.

In Singapore, many participants identified performance targets as a motivation-raising instrument. For example, Adeline said that:

> Without the target-setting of, 'You should at least try to aim for this,' I'm someone who very easily lowers my expectations to just suit whatever is enough to tide me over.

Middle manager Eleanor (Singapore) noted a change from her early days in her school, when 'we were very much like a *kampung* [i.e. village community]', to a recent push from the school board and the Ministry to enter a partnership with another school:

> And of course if you're pushed toward partnership, then who wants to marry you if your standards are not so high, right? (laughter) So we had to say, 'Okay, make sure that we get all these grades,' and we pushed our students to that level.

Thus, motivation-raising targets can operate at either the individual level, as in Adeline's account, or at the school level, as in Eleanor's.

Besides raising motivation levels, accountability instruments also reorient teachers' motivation and effort toward particular priorities. For most Finnish participants, this meant aligning their teaching with the curriculum. As Emilia said:

> Of course, I follow the curriculum. That's what we do. […] But how I actually choose to execute it in my classroom, that's up to me. And the main thing is that I can check at the end of the day that, 'Okay, I've covered this. And I think my students have learned how this is done.' But it's not something that I think about that much. (laughter)

Unlike this relatively loose alignment to accountability standards, Singaporean participants described more demanding alignments. According to Sonia:

> The EPMS is something that drives all the teachers in Singapore. A lot of us peg ourselves to the targets set for us, set by us, according to this EPMS framework, just to ensure that we do not end up being penalised. Because the penalty system if you don't

> meet their criteria can be rather severe, in my opinion.

For example, Sonia noted that 'these accountability instruments kind of force you to try out new things' because 'managers look favourably' on teachers who try new pedagogical tools. More constructively, several participants said that EPMS structures give them useful insight into where they should channel efforts to improve their work.

However, teacher accountability instruments can have negative side effects. Although Sonia (Singapore) said that a positive session with her reporting officer can make her feel 'affirmed', she added that:

> Sometimes, if you've just come out of a very *bad* review session—and when I say 'very bad' I mean that you feel the feedback is unfair—it does affect your motivation quite a bit. And sometimes you just feel like, 'Okay, I wish I could just quit and resign!'

Thus, EPMS instruments that aim to develop teachers' practice and raise their motivation can instead lower teachers' motivation. Moreover, Sonia viewed the 'pitting of one teacher against another' in the EPMS ranking process as 'the major flaw' of the system, because it makes teachers' work 'very, very stressful'. Most Singaporean participants similarly highlighted the ambivalent effects of EPMS, as well the pressure generated by its competitive rankings. As reporting officer Jane noted,

> It feels good to affirm someone, to help someone to grow. But to monitor a person and to give an adverse performance grade, it's very draining, honestly. It's draining.

Hence, EPMS rankings place considerable pressure not only on those receiving the grades, but also on those allocating them.[35]

In Finland, some participants mentioned that they can be demotivated not by onerous or unfair feedback, but rather by the absence of any feedback. In Hannele's words:

> I would like to have more, 'Thank you,' or, 'You did well.' But you can understand that these teenagers, they don't normally [...] say that. So I'm hoping that the headmaster would sometimes come to the classroom, and know more about your work, and give you feedback. Good or bad.

[35] To alleviate such pressure, some reporting officers award adverse D grades not strictly on the basis of performance, but based on the anticipated consequences of such grades or other justifications. For example, the Singapore Teachers' Union (2009) observed that D grades may unexpectedly and undeservedly be awarded to teachers who are transferring to another school (and thus will no longer encounter their former reporting officer face-to-face), teachers who are approaching retirement (and thus will not be impacted by promotion-related penalties), or teachers who have been on maternity leave for part of the year (and thus had not contributed as much to annual school outcomes as their colleagues had). The latter strategy was corroborated by one of my Singaporean friends, who said that her mother, a long-time reporting officer, always felt relieved whenever any of the teachers who report to her went on maternity leave, because this meant that she could award her expected quota of D grades with fewer qualms.

Päivi mentioned similar sentiments among some of her colleagues, while saying that she did not feel such a need herself, because she valued feedback from 'pupils and their mothers and fathers' over feedback from the principal. However, Liisa cautioned that parental feedback is 'not always positive', having witnessed Finnish colleagues who 'got such bad feedback' at a parents' day 'that they were crying afterwards'.

Besides potentially causing undue stress, another side effect is that teacher accountability instruments can impose priorities that conflict with teachers' personal goals. As shown in Table 5.7, this was more prevalent among Singaporean participants. (Similar observations have been made in prior studies of Singaporean teachers, e.g. Loh, 2016; Loh & Hu, 2014.) According to Timothy:

> EPMS has many different criteria, right. So, as with many teachers, I find it a bit hard to focus on what, to me, is the essence of what I'm trying to do, when there are expectations of you in other areas. Like, 'Oh, have I shared this with someone yet?' And then, 'Do I need to be a "knowledgeable other" to help this other colleague doing blah blah blah?' […] Or, 'Have I made sure that I have enhanced the character of my CCA [co-curricular activity] students?' […] By nature, the job of a teacher now—at least in our school system—very much requires multitasking.

Singaporean participants were especially critical of the numerous non-teaching requirements of their jobs. As Geok Ling put it, 'I didn't come into this job to be an event organiser.' However, Mark accepted that such organisational responsibilities were part of the job, noting that, 'MOE calls us education officers—not teachers—in their official parlance.' He estimated that teaching-related work 'probably only takes up 40 percent of a teacher's time', with the balance going to committee work, co-curricular tasks, and the like. Accountability-related expectations also affect time allocations within the classroom. Adeline noted that, in a 'time crunch' when she has to 'meet all these teaching expectations', then her teaching can get 'a bit rushed', or she may decide 'not to do a group activity, instead I will just do frontal teaching'. More generally, several participants bemoaned the volume of paperwork they faced, whether from EPMS structures (e.g. filling in EPMS forms or compiling students' work to be reviewed by a reporting officer) or other administrative processes (e.g. Jeffrey enumerated 12 required steps for booking a bus for a student excursion).

While Finland's lighter accountability system meant that most participants did not mention any tension between personal priorities and accountability expectations, Maarit noted that her work as an assistant headteacher sometimes impinged on the time that she would have otherwise used for lesson planning. She added that teachers had to fill in '*all* kinds of forms on all kinds of students', which 'takes a lot of effort'. Compatriot Emilia likewise mentioned that she was required to fill in a 'massive amount of different questionnaires', which sometimes 'gets quite tiring'.[36] Additionally, Juhani said that some Finnish teachers inflate students' grades in order to appease parental demands—another perverse side effect of accountability.

Apart from the motivation-related effects that I focus on in this section, Singaporean participants identified some other side effects from their country's approaches to teacher accountability, which are not included in Table 5.7. For example, Timothy noted that the competitive EPMS rankings can reduce collegiality among teachers—although, as mentioned in the previous section, Andy and some other interview participants would dispute this. Also, the added pressure of higher-level responsibilities can make some teachers reluctant to be promoted (as noted by Sonia, Jeffrey, Eleanor, and Timothy), and prompt some managers to coerce their subordinates into additional tasks that bolster the manager's performance ranking (as noted by Joseph and Andy).

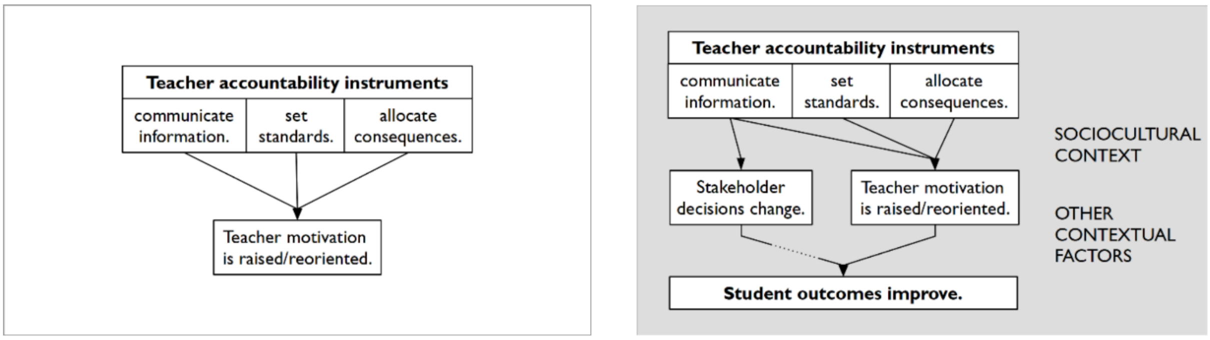
5.4 Mechanisms, perceptions, and pathways in teacher accountability

In addressing the question of *how* teacher motivation mediates the influence of teacher accountability instruments on student outcomes, I adopt a realist perspective that looks for the mechanisms that actually generate these changes in teacher motivation. (For the theoretical basis of this approach, see Sections 2.5 and 3.1.) I posit three key mechanisms through which accountability instruments aim to influence teacher motivation: setting standards, collecting information, and allocating consequences. Each of these mechanisms influences teacher motivation in different ways, as discussed in Section 2.5. In this section, I explore interview participants' experiences of how each mechanism can influence teacher motivation, as indicated in Figure 5.5. Accordingly, I place less emphasis on comparison between Finland and Singapore, and more emphasis on how teacher accountability mechanisms affect teacher motivation across both contexts. (As noted in Section 2.5, these mechanisms that link accountability instruments to changes in teacher motivation constitute only one step along the pathway to influencing student

[36] Emilia said that some of the questionnaires came from the central government, whereas some came from the municipality, and others were only for Swedish-speaking schools within her municipality. Some questionnaires were also issued by external organisations that were conducting research. I do not know definitively why teachers in Emilia's school faced so many questionnaires while most other interview participants did not. Possible reasons include the management approach of her municipal officers; as well as the fact that she taught at a Swedish-speaking school. Since there are far fewer Swedish-speaking than Finnish-speaking schools, the former may be more likely to be selected for sample-based reporting processes.

outcomes. Unfortunately, given time and space constraints, I was unable to collect any empirical data that connect changes in teacher motivation to changes in student outcomes. I address this evidence gap further in Section 5.5 and in Chapter 7.)

Figure 5.5 *Relationship between Section 5.4 and the overall conceptual framework*


Teacher accountability instruments
communicate information.
set standards.
allocate consequences.
Teacher motivation is raised/reoriented.
Teacher accountability instruments
communicate information.
set standards.
allocate consequences.
Stakeholder decisions change.
Teacher motivation is raised/reoriented.
Student outcomes improve.
SOCIOCULTURAL CONTEXT
OTHER CONTEXTUAL FACTORS

This section

Overall conceptual framework

Table 5.8 summarises the accountability mechanisms that each interview participant mentioned, grouped by whether the mechanism affected teacher motivation in intended or undesirable ways. As such, this table is a form of cross-tabulation of Table 5.6, which grouped teacher accountability instruments by their underlying mechanisms, and Table 5.7, which categorised the effects of teacher accountability instruments. Crucially, every motivational change that participants identified could be traced to one of these three mechanisms, or to two or more of these mechanisms operating in concert. This suggests that the mechanisms presented in the conceptual framework may be an adequate characterisation of how teacher accountability instruments influence teacher motivation, at least in these two countries (given that the interviews reached saturation in each country).

Among Finnish participants, all identified effects from instruments that use the standard-setting mechanism, as shown in Table 5.8. Such standards can have considerable sway over the direction of teacher practice, even in the absence of formal monitoring or rewards and penalties. Liisa said that:

> Now we have a new curriculum in Finland. […] The principal is not directing the work in such a way that, 'You have to meet at 2 o'clock tomorrow to discuss this new curriculum.' But we just know that we're supposed to do it. And then I gathered people together and we discussed it and selected books for it. So we're very independent.

Standards-setting instruments can exert a similarly powerful effect on teacher practice in Singapore. Maggie noted that Singaporeans like 'knowing exactly what is expected', and that 'they *really* like to do it to the letter […] even if it's optional'. Maggie viewed this desire to comply exactly with standards as 'very unhealthy, really'. Thus, Liisa (Finland) and Maggie (Singapore) represent two different contexts, and two different opinions on the merits of complying with standard-setting instruments, but they both recognise the potential efficacy of these instruments in orienting teachers' work. Standalone standard-setting instruments operate passively in concert with teachers' pre-existing motivations—as in Liisa's (Finland) account of changing her practice in response to the new curriculum simply because she knew that she was 'supposed to do it'.

Table 5.8 *The mechanisms underlying teacher accountability instruments, by their intended and unintended effects on teacher motivation in Finland and Singapore*

● The teacher accountability instruments that use this mechanism can have these effects.
◑ These effects exist, but such accountability instruments are self-initiated by teachers.
○ These effects exist, but they result from the absence rather than the presence of such accountability instruments.

	FINLAND												SINGAPORE											
	Anneli	Liisa	Emilia	Kristiina	Masa	Satu	Maarit	Hannele	Helena	Antero	Päivi	Juhani	Adeline	Jeffrey	Maggie	Joseph	Peter	Andy	Sonia	Timothy	Mark	Eleanor	Geok Ling	Jane
Intended effects from:																								
Setting standards	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Communicating information		●	●	●	◑	●	◑		◑	●	◑	◑	●	●	●	●	●	●	●	●	●	●	●	
Allocating consequences													●		●	●	●	●	●		●	●	●	●
Side effects from:																								
Setting standards					●		●				●		●	●	●	●		●	●	●	●	●	●	●
Communicating information		●						○			○	◑	●	●	●	●	●	●	●	●	●	●	●	●
Allocating consequences		○		○		○				●				●	●	●	●	●	●	●				●

In turn, informational instruments influence teacher motivation via their consciousness that stakeholders are actively comparing them to a set of standards, as noted in Section 2.5. When asked how accountability instruments affect his work, Timothy (Singapore) responded:

> It's always at the back of one's mind. [...] Yeah, because you know that there are other people who are watching you, so to speak. [...] I would consider it a negative motivation, or not the most desired form of motivation. But at least it helps to spur you on when you are drained.

Beyond my field interviews, a similar observation was made by another Singaporean teacher: 'I feared to let my guard down [in front of my reporting officer] as I felt that I was constantly being assessed' (Ahmad, 2016, p. 111). Finnish participant Maarit offered a more benign example of the informational mechanism:

> A week ago, the special education teacher advised me that maybe these two students shouldn't sit so near each other. Because I'm hopeless at doing very strict seating orders. (laughter) So that was a hint. Of course I knew it, but I just didn't care. (laughter) But now I try to be more strict, so that they aren't sitting together.

Thus, the awareness that a colleague was aware of her non-optimal seating arrangements prompted Maarit to channel more effort in that direction.

Two observations are worth noting at this point. Firstly, as with the standard-setting mechanism, the effects of the informational mechanism are not uniformly positive. For example, Liisa (Finland) said that informal monitoring by colleagues can lead to excessive conformity to 'certain ways' in which teachers are expected to behave, because 'it's like there are a lot of little policemen around you'. She added that it was 'easier' to conform to behavioural expectations, at least 'if you want to be friends with your co-workers'. Additionally, Liisa mentioned that, despite believing in the value of integrating technology into learning, she no longer assigned homework that required the use of a computer, because one parent had protested angrily because they did not want their child to use computers at home. Similarly, Finnish participants Masa and Juhani, as well as Singaporean participant Maggie, reported making what they viewed as peripheral modifications to their teaching practice in order to accede to parental feedback.

Secondly, from interview participants' accounts, there is no clear distinction between the mechanisms underlying instruments that communicate information informally (and/or to colleagues), and those that communicate information for official or codified purposes (and/or to managers and external stakeholders). As noted in Section 2.5, I had initially posited a mechanism that affected teachers' motivation via their desire to maintain professional reputations and collegial camaraderie, as distinct from the main informational mechanism. I anticipated that the reputational mechanism would be associated with informal, social monitoring, whereas the main informational mechanism would be linked to formalised instruments. However, in the interview data there was no obvious analytical difference between the motivational mechanisms in participants' descriptions of reluctantly changing their practice in response to informal monitoring (as in Liisa's account of how 'it's like there are a lot of little policemen around you'), as compared to monitoring linked to codified instruments (as in Timothy's account of how 'you know that there are other people who are watching you, so to speak'). There was also little difference between participants' descriptions of instances where they welcomed accountability-related feedback that was informally offered by fellow professionals, compared to instances where such constructive feedback emerged from formal structures or from managers and parents. Accordingly, I concluded that all of these instruments worked via teachers' desire to be regarded favourably when information about their practice, whether formally or informally collected, is benchmarked against stakeholder expectations, whether the stakeholders are colleagues or others.

The final mechanism through which accountability instruments can influence teacher motivation entails the allocation of consequences for teacher practice. The force of such instruments arises from the desire to gain rewards and avoid punishments. As Geok Ling remarked on Singapore's EPMS:

> You know that, 'Okay, they're going to grade me like this, and this is how I can get more because of the bonus.' So people will drive their behaviour towards the criteria. […] But you can argue that it's a necessary devil; it's a double-edged sword. While it may raise the standard and it motivates some, because they see it as an affirmation, it will also demoralise some.

Other Singaporean participants echoed her observations both on the strong influence of EPMS consequences, and on their varied effects on teacher motivation. In contrast, as shown in Table 5.8, most Finnish participants did not mention any consequence-related effects, since such instruments are sparse in their context. However, Antero noted that Finland's salary supplements, though small, could make colleagues 'a little bit jealous, of course'. On the other hand, compatriots Kristiina, Satu, and Liisa mentioned that the absence of performance-based consequences (and the fact that many teachers hold permanently tenured positions) meant that some teachers lacked the motivation to work hard.

Teacher accountability instruments often invoke multiple, interacting mechanisms for influencing teacher motivation. Singaporean participant Peter, who had recently been promoted to a middle-management role, said that:

> If my job title is in some ways an accountability instrument, having that subject head position now is something that's at the back of my mind. […] It has changed my outlook of who I am as a teacher. […] Because I know that my performance is going to be evaluated in the lens of being a subject head, I guess that shifts my focus a little when it comes to how much I'm involved in things that aren't subject-related.

Thus, the expectations (standard-setting) associated with being a subject head alongside the awareness that his 'performance is going to be evaluated' (information-communicating) had influenced Peter to the extent of 'chang[ing] [his] outlook of who [he is] as a teacher'. This is noteworthy since Peter explicitly said elsewhere that he did not value the consequences allocated under the EPMS system.[37] Nonetheless, one such consequence—his promotion to subject headship—operated in tandem with other accountability instruments to channel his motivation in accordance with EPMS performance standards.

The linchpin: how teachers regard accountability instruments

As shown above, teacher accountability instruments can shape teacher motivation via different mechanisms, and these mechanisms can have substantial effects. However, as argued in Section 2.4, a teacher accountability instrument will only influence teacher motivation as intended if teachers regard the instrument as sufficiently persuasive. This persuasiveness can come from a range of sources, such as: regarding performance standards as being aligned with one's personal beliefs about teaching, valuing the esteem of the stakeholders who monitor teacher practice, or desiring the offered rewards. To illustrate, Sonia (Singapore) said that Ministry expectations of teachers 'converge quite a bit' with what she prioritises in her profession, thus indicating that at least some of the government's standard-setting instruments are meaningful to her. At another point in the interview, she said, 'Singapore has this idea of pride, you know. If a teacher were to get a D grade, there is a lot of hurt to their pride.' This indicates her expectation that the typical Singaporean teacher would regard EPMS performance grades as a compelling informational instrument to which they want to measure up. Furthermore, when discussing performance bonuses, Sonia said, 'Let's face it: over here, people are money-motivated—we need the money.' This suggests that EPMS bonuses are a desirable incentive for good performance.

Potential influence notwithstanding, teachers have diverse views of accountability instruments. This diversity is apparent in their heterogeneous responses to the same set accountability instruments. When discussing Finland's light-touch approach to teacher accountability, Satu observed that:

> The freedom for me is like, 'I can do everything.' (laughter) 'I can make a hundred exercises for the students if I want to.' And some of my colleagues think that even one exercise is too much. So the freedom has two sides.

Thus, a single accountability instrument can have divergent effects on teacher motivation. Similar encounters with colleagues who put in minimal effort were related by Finnish participants Maarit, Liisa, and Hannele, as well as Singaporean participants Jane, Peter, Maggie, Joseph. Also,

[37] Specifically, Peter said elsewhere in the interview that informal affirmations from students give him 'greater validation than seeing that B grade for my performance at the end of the year'. Furthermore, in responding to the hypothetical question about Singapore adopting Finland's teacher accountability approach, he said that the removal of performance bonuses 'wouldn't make a difference for me; I would want to teach well anyway'.

Joseph noted that different teachers may direct their energies toward different priorities, even under the same EPMS standards:

> It depends on individual teachers and their convictions. For me, I believe that nurturing the child's character is more important. So if I need to, I would focus more on the child rather than on the lesson.

He drew a distinction between teachers who join the profession because they intrinsically value teaching, and those who 'are in it for the money only'. Thus, whether in Singapore's extensive in-service accountability system or Finland's front-loaded, standards-focused approach, teachers' personal perspectives on accountability instruments shape their responses to accountability instruments—and, consequently, the efficacy of these instruments within the school system.

One implication of this is that teachers who do not genuinely agree with a standard-setting accountability instrument may only comply with it superficially. As Maggie (Singapore) said,

> Let's say you have a top-down policy implementation. If you didn't trust [the teachers] with the discussion beforehand, then when you do implement it, they would just do whatever it takes to survive, and it might not turn out the way you want it. Like, they will modify it, just to placate you and show you some semblance of what you want to see.

Singaporean participant Eleanor agreed. When discussing her school's attempt to promote a new pedagogical approach, she noted that 'you need to have buy-in' from teachers. In the absence of such buy-in, that teachers would follow managers' recommendations during lesson observations 'just for show', but would revert to routine pedagogies at other times, as Eleanor observed. Such superficial implementation can occur partly because teachers, like other client-facing public servants, serve multiple goals that cannot easily be monitored, as noted in Section 2.1. Moreover, teachers often face more expectations than can be feasibly fulfilled—and they differ in how selectively they prioritise their work. Päivi (Finland) noted that some of her younger colleagues felt exhausted from attempting to meet every stated standard and expectation, whereas her stance was that, 'you do those things that you feel are important, and you do those things well', because 'no one can make [her] job or [her] situation any worse' if she failed to meet their excess expectations. (Päivi added the caveat that 'if you do something very, very wrong, then it's okay [for other stakeholders] to say that you can't do this work', i.e. that you should leave the profession. In her view, professional autonomy should be enjoyed and safeguarded, but it was not unlimited.)

In turn, consequence-based accountability instruments may fail to deliver their intended effects if teachers do not regard the incentives as sufficiently appealing. For example, Jane (Singapore) said that:

> Sometimes the 'smart' ones will think, 'Why do you want to put in so much effort and time? I might as well cruise a little, get a C grade, and give tuition [i.e. private tutoring].'

Elsewhere, Jane commented that despite the magnitude of EPMS performance bonuses, 'tuition is much more lucrative'. (Peter made a similar remark, as quoted in Section 5.2 above.) Estimates from the Singapore government's Household Expenditure Survey 2017–18 suggest that private tuition was a S$1.4 billion (approximately US$1 billion) industry in 2018, with the average household spending more money on private tuition than on clothing (Singapore Department of Statistics, 2019; Teng, 2019). Besides financial considerations, Jeffrey (Singapore) noted that 'some teachers aren't keen to be promoted up the career ladder', because 'as a senior teacher you have to take on the responsibility of mentoring others, which most people can't be bothered with because they'd rather just teach' (see Ang, 2016, for a similar sentiment). Thus, for such teachers, the prestige and salary increment of a promotion does not outweigh the undesirable burden of responsibility.

Additionally, both consequential and informational accountability instruments may prove ineffective when they are perceived as unfair. When asked how the EPMS affects her motivation, Maggie said:

> When it's used correctly, it motivates me a lot, regardless of whether I just get a normal C grade or not. But when you don't feel it's fair, […] then you feel trapped. You feel that whatever effort you put in is not worth it.

These sentiments were echoed by compatriots Timothy, Sonia, Peter, and Joseph. I discuss such perceptions of fairness and their interaction with Singapore's sociocultural context in Section 6.3.

The above examples are not a comprehensive catalogue of ways in which teacher accountability mechanisms fail to have their intended effects. Rather, they demonstrate that these mechanisms are contingent on the perspectives of those whom they seek to influence.

An alternative pathway: changing stakeholder decisions

As detailed above, teacher accountability instruments have the potential to affect the level and orientation of teacher motivation extensively. However, as posited in my conceptual framework, there is another pathway by which teacher accountability instruments can affect student outcomes. Besides the teacher motivation pathway, accountability instruments can also communicate information on teacher practice to stakeholders, who then use this information to make decisions that aim to optimise student learning. Whereas changes in teacher motivation can result from the informational, consequential, or standard-setting mechanisms, changes in stakeholder decisions depend primarily on the informational mechanism, as noted in Section 2.5.

I did not explicitly ask participants about this second pathway, but some addressed it nonetheless. Satu (Finland) mentioned some municipal feedback mechanisms, remarking that:

> It's just a way to transfer the results to higher levels, but it doesn't come down to me again. […] The municipality gives the money and the guidelines about what work you have to do. So they have to follow up about whether everything is okay, and if there are things to do better, in the future. It's important. But most of the teachers continue teaching the same way, and it doesn't affect them so much.

Similarly, Andy (Singapore) said:

> I don't think that [the EPMS] actually has much influence on me. I understand its administrative purpose, as a record of what I have done. […] To me, it's just paperwork to fulfil that administrative function.

Thus, Satu and Andy recognised that certain stakeholder-oriented teacher accountability instruments can communicate information for useful system-level purposes, even if such instruments have no direct effect on their work.

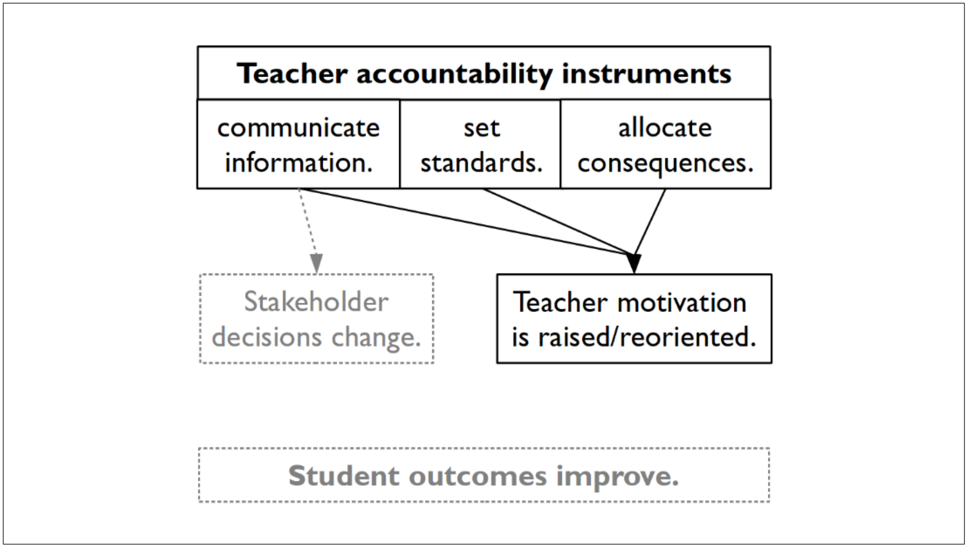
5.5 Discussion

Overall, the two empirical sources discussed in this chapter paint different pictures. The TIMSS 2015 mediation analysis does not provide any clear evidence for the outcome-oriented aspect of RQ2, i.e. '*to what extent* does teacher motivation mediate the influence of teacher accountability instruments on student outcomes?' Adding teacher motivation to model 2 does not noticeably alter the relationship between teacher accountability and student outcomes; thus suggesting that teacher motivation does not, in fact, play a mediating role here. However, the model 3 results do suggest that teacher accountability may, in some contexts, be associated with teacher motivation, thus giving a basis—albeit a weak one—for the first step of the mediation pathway. I examine this first step more closely in the next chapter, exploring the relationships between teacher accountability, teacher motivation, and sociocultural context.

Turning to the second empirical source, the interviews with Finnish and Singaporean teachers shed light on the process-oriented aspect of RQ2, i.e. '*how* does teacher motivation mediate the influence of teacher accountability instruments on student outcomes?' As with the statistical analysis, the interviews do not offer evidence for the relationship between teacher motivation and student outcomes. However, unlike in the statistical analysis, this shortcoming of the interviews was by design, given the time and space constraints of a PhD project. Instead, the interviews offer considerable evidence that teacher accountability instruments can affect the level and orientation of teacher motivation, with every interview participant describing some effects. Moreover, the interviews clearly demonstrate the three accountability mechanisms posited in the conceptual framework, while illustrating how these mechanisms are contingent on teachers' perspectives. The interviews also indicate agreement from some participants that certain teacher accountability instruments aim to influence stakeholders' teacher-related decisions rather than teacher motivation itself. However, they did not give convincing evidence of teacher accountability instruments beneficially changing stakeholder decisions—although, again, the interviews were not designed for this aim.

Figure 5.6 summarises the evidence for the conceptual framework from these ILSA and interview analyses. The evidence in this chapter makes the case that teacher motivation can be changed via the three mechanisms underlying teacher accountability. However, my data cannot credibly connect changes in teacher motivation to improvements in student outcomes.

Figure 5.6 *Evidence for the conceptual framework from the RQ2 analysis*


Teacher accountability instruments
communicate information.
set standards.
allocate consequences.
Stakeholder decisions change.
Teacher motivation is raised/reoriented.
Student outcomes improve.

Hence, this chapter's answer to RQ2 may be that (a) it is unclear how much teacher motivation mediates the influence of accountability instruments on student outcomes; but (b) teacher accountability instruments have the potential to influence teacher motivation, and mechanisms for such influence can include setting standards that orient teachers toward particular goals, by communicating information that activates teachers' desires to compare favourably with expectations about their practice, and by allocating performance-based consequences that incentivise certain efforts or practices.

These interviews complement prior evidence that accountability instruments can play a positive causal role in raising teacher effort (Lavy, 2009; Macartney et al., 2018), as discussed in Section 2.4, and in reorienting teachers' decisions toward policy goals (Bassok et al., 2016; Mausethagen, 2013), as noted in Section 2.5. However, my analysis also echoes the finding that accountability instruments can raise teachers' stress levels (von der Embse et al., 2016) besides triggering other undesirable side effects, as discussed in Section 2.1.

Still, it is important to note that some accountability instrument may fail to have any effect on teacher motivation and practice in certain settings (e.g. Yuan et al., 2013). As observed in the Finland and Singapore interviews, the degree to which intended and unintended effects may be triggered depends partly on teachers' subjective perspectives and priorities—a finding that also emerged in some studies reviewed in Section 2.4 (e.g. Broekman, 2016; Mizel, 2009; Müller & Hernández, 2010; Narwana, 2015).

Another possible explanation is that teacher accountability instruments primarily affect student outcomes by influencing the profile of people who choose to become teachers, rather than by influencing the level or orientation of teacher motivation in-service. That is, my framework may be mistaken in positing a teacher motivation pathway, because the relationship between teacher accountability instruments and student outcomes may instead occur primarily via a self-selection pathway (Biasi, 2018; Lazear, 2000, 2003; see also Section 2.5 in this thesis). Some interview participants pointed to such a self-selection pathway in their responses to the hypothetical question. When asked what would happen if Finland adopted Singapore's EPMS, Masa said:

> If that system was in place, it would attract a whole different kind of person. With different goals, different ambitions that I don't think are in line with what Finnish teachers currently consider to be their job.

This sentiment was echoed by Helena (Finland) and by Andy and Jeffrey (both Singapore).

Despite the theoretical plausibility of a selection effect that dominates any motivational effects, this selection effect was not evident in my interview data. Table 5.9 summarises participants' responses to the question, 'Why did you decide to become a teacher?' Responses were coded based on the Factors Influencing Teaching Choice framework (FIT-Choice; Watt et al., 2012), with modifications made inductively during coding, as noted below the table.

Table 5.9 *Summary of interview participants reasons for becoming a teacher*

	FINLAND												SINGAPORE											
	Anneli	Liisa	Emilia	Kristiina	Masa	Satu	Maarit	Hannele	Helena	Antero	Päivi	Juhani	Adeline	Jeffrey	Maggie	Joseph	Peter	Andy	Sonia	Timothy	Mark	Eleanor	Geok Ling	Jane
Perceived teaching ability			●						●	●														
Intrinsic value	●	●	●	●	●	●			●		●		●		●	●	●	●			●		●	
Personal utility value		●	●						●			●			●	●		●		●	●			●
Social utility value														●			●		●			●	●	
Work with children/adolescents	●							●		●	●													
Prior experiences		●		●	●						●		●				●		●		●	●		
Social influences													●	●	●	●								
Interest in subject			●			●	●	●		●		●		●						●		●		

Note. Shading demarcates groups of participants who entered the profession under different phases of the country's teacher accountability approach. Categories were based on the Factors Influencing Teaching Choice (FIT-Choice) framework (Watt et al., 2012), with some modifications added inductively while coding the data. Modifications were: (a) broadening the 'Intrinsic value' category (originally: 'I am interested in teaching'; 'I like teaching') to include specific aspects of the profession that do not fall under 'Personal utility value', such as 'I wanted to interact with people' and 'I wanted to have autonomy over my work'; (b) broadening the 'Personal utility value' category (originally: 'Job security'; 'Time for family') to include the availability of teacher training programmes and scholarships; (c) broadening the 'Prior teaching and learning experiences' category (originally: 'I have had inspirational teachers'; 'I have had good teachers as role models'; 'I have had positive learning experiences') to a general 'Prior experiences' category that includes positive teaching experiences in non-classroom contexts as well as experiences in other fields to which teaching compared favourably; (d) adding an 'Interest in subject' category for instances where passion for the subjects taught was a key motivating factor.

As with the other interview matrices, respondents are sorted, within countries, from those who most recently entered the profession (i.e. Anneli and Adeline), to the longest-serving teachers (i.e. Juhani and Jane). Among Finnish participants, two (Päivi and Juhani) became teachers prior to the mid-1990s shift away from school inspections and other standardised tools for monitoring teachers' classroom practice (Müller & Hernández, 2010; Simola et al., 2017). Another seven entered the profession before the 2001 release of the first PISA results that propelled Finnish education to international admiration, thus silencing local criticism of Finnish teachers (Simola, 2014); while three participants (Anneli, Liisa, and Emilia) joined the profession after these two major events. Among Singaporean participants, three (Eleanor, Geok Ling, and Jane) became teachers prior to introduction of EPMS; six entered the profession when EPMS reporting was in its first iteration, which had more complicated performance standards than the current version, and which emphasised numerical targets and ratings; and three (Adeline, Jeffrey, and Maggie) have only experienced the most recent version of EPMS reporting, which was implemented in 2014 (Kan, 2014; Low & Tan, 2017). These groupings are demarcated by shading in Table 5.9.

Within each country, there were no clear differences across the groupings in participants' reasons for becoming teachers. Among Singaporean participants, the only discernible difference is that those who entered the profession during the most recent iteration of EPMS, together with Joseph from the middle group, all identified social influences in their decisions. In the FIT-Choice framework, social influences denote friends, family members, or former colleagues thinking that you should become a teacher (Watt et al., 2012). There are no obvious reasons why Singapore's current accountability approach would be favoured by people who are more inclined toward social influences. However, it is possible that these participants were more attuned to social influences since they were relatively young. Among Finnish participants, the only discernible difference between groups is that those who became teachers more recently all identified the intrinsic value of the profession as a motivation. However, this was the most commonly cited reason among both Finnish and Singaporean participants. Moreover, these recent Finnish entrants were no less likely to also mention the personal utility (i.e. extrinsic) value of the profession as well. Hence, these within-country comparisons do not offer evidence of a selection effect with changes in teacher accountability instruments over time.

A between-country comparison does show some differences in reasons for becoming a teacher—but these run counter to the differences that would be expected given their respective accountability approaches. Specifically, some Singaporean participants (but no Finnish participants) mentioned social utility value; whereas some Finnish participants (but no Singaporean participants) mentioned their perceived teaching ability. This is contrary to a self-selection framework, in which Singapore's performance-based consequences would presumably attract people who believe in their abilities to teach, and Finland's lack of material rewards would presumably attract people with altruistic (i.e. social utility) goals.[38] In addition to these differences, some Singaporean participants, but no Finnish participants, mentioned social influences; whereas some Finnish participants, but no Singaporean participants, mentioned the desire to work with children/adolescents. However, there are no obvious connections between these reasons and either country's configuration of teacher accountability structures.

That said, none of this invalidates the possibility of a self-selection effect. Indeed, Liisa (Finland) and Adeline (Singapore) mentioned accountability-related job characteristics in their career

[38] However, this difference is aligned with Finnish and Singaporean participants' respective concepts of motivation. I analyse these concepts of motivation in Section 6.3.

decisions. Liisa, who had previously worked in a large corporation, said that she moved into teaching because she wanted a role 'that was more in [her] control, and that was more equal', which aligns with the professional autonomy and egalitarian outlook of the Finnish teaching profession. Adeline did not mention any accountability-related elements when I asked why she decided to become a teacher, but her answer to the Finland hypothetical emphasised the appeal of EPMS career tracks in her professional decisions. When asked how she would react to Finland's teacher accountability approach, she said,

> I think I might not stay in teaching as a lifelong career. For myself, I entered teaching on the MOE teaching scholarship, and so I entered knowing that, 'Oh, okay, after two and a half years, I will get another opportunity to see the other side of education in Ministry work.' And so the thought of being in my current position for a long, long time—maybe I'm like the semi-millennial generation, which makes me feel like I probably might think of switching out to see something else down the road.

Thus, Liisa's and Adeline's observations quoted here suggest that teachers' decisions to enter the profession may depend, to some extent, on accountability instruments. Furthermore, my interview samples were small, which particularly hampers the within-country comparison. Also, as noted in the methodology chapter, the samples may be biased because they comprise teachers who were sufficiently motivated to take the time to speak to a researcher about their work.

Still, weighing the evidence solely from these interviews yields more extensive support for a motivation pathway through which accountability instruments influence teacher practice than for a self-selection pathway. I am not denying the possibility of a selection effect, but rather arguing that this possibility cannot in itself disqualify the evidence in Sections 5.3 and 5.4 for an in-service motivational effect. Further, I am arguing that it is important to pay attention to the intended and unintended effects of teacher accountability instruments on teacher motivation, and to the mechanisms underlying these effects.

Chapter 6: The influence of sociocultural context on teacher accountability and teacher motivation

Thus far, I have used cross-country survey data in Chapter 4 to show that the relationship between teacher accountability instruments and student outcomes may be affected by some aspects of sociocultural context. In Chapter 5, I used interview data to show that teacher accountability instruments can considerably but contingently influence teacher motivation. In this chapter, I turn to the third research question: *to what extent does the influence of teacher accountability instruments on teacher motivation depend on sociocultural context?* (As noted in Section 3.1 and at the beginning of Chapter 4, the term 'influence' here does not imply causal inference in a statistical sense, but rather in a realist sense.)

I explore this question using both cross-country surveys and teacher interviews. As with RQ1, I use a moderation model to analyse the cross-country survey data, comparing the moderating influence of sociocultural constructs to the influence of other contextual variables (Section 6.1). The main dataset for this analysis is TALIS 2013. Next, I analyse interview participants' experiences of their respective sociocultural contexts (Section 6.2). Then, bringing the key concepts together, I examine how sociocultural context in Finland and Singapore influences the effects of accountability instruments on teacher motivation (Section 6.3). Finally, I discuss these observations and weigh the evidence for and against an alternative explanation (Section 6.4).

Articulating and analysing sociocultural patterns is hardly a straightforward process, as discussed in Chapter 3. In the first place, culture is not only a construct, but a multifaceted and malleable one. For example, Eleanor commented on the interplay between different sociocultural patterns that have been variously prioritised in Singapore:

> Singapore is interesting, because while we try so hard to hold on to heritage and culture, the new culture that we have inadvertently created—because we were so focused on progress and meritocracy—is of course achievement and excellence. And sometimes these old values get sidelined because we're trying to just get ahead, you know, to just get to the next step.

Moreover, there can be a gap between the cultural values that people espouse and how they actually act. (For one example, see Juhani's and Päivi's observations about Finland's egalitarianism in Section 6.2 below.) As noted in Section 3.6, this epistemological gap between articulations and reality poses a challenge for the self-report data in both the cross-country surveys and the field interviews. This challenge is heightened by the fact that many aspects of sociocultural context can be tacit, taken for granted, or subconscious. As Anneli (Finland) reflected, 'There are so many things I want to say, but I don't know how to put them in words. […] It feels like I'm blind to my own culture, somehow.' The cross-cultural nature of this research project adds further complications, because my interpretations of participants' sociocultural descriptions may differ from their own.

With these challenges in view, I draw on multiple data sources and perspectives in an attempt to construct observations that accurately represent the contexts in question. Some of these comparative perspectives are internal to my main empirical sources. The analysis in this chapter includes a comparison between interview participants' views and statistical aggregates from the WVS/EVS and Hofstede surveys. Also, I have been aided greatly by the observations of interview participants who have spent substantial amounts of time in other countries, and consequently had both insider and outsider lenses on the sociocultural context in which they taught. Such participants include Mark and Masa, who moved from the United States to Singapore and Finland, respectively, as adults. Additionally, Maggie grew up in Malaysia, completed secondary school in Singapore, attended university in the United States, married someone from another Asian country, and currently teaches in Singapore. Helena is Finnish and spent her childhood in Finland, but completed her 'A' levels and undergraduate studies in the United Kingdom, before returning to Finland for teacher training. Also, as noted in Section 3.4, one of the Finnish participants recently left the classroom and to work in international education. To corroborate the interviews, I also cite prior research on these contexts.[39]

Notwithstanding these methodological challenges of analysing sociocultural context—and, as noted in Section 3.5, the ethical implications of doing so inadequately—I believe that such analysis is worthwhile. Whatever the epistemological complications of studying sociocultural context, I take the stance that sociocultural context is ontologically real and has real implications for teacher practice and educational quality (as discussed in Section 3.1 on the ontological and epistemological commitments of this thesis). As long as the interview participants were honestly sharing their sociocultural articulations with me, then even if these articulations may be

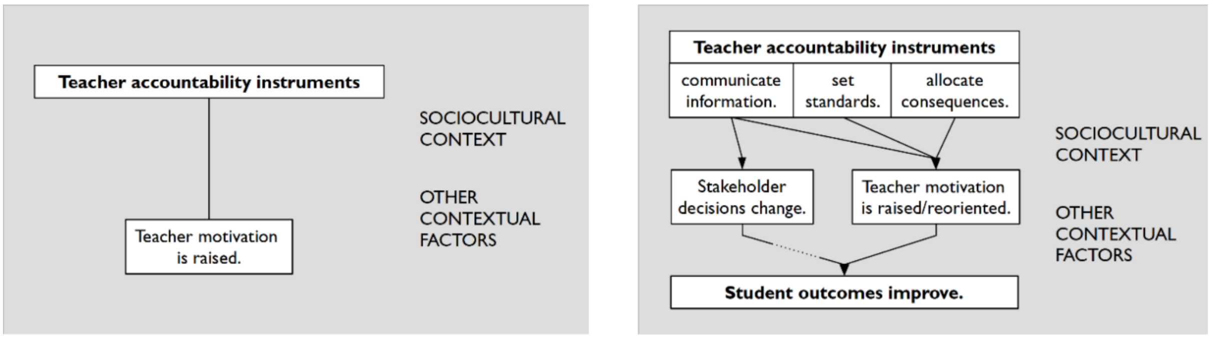
[39] Unfortunately, the recent English-language sources that I could find on Finnish culture were mainly journalistic accounts written by advocates for certain Nordic principles and practices (e.g. Booth, 2014 a British journalist who lives in Denmark; and Partanen, 2016 a Finnish journalist who has naturalised in the United States). Nonetheless, despite their evident admiration for Finnish/Nordic culture, these accounts were also critical of some aspects of this culture. Other non-scholarly analyses that I cite on Finnish education include Crehan (2016), Ripley (2013), and Walker (2017), none of whom are Finns, but all of whom have firsthand experience of Finnish schools.

somewhat idealised or otherwise divergent from participants' behaviour, then these mental articulations are nonetheless as real as manifest behaviour—and, moreover, have a real likelihood of shaping participants' responses to accountability instruments.

6.1 TALIS 2013 analysis: Teacher accountability, teacher motivation, and moderating variables

With the cross-country surveys, I use model 3 to explore RQ3 from a statistical standpoint. As outlined above in Section 3.3 and indicated in Figure 6.1 below, model 3 is a multilevel moderation model. It tests the extent to which the relationship between teacher accountability instruments (dependent variable) and teacher motivation (independent variable) is moderated by various contextual factors, including some sociocultural constructs. The structure of model 3 is similar to that of model 1, which I used in Section 4.3 to examine the extent to which sociocultural context moderates the relationship between teacher accountability and student outcomes—except that the outcome variable in model 3 is instead teacher motivation. In this model 3 analysis, the main dataset is TALIS 2013, which has self-reported data on teacher job satisfaction data matched with school-level data on teacher accountability instruments. (For a description of the teacher job satisfaction scale, see Section 5.1. For the construction of the teacher accountability scale, see Section 4.1.)

Figure 6.1 *Relationship between Section 6.1 and the overall conceptual framework*


Teacher accountability instruments
Teacher motivation is raised.
SOCIOCULTURAL CONTEXT
OTHER CONTEXTUAL FACTORS
Teacher accountability instruments
communicate information.
set standards.
allocate consequences.
Stakeholder decisions change.
Teacher motivation is raised/reoriented.
Student outcomes improve.
SOCIOCULTURAL CONTEXT
OTHER CONTEXTUAL FACTORS

This section

Overall conceptual framework

As with the cross-country analysis using TIMSS 2015 data in Chapter 5, the TALIS 2013 sample has a relatively small country sample size (n=29). Given the number of country-level variables in the model, there is a risk of overfitting. (For a discussion of the multicollinearity that may result from such overfitting, examined primarily with TIMSS 2015 data but also with reference to this

TALIS 2013 dataset, see Appendix A.) To mitigate this risk, I present estimates from six separate regressions, each of which had one of the sociocultural constructs.

The main results for these model 3 regressions are shown in Table 6.1. Column (a) shows the distribution of variance across each level of the model, column (b) shows parameter estimates for all variables except the sociocultural terms, and columns (c)–(h) give results for models with each of the six sociocultural constructs and their associated interaction terms. Note that the teacher job satisfaction outcome variable has been standardised such that its mean=0 and SD=1 in the full TALIS 2013 sample.

As shown in Table 6.1, these regressions indicate that more extensive teacher accountability is associated with lower levels of teacher motivation, as shown in the negative parameter estimates on the country-level teacher accountability term. These estimates are sizable, ranging from -0.168, for the confidence in institutions model in column (c), to -0.263, for the civic norms model in column (e). (For reference, the teacher job satisfaction variable has been standardised, as noted above; and the country-level $\overline{\text{Accountability}}$ weighted mean had a standard deviation of 0.58, as shown in Table 3.4.) The coefficient on the school-level teacher accountability differential is also consistently negative, though smaller, at approximately -0.06 across columns (b) to (h). However, these estimates for the $\overline{\text{Accountability}}_c$ and $(\text{Accountability}–\overline{\text{Accountability}})_{sc}$ terms were significant at the 5-percent level in fewer than half of these regressions. In short, higher levels of the TALIS teacher accountability scale are associated with lower levels of teacher job satisfaction, but this association is not consistently significant.

Additionally, there is some limited evidence that certain aspects of sociocultural context moderate the relationship between teacher accountability instruments and teacher motivation. This relationship is significantly moderated by confidence in institutions and social trust, as seen in the significant interactions between $\overline{\text{Accountability}}$ and the respective sociocultural constructs in columns (c) and (f). Given that the scales for confidence in institutions and social trust were derived from factor analysis and thus have standard deviations approximately equal to 1, it is notable that the coefficients on their interaction terms are of a similar order to the coefficients on the main $\overline{\text{Accountability}}$ terms. Specifically, the $\overline{\text{Accountability}}$*confidence in institutions coefficient was -0.131, slightly but insignificantly smaller than the $\overline{\text{Accountability}}$ coefficient of -0.168 in the same regression; and the $\overline{\text{Accountability}}$*social trust coefficient was -0.167, comparable to the $\overline{\text{Accountability}}$ coefficient of -0.173 in the same regression.

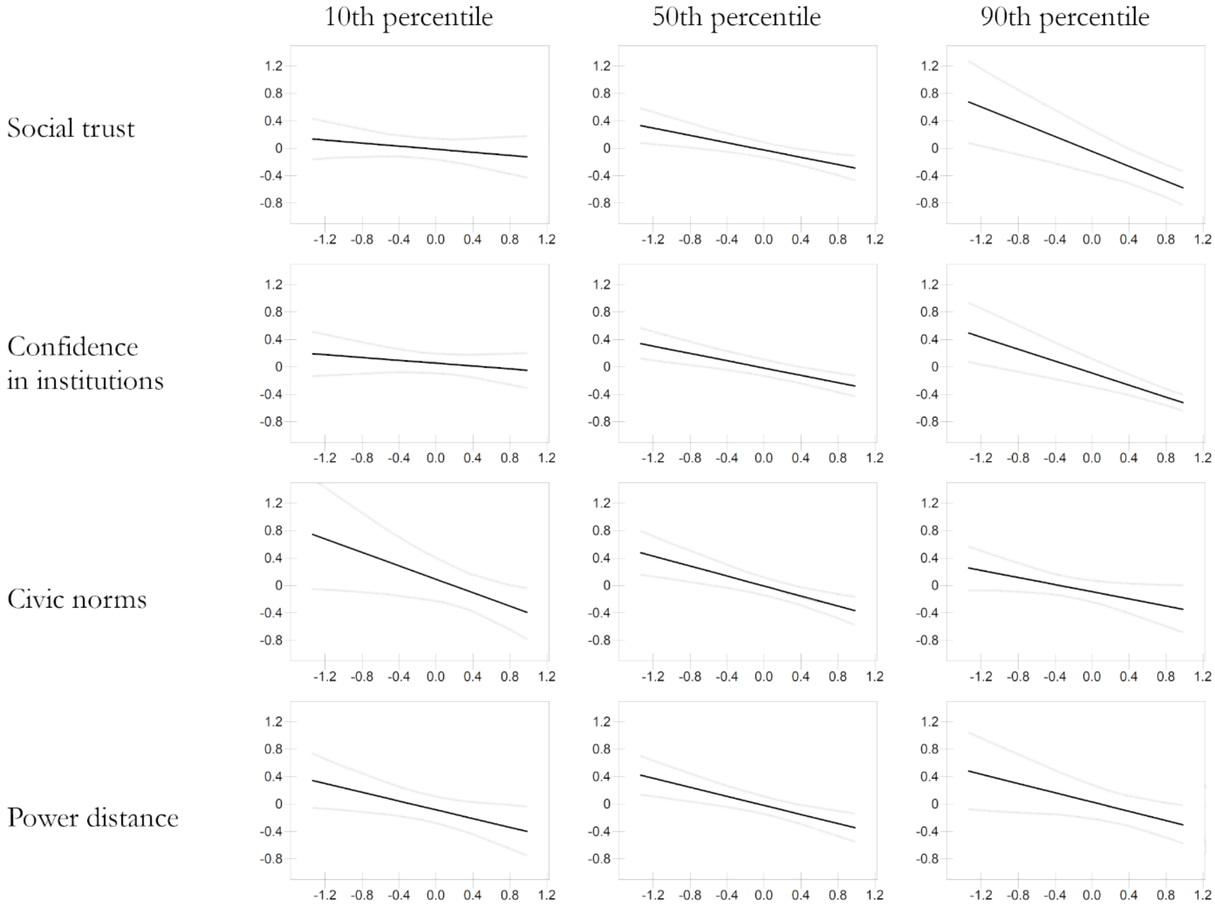
Table 6.1 *Model 3: results for multilevel regressions of TALIS 2013 teacher job satisfaction against different sets of predictors*

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
	Variance components	*No sociocultural*	*Confidence in inst'ions*	*Civic networks*	*Civic norms*	*Social trust*	*Power distance*	*Uncertainty avoidance*
Teacher level								
Constant$_{tsc}$	0.002	-0.227*	-0.214*	-0.183	-0.208*	-0.223*	-0.224*	-0.212*
	(0.057)	(0.092)	(0.092)	(0.103)	(0.098)	(0.088)	(0.095)	(0.098)
Teaching experience$_{tsc}$		0.002*	0.002*	0.002*	0.002*	0.002*	0.002*	0.002*
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
School level								
School autonomy$_{sc}$		0.271**	0.265**	0.272**	0.269**	0.270**	0.272**	0.271**
		(0.081)	(0.081)	(0.080)	(0.080)	(0.080)	(0.080)	(0.081)
(Accountability–$\overline{\text{Accountability}}$)$_{sc}$		-0.063	-0.064*	-0.061	-0.066*	-0.068*	-0.066	-0.056
		(0.034)	(0.031)	(0.033)	(0.032)	(0.031)	(0.035)	(0.034)
Country level								
$\overline{\text{Accountability}}_{c}$		-0.218*	-0.168	-0.204	-0.263*	-0.173	-0.229*	-0.204*
		(0.113)	(0.087)	(0.168)	(0.126)	(0.098)	(0.117)	(0.094)
GDP$_{c}$		-0.001	0.022	-0.037	0.016	-0.023	0.017	0.016
		(0.038)	(0.048)	(0.041)	(0.029)	(0.082)	(0.042)	(0.051)
[Sociocultural]$_{c}$			-0.057	0.199	-0.085	-0.012	0.042	0.032
			(0.05)	(0.127)	(0.097)	(0.08)	(0.069)	(0.053)
Interactions with (Accountability–$\overline{\text{Accountability}}$)$_{sc}$								
*Teaching experience$_{tsc}$		0.000	0.000	0.000	0.000	0.000	0.000	0.000
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
*School autonomy$_{sc}$		0.088*	0.091*	0.084*	0.089*	0.097*	0.099*	0.082
		(0.044)	(0.043)	(0.042)	(0.042)	(0.041)	(0.042)	(0.043)
*GDP$_{c}$		-0.011	-0.017*	-0.011	-0.005	-0.004	0.001	-0.014
		(0.010)	(0.008)	(0.011)	(0.01)	(0.017)	(0.009)	(0.01)
*[Sociocultural]$_{c}$			0.025	-0.009	-0.019	-0.017	0.037	-0.008
			(0.013)	(0.017)	(0.014)	(0.025)	(0.012)	(0.015)
Interactions with $\overline{\text{Accountability}}_{c}$								
*Teaching experience$_{tsc}$		0.004*	0.004*	0.004*	0.004*	0.004*	0.004*	0.004*
		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
*School autonomy$_{sc}$		-0.139	-0.153	-0.130	-0.145	-0.126	-0.144	-0.144
		(0.084)	(0.081)	(0.080)	(0.083)	(0.082)	(0.082)	(0.078)
*GDP$_{c}$		0.010	0.029	0.006	-0.002	0.028	-0.005	0.040
		(0.044)	(0.047)	(0.062)	(0.036)	(0.083)	(0.048)	(0.046)
*[Sociocultural]$_{c}$			-0.131*	0.040	0.106	-0.167*	-0.010	0.093
			(0.064)	(0.243)	(0.141)	(0.073)	(0.087)	(0.055)
Variance parameters								
Teacher	0.833	0.832	0.832	0.832	0.832	0.832	0.832	0.832
School	0.078	0.075	0.075	0.075	0.075	0.075	0.075	0.075
Country	0.093	0.070	0.062	0.056	0.067	0.064	0.069	0.065
-2*loglikelihood	234853.97	234661.27	234650.67	234654.52	234655.90	234656.32	234649.29	234658.52
Likelihood ratio test, compared to (b)	—	—	10.61*	6.75	5.37	4.95	11.98**	2.75
(df)			(3)	(3)	(3)	(3)	(3)	(3)
VPC (country-level)	0.093	0.072	0.064	0.058	0.069	0.066	0.071	0.067

Note. VPC=variance partition coefficient. N(teachers)=79 252; N(schools)=5 259; N(countries)=29.
*$p < 0.05$. **$p < 0.01$.

Figure 6.2 illustrates these relationships by showing predicted levels of teacher job satisfaction against $\overline{\text{Accountability}}$ at the 10th, 50th, and 90th percentiles of different sociocultural constructs, with all other variables held constant at their means. (This figure is similar to Figure 4.6 for model 1.) In addition social trust and confidence in institutions, i.e. the two sociocultural constructs that significantly interacted with $\overline{\text{Accountability}}$, I also show predicted teacher job satisfaction levels for (a) civic norms, which significantly interacted with $\overline{\text{Accountability}}$ in model 1 (Section 4.3), but did not interact significantly with $\overline{\text{Accountability}}$ in this analysis (parameter estimate=-0.106, SE=0.141); and (b) power distance, as an example of a sociocultural construct that did not discernibly interact with $\overline{\text{Accountability}}$ (parameter estimate=-0.010, SE=0.087).

Figure 6.2 *Predicted TALIS 2013 teacher job satisfaction (and 95% confidence intervals) against $\overline{\text{Accountability}}$, for the 10th, 50th, and 90th percentiles of each contextual moderator*



Note. Each row shows predicted teacher job satisfaction against $\overline{\text{Accountability}}$ for the 10th, 50th, and 90th percentile of the named contextual predictor. All other variables are held constant at their means. Rows are sorted by decreasing magnitude of the p-value on the coefficient of corresponding interaction term. Each row of predictions is based on a regression in Table 6.1: social trust from column (f) of the table, confidence in institutions from column (c), civic norms from column (e), and power distance from column (g).

The downward slopes in most of these predicted plots reflect the negative overall relationship between teacher job satisfaction and teacher accountability instruments. However, this relationship is not uniform. The negative slope between job satisfaction and accountability does not appear to vary with different levels of power distance. For civic norms, this negative association varies slightly (reflecting the nontrivial coefficient on the $\overline{\text{Accountability}}$*civic norms interaction term), but this variation is well within the very wide confidence intervals, especially at the 10th percentile of civic norms (reflecting the large standard error on the coefficient). However, the negative association between teacher job satisfaction and teacher accountability varies noticeably with social trust and confidence in institutions. At the 10th percentiles of social trust and of confidence in institutions, teacher job satisfaction does not meaningfully vary with $\overline{\text{Accountability}}$. In contrast, at the 90th percentiles of social trust and confidence in institutions, teacher job satisfaction decreases considerably as teacher accountability increases.

To make these differential effects more concrete, at the 0th percentile of teacher accountability (i.e. the leftmost ends of each plot), an average teacher in a hypothetical country with social trust at the 90th percentile would be expected to have a job satisfaction score that is 0.55 points (≈0.55 SD) higher than a similar teacher in a hypothetical country at the 10th percentile of social trust, holding all other variables constant at their means. Conversely, at the 100th percentile of teacher accountability (i.e. the leftmost ends of each plot), the difference in these two teachers' job satisfaction scores would be 0.46 points, but in favour of the teacher in the low-trust country. Nevertheless, it is important to note that the confidence intervals overlap across each of these plots at different levels of social trust, even at the extreme ends of each plot. That is, even though the moderation effects from social trust described in this paragraph may be fairly large, they are not significantly different from each other when all parameters are considered concurrently.

Thus far, the answer to RQ3 from the TALIS 2013 analysis appears to be that the relationship between teacher accountability instruments and teacher motivation may depend to a limited degree on some aspects of sociocultural context. However, there are some indications that this conditional relationship may not be robust. Besides relatively imprecise estimates (as manifested in the overlapping confidence intervals described above), there is a possibility that the coefficient on the social trust interaction term may be inflated due to multicollinearity, as indicated by the large increases in the standard errors for the GDP and $\overline{\text{Accountability}}$*GDP terms in moving from column (b) to the social trust model in (f). That said, the confidence in institutions model in column (c) does not show any similar signs of multicollinearity.

Sensitivity checks

To test the robustness of these sociocultural moderations, I re-estimated the models using the TIMSS and PISA datasets that were also used for sensitivity checks in Chapter 5. Specifically, I estimated different versions of model 3 for: the main TIMSS 2015 dataset, matched with $\overline{\text{Accountability}}$ from PISA 2015; a TIMSS 2015 dataset with matched $\overline{\text{Accountability}}$ from PISA 2012; the full PISA 2012 sample; and an OECD-only PISA 2012 subsample. Similar to TALIS 2013, the TIMSS 2015 proxy for teacher motivation was a job satisfaction scale derived from teacher-reported items. For PISA 2012, the proxy was a teacher morale scale derived from principal-reported items. As noted in Section 5.1, all of these scales were standardised (although the PISA 2012 teacher morale variable was scaled such that mean=0 and SD=1 among OECD countries, not across the wider sample).[40]

These sensitivity checks show that the TALIS 2013 results described above are far from robust across comparator datasets. Table 6.2 summarises the sensitivity checks, showing the significance and direction of the main $\overline{\text{Accountability}}$ parameter as well as the interactions between $\overline{\text{Accountability}}$ and each sociocultural construct. As with Table 6.1, each regression summarised in this table included only one sociocultural construct and its associated interaction term. Accordingly, every cell in the 'Interactions with $\overline{\text{Accountability}}$' section of table corresponds to separate regression with a distinct dataset-sociocultural construct pairing.

As shown in Table 6.2, there is no consistency across these datasets either in the relationship between teacher accountability instruments and teacher motivation, or in the sociocultural constructs that moderate this relationship between accountability instruments and motivation. As reported above, social trust and confidence in institutions significantly moderated the effects of accountability instruments in the main TALIS regressions. However, social trust was not a significant moderator in any of the TIMSS 2015 and PISA 2012 regressions. While confidence in institutions was a significant moderator for both the main TALIS 2013 data and the TIMSS 2015

[40] I considered running sensitivity checks with the 2008 cycle of TALIS, but decided that this would be quite resource-intensive without a proportionate gain in analytical insight. While TALIS 2008 included 9 out of the 14 questionnaire items that I used in constructing the TALIS 2013 teacher accountability scale, the response categories for 6 out of those 9 items in 2008 were binary, unlike their ordinal 2013 counterparts. Hence the TALIS 2008 dataset would have required a separate scale construction procedure for teacher accountability instruments. Additionally, the TALIS 2008 data predated the other educational datasets by several years, and thus would have required the construction of social capital scales from earlier WVS/EVS cycles. Given that the lack of robustness in the model 3 results was already apparent in comparisons of the TALIS 2013, TIMSS 2015, and PISA 2012 data, I deemed it unnecessary to conduct additional sensitivity checks using the TALIS 2008 data.

dataset that was matched to Accountability from PISA 2015, these moderation effects ran in opposite directions.

Table 6.2 *Summary of the direction and significance of the main effect and moderated effects of Accountability on teacher motivation*

	TALIS 2013	TIMSS 2015		PISA 2012	
	Full sample	*Accountability* from PISA 2015	*Accountability* from PISA 2012	Full sample	OECD countries
Main effect of Accountability across the 6 models					
Direction	–	Varies	+	Mostly +	Mostly –
Significance	Varies	Not significant	Not significant	Not significant	Not significant
Interactions with Accountability in each of the 6 models					
*Confidence in institutions	–		+		
*Civic networks					
*Civic norms		+ +			
*Social trust	–				
*Power distance				+ +	+
*Uncertainty avoidance					
N (countries)	29	23	22	52	36

Note. The teacher motivation variable for TALIS 2013 and TIMSS 2015 is teacher job satisfaction (teacher-reported scale); and for PISA 2012, teacher morale (principal-reported scale). For interactions with Accountability, 1 symbol (+/–) indicates p<.05; 2 symbols (+ +/– –) indicate p<.01. Full results can be provided upon request.

For a finer-grained view, Table 6.3 shows parameter estimates for selected predictors across the TALIS, TIMSS, and PISA regressions, for the sociocultural moderators that were significant in the main TALIS regressions above. The first three columns replicate columns (b), (c), and (f) from Table 6.1. The other columns present analogous results for TIMSS 2015 with Accountability from PISA 2015, and for the full PISA 2012 sample, respectively. Each column (b) shows a regression without any sociocultural terms; column (c) shows the model with confidence in institutions terms, and column (f) shows the model with social trust terms.

As shown in Table 6.3, the effect of the main (i.e. non-interacted) accountability term is inconsistent across the datasets, with relatively large negative coefficients in TALIS, but much smaller effects in TIMSS and PISA. Moreover, this accountability term is only significant in the TALIS model without any sociocultural variables.

More pertinent for RQ3 is the fact that the interaction between teacher accountability and social trust has a sizable and negative coefficient across all three datasets, although it is only significant for TALIS. This Accountability*social trust coefficient is also large and negative, though insignificant, in the TIMSS 2015 dataset matched with accountability data from PISA 2012 (which is not shown in Table 6.3, but results are available upon request). However, in the PISA 2012 dataset with only OECD countries (also not shown in Table 6.3), the coefficient of the $\overline{\text{Accountability}}$*social trust term is, instead, positive, though insignificant. Such inconsistency was also present in the confidence in institutions regressions. While the $\overline{\text{Accountability}}$*confidence term was significant in the TALIS regression, it was smaller and insignificant for TIMSS and PISA—especially for PISA, where the coefficient was negligible.

Table 6.3 *Model 3: sensitivity checks for confidence in institutions and social trust (showing selected variables only)*

	TALIS 2013			TIMSS 2015			PISA 2012		
Predictors (selected)	(b1)	(c1) Conf. in institut'n	(f1) Social trust	(b2)	(c2) Conf. in institut'n	(f2) Social trust	(b3)	(c3) Conf. in institut'n	(f3) Social trust
Accountability	-0.218* (0.113)	-0.168 (0.087)	-0.173 (0.098)	0.023 (0.137)	-0.013 (0.135)	0.074 (0.130)	0.050 (0.104)	0.040 (0.098)	0.035 (0.119)
$\overline{\text{Accountability}}$ *Teaching experience	0.004* (0.002)	0.004* (0.002)	0.004* (0.002)	0.007 (0.006)	0.007 (0.006)	0.007 (0.006)			
$\overline{\text{Accountability}}$ *[School context]+	-0.139 (0.084)	-0.153 (0.081)	-0.126 (0.082)	0.008 (0.052)	0.008 (0.052)	0.009 (0.053)	-0.174** (0.037)	-0.173** (0.038)	-0.171** (0.038)
$\overline{\text{Accountability}}$ *GDP	0.010 (0.044)	0.029 (0.047)	0.028 (0.083)	-0.110 (0.077)	-0.067 (0.111)	0.021 (0.188)	-0.037 (0.059)	-0.021 (0.073)	-0.016 (0.051)
[Sociocultural]		-0.057 (0.05)	-0.012 (0.08)		0.07 (0.069)	0.070 0.142		0.027 (0.051)	0.098* 0.049
$\overline{\text{Accountability}}$ *[Sociocultural]		-0.131* (0.064)	-0.167* (0.073)		-0.090 (0.197)	-0.451 (0.285)		-0.009 (0.093)	-0.117 (0.193)
N Teachers		79 252			6 147			—	
Schools		5 259			3 761			14 840	
Countries		29			23			52	

Note. The teacher motivation variable for TALIS 2013 and TIMSS 2015 is teacher job satisfaction (teacher-reported scale); and for PISA 2012, teacher morale (principal-reported scale). The TIMSS 2015 data here are matched with country-level weighted means for teacher accountability instruments from PISA 2015. The PISA 2012 dataset is the full sample. Full results can be provided upon request.
+For TALIS 2013 and PISA 2012, school context=school autonomy. For TIMSS 2015, school context=school resources.
*p < 0.05. **p < 0.01.

It is also worth noting that likelihood ratio tests for the TIMSS 2015 and PISA 2012 data that compared the respective models with confidence in institutions and social trust to the model without any sociocultural predictors—i.e. comparing columns (c2) and (f2), in turn, to column (b2) for TIMSS, and corresponding comparisons for PISA—did not find significant improvements in model fit from adding the main and interaction terms for either confidence in institutions or social trust. For the TALIS 2013 models, the addition of the confidence in institutions terms improved model fit slightly but significantly. However, the change in model fit for social trust was insignificant, as shown in the likelihood ratio test in Table 6.1.

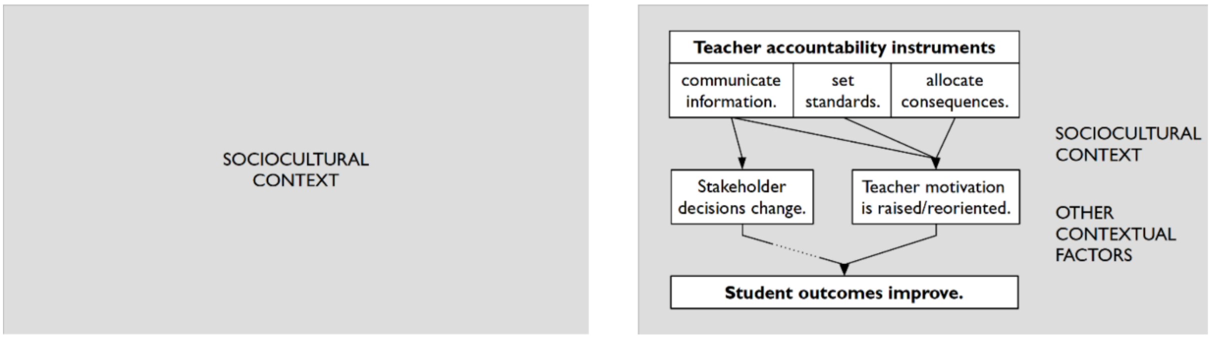
The lack of consistency between the TALIS and PISA results may be understandable, since their respective proxies for teacher motivation captured different constructs (teacher job satisfaction vs. teacher morale) and were derived from different categories of survey respondents (teachers vs. principals). However, the TALIS and TIMSS analyses both used teacher-reported job satisfaction to proxy for teacher motivation. Although these scales were derived from different questionnaire items, as described in Section 5.1, they are conceptually similar. Moreover, the correlation between the country-level weighted means of these TALIS and TIMSS teacher job satisfaction scales was moderately strong (r=0.64 for 15 countries, as shown in Table 5.2). Hence, the lack of consistency between the TALIS and TIMSS analyses does cast doubt on their overall robustness. (While the TALIS and PISA teacher accountability scales were also constructed in separate IRT analyses using different questionnaire items, the country-level correlations between these scales were strong, as reported in Section 4.1. Specifically, r=0.86 for the 35 countries with teacher accountability data for both PISA 2012 and TALIS 2013, and r=0.77 for the 34 countries with data for both PISA 2015 and TALIS 2013. Hence, inconsistency of the teacher motivation results is unlikely to stem primarily from the different teacher accountability measures. It would be theoretically possible to test whether the TIMSS 2015 results are robust to the substitution of the country-level teacher accountability measure from PISA with the TALIS 2013 accountability measure. However, this is not feasible in practice, since there is an overlap of only 11 countries between the TIMSS 2015 and TALIS 2013 datasets, giving too small a sample size at the country-level for the multilevel modelling.)

It is difficult to disentangle the degree to which these inconsistencies are due to measurement issues, different teacher motivation proxies, sample design, or real-world variegation. However, as they stand, these datasets offer little clear evidence that national sociocultural context affects the relationship between teacher accountability and teacher motivation.

6.2 Sociocultural context in Finland and Singapore

In contrast to the cross-country surveys, the field interviews offered convincing evidence that sociocultural context influences the relationship between teacher accountability and teacher motivation, at least for the 24 teachers whom I interviewed. I drew on these interviews alongside secondary sources in Chapter 5 to outline Finland's and Singapore's approaches to teacher accountability, and how these approaches can affect teacher motivation. I now return to these interviews to trace how sociocultural context affects this relationship between accountability instruments and teacher motivation. I begin in this section by identifying sociocultural patterns in each country that are salient to the workings of teacher accountability instruments, as indicated in Figure 6.3.

Figure 6.3 *Relationship between Section 6.2 and the overall conceptual framework*


SOCIOCULTURAL CONTEXT
Teacher accountability instruments
communicate information.
set standards.
allocate consequences.
Stakeholder decisions change.
Teacher motivation is raised/reoriented.
Student outcomes improve.
SOCIOCULTURAL CONTEXT
OTHER CONTEXTUAL FACTORS

This section

Overall conceptual framework

To frame this analysis, I would like to emphasise, firstly, that Finland's and Singapore's sociocultural contexts—or, indeed, any national sociocultural context—are too complex to be classified along a set of unidimensional traits. Unidimensional traits facilitate the cross-country statistical analyses in this thesis. However, such traits are also reductive to the point of obscuring certain contextual characteristics that should inform policy design. To illustrate, one research programme in cultural psychology identifies different degrees of independence and interdependence in cultures (Markus & Conner, 2013; Markus & Kitayama, 1991). In this framework, people in independent cultures tend to view themselves as individual, autonomous, and equal; whereas those in interdependent cultures tend to view themselves as relational and guided by tradition and hierarchy. In some senses, Finland's culture is highly independent, and Singapore's is highly interdependent. This is evident in Finnish participant Maarit's assertion that 'our headmaster is, really, one of us' (echoed by Anneli, Liisa, and Emilia) and a Singaporean participant's recollection of their colleagues in an MOE department being 'mortally afraid' of visits from their top-level superior because 'there was this belief that someone in that position needed to be treated in certain ways'. However, Timothy (Singapore) also asserted that 'no matter what, Singaporeans are very me-first'—which other Singaporean participants agreed with, and which is hardly a hallmark of interdependence. Also, as I discuss below, Finns are highly attuned to social responsibility, thus departing from extreme independence in ways that are interlinked with accountability.

The fact that Finland and Singapore each combine independent and interdependent traits does not contradict Markus, Conner, and Kitayama (ibid), who argue that every individual and every cultural setting combine different degrees of independence and interdependence. That said, the sociocultural patterns influencing teacher accountability in Finland and Singapore include some characteristics that do not fit neatly under either independence or interdependence, nor under any one of the social capital or Hofstede scales that I used in the statistical analysis (see Figure 3.2 for box plots showing Finland's and Singapore's scores on these scales). In this section, I sketch a portrait of the salient features of each country's sociocultural context, based on interview participants' accounts. One component of each portrait is a discussion of the degree to which participants agreed with key observations from the WVS/EVS and Hofstede datasets. While participants broadly endorsed these statistical aggregates, they also described more nuanced patterns that the standardised surveys were not designed to capture.

Note, however, that I asked participants about the WVS/EVS and Hofstede data at the end of each interview, as shown in the interview guide (see Table 3.6). Besides participants' view on these survey-based measures, this section draws heavily on their less-structured observations on sociocultural context, whether tangential remarks along the way or specific responses to pertinent questions, especially #8 in Table 3.6, i.e. 'Can you tell me about aspects of [Singaporean/Finnish] culture that affect [Singapore's/Finland's] education system?' In line with Tomlinson's (1989) argument for 'hierarchical focusing', I asked the more general questions such as #8 before posing the specific questions based on the WVS/EVS and Hofstede data, in order to minimise undue influence on participants' articulated views. (See Section 3.4 for more details.)

Secondly, although there are many contrasts between Finland's and Singapore's sociocultural contexts, it would be inaccurate to regard them as polar opposites. This is partly because culture is complex and multidimensional, as noted above (see also Section 2.2; and Anderson-Levitt, 2012). Furthermore, there are important similarities between these two contexts. For example, Singaporeans are commonly (and not wholly inaccurately, as I discuss below) regarded as deferential to authority. However, Finns are also 'usually quite obedient', in Emilia's words (echoed by many Finnish participants; see also Simola et al., 2017). Another relevant similarity is that people in both contexts value education highly, whether for nation building or for individual mobility (see Barr & Skrbiš, 2008; Dimmock & Tan, 2016; Jones, 2019; Ng, 2013 on Singapore; and Chung, 2009; Klinge, 1993; Simola et al., 2017 on Finland). This was noted by both American interview participants. Mark, who taught in Singapore, argued that one reason why

Singapore, Finland, and Shanghai are educationally successful is because 'there's a societal expectation that students will do well in school', and this 'can't be underestimated'. Similarly, Masa commented that Finland's traditionally literary culture has 'a huge effect that [he doesn't] think has been measured, or even appreciated' on educational outcomes.

Similarities notwithstanding, there are clear differences between Finland's and Singapore's respective sociocultural contexts, as I show below. For convenience, I sum up the relevant sociocultural patterns in Finland as 'complementary responsibility', and those in Singapore as 'managed meritocracy'.

Complementary responsibility in Finland

Table 6.4 summarises Finnish participants' responses to the penultimate set of questions in every interview, in which I asked whether the aggregate WVS/EVS and Hofstede data in my statistical analysis accurately represented their sociocultural context. In each country, I focused on three sociocultural constructs for which their country was in either the highest or lowest quartile. For Finland, these were: high social trust, low power distance, and high adherence to civic norms.

Table 6.4 *Interview participants' level of agreement with indicators about Finland from the sociocultural surveys*

● agree ◑ partially agree ○ disagree

	FINLAND											
	Anneli	Liisa	Emilia	Kristiina	Masa	Satu	Maarit	Hannele	Helena	Antero	Päivi	Juhani
High social trust	●	●	●	●	●	●	●	●	●	●	●	●
Low power distance	●	●	●	●	●	●	●	●	●	●	◑	◑
Not justifiable to breach civic norms*	●	●	●	●	◑	●	●	●	●	●	●	●

*This question was phrased as: 'Would you agree that most Finns think that you should follow the official system rather than trying to find loopholes or shortcuts for your own benefit?'

As shown in Table 6.4, almost all Finnish participants endorsed these sociocultural characteristics. They agreed unanimously about high social trust, with most participants giving unqualified agreement. In Antero's words:

> People can leave their doors open. Or, for example, I leave the keys in my car. […] If I have an emergency situation, I can give my wallet, my keys, my phone, my kids to a stranger who will help me.

While acknowledging that 'there are some people who will steal them', Antero stated that 'the common view is that we can trust'—a view that every participant affirmed. To illustrate this using data from the European Values Study 4, when given a choice between the statements 'Most people can be trusted' and 'You can't be too careful in dealing with people', 65.1% of Finnish respondents chose the former. The proportion of trust-oriented respondents in Finland was exceeded only by Nordic counterparts Denmark (76.1%), Norway (74.2%), and Sweden (70.1%). For comparison, the lowest proportion of trust-oriented respondents was 4.7%, in the Turkish Cypriot Community of Northern Cyprus (EVS, 2016b). Separately, an analysis of the United States' General Social Survey found that having ancestors who immigrated to the United States from Nordic countries was associated with being more likely to say that most people can be trusted (Uslaner, 2008). Admittedly, there are limits to how informative such self-reported data are. For corroboration, although Finland was not included in Cohn et al. (2019) civic honesty experiment with 17,000 'lost' wallets in 40 countries, Nordic neighbours Norway, Denmark, and Sweden ranked among the top five places where a 'lost' wallet was most likely to be returned, whether or not it contained any money.

When asked about the low power distance measured in the Hofstede survey—i.e. a preference for equal power distributions rather than hierarchies—most participants strongly agreed. This sociocultural egalitarianism is mirrored in Finland's relative economic equality. During the ten most recent years for which Gini data are available (2005–2015), Finland consistently had one of the ten most equal income distributions in the world (World Bank, 2019). Finland also had the third highest score in Oxfam's 2018 Commitment to Reducing Inequality index (Lawson & Martin, 2018). Klinge (1984, 1993) suggests that this egalitarianism is longstanding, due in part to the absence of a large landholding aristocracy. However, although Liisa said that 'equality is a Finnish thing', she also noted that it could cause inefficiencies: since all teachers 'are on the same level', this leads to the 'big problem' that 'one principal has a lot of people to manage'. More controversially, Juhani argued that the stated preference for egalitarianism may not reflect actual practice:

> I think that's what they *want*. But that's not how they act. We want equality, yes. And power should be divided equally. But every time there is a problem anywhere, we need that hierarchy, and a person who is in a high enough position to say something. […] I think we still have not gotten rid of the centuries of Swedish and Russian bureaucracy.

Likewise, Päivi observed that Finns say they want horizontal power distributions 'as much as possible', but that in practice they 'look up': 'We want to have, "*You* are doing wrong. This is not our fault. It's *yours*." (laughter) We want to have someone to blame.'[41]

This tension between egalitarianism and a preference for orderly delegation was also apparent in participants' views on Finnish adherence to civic norms. Most participants agreed with Emilia, in that 'it's very ingrained into Finnish culture to just follow rules, and to do it by the book'. However, Anneli gave a more qualified view, saying, 'We trust the system, *as long as it feels somewhat fair, and reasonable*—which is usually true, but not always [emphasis added]'. American-born Masa's reservations went further: he argued that Finland has 'creative flexibility in rules', perhaps from its history of conforming superficially to appease 'old Soviet influences'. Examples that Masa cited include: when a suitable candidate has been identified for a job vacancy before the vacancy has been advertised, school leaders may display the vacancy advertisement behind a plant or another obstruction in the teachers' room in order to superficially fulfil the requirement for job vacancies to be displayed publicly; and when families visit their remote lakeside cabins in the summer and need to dispose of their rubbish, they save it for bonfires during the traditional Juhannus (midsummer) celebration rather than building illegal fires solely for rubbish disposal. However, he emphasised that the underlying intention is to streamline the system rather than to abuse it, and that 'it doesn't seem to be really hurting anybody':

> There's nothing really wrong [with this flexibility about norms], but it's a common-sense approach […] And yes, there is a rule, we can follow that rule, but we can still do pretty much whatever we want. And to me, that feels like a very Finnish thing.

Masa's observations about creative compliance were not simply an outsider's contrarian view. Rather, they align with other participants' descriptions of how some teachers perceive the national curriculum, as I discuss in the next section.

Crucially, these sociocultural patterns interact closely with each other. When I asked Helena whether Finns trust others to behave fairly, she said:

> Yes, I think that's the first thing. If someone asks me what's special about the Finnish school system, I think the core thing is trust. […] I've met hundreds of visitors from different countries. […] And the one thing which they're always equally amazed about

[41] It is also important to note that Finland is not a utopia without any socioeconomic differentiation. Helena noted that intergenerational inequalities do affect education: 'We have research information that students whose parents have academic [rather than vocational] schooling do better academically. […] So it would be lying to say that doesn't exist.' Additionally, Hanifi's (2013) analysis of Statistics Finland's time use survey found another unequal distribution: people of higher socioeconomic status were more likely to participate in volunteer work and to agree that most people can be trusted, indicating association between socioeconomic privilege and social capital.

> is that I can talk to my boss as an equal. […] But that's also to do with the trust, I think. That you trust each other to do the job.

When asked whether most Finns believe in following the official system rather than seeking loopholes or shortcuts, she said:

> I think that boils down to trust again. You're more willing to follow the system if you trust that there's a good reason for that system to exist. And if you feel that police are trustworthy, that they're good people who are trying to do a job that benefits the whole society, then you will do what the police ask you to do.

Other participants similarly believed that these sociocultural patterns were mutually reinforcing. Like Helena, Anneli linked civic norms adherence to trust: 'I do trust the system, and I believe that somebody *smarter* than me, or more acquainted with the system, has decided that this is the system that works best.' In turn, Kristiina linked norms adherence to egalitarianism: 'If everybody follows the rules, then we are following the same rules, and we're treated equally as well.' This interaction between trust, egalitarianism, and civic responsibility is also stressed by Berggren and Trägårdh (2010) as well as Partanen (2016) in their discussions of Nordic society and culture. In their view, Nordic social responsibility enables a strong individual autonomy, because the trust ethic alongside egalitarian welfare policies serve to liberate individuals from dependency, such that children need not depend on family finances for their education, and employers are free to fire undesirable employees without the latter losing basic livelihood necessities or healthcare coverage.

Finland's blend of trust, egalitarianism, and civic mindedness constitutes a sociocultural context that I dub 'complementary responsibility'. This sense of responsibility is far-reaching, and participants viewed it as being productive. Masa recalled a German colleague who believed that Finland's PISA success derived not from its school system, but from its culture—'the idea that we're going to try and take care of everybody, […] responsibility on a societal level'. I characterise such responsibility as 'complementary' rather than 'mutual' or 'reciprocal' because it is not premised on transactional tit-for-tat. Neither is there the expectation that everyone in a particular setting should jointly pay attention to the same obligations. Rather, there is a belief that (almost) everyone, in their respective stations, will autonomously act for the common good. In Juhani's words:

> Finnish society is stable. And it has never actually listened to others. We are stubborn. And we like to do things in our own way. And we tend to keep our own way within us. We don't talk about it to others. And we expect everybody knows their roles, because, 'Come on, we are Finns, hey.'

This idea of everyone having respective, complementary roles aligns with Juhani's and Päivi's previous comments about the desire for clear leadership despite a preference for egalitarianism. The notion that, within these distinct roles, people may 'do things in [their] own way' aligns with Masa's observations about 'creative flexibility' in fulfilling shared norms.

Remarkably, the expectation that 'everybody knows their roles' extends to children. The interview participant who currently works in international education contrasted their stint overseas where 'you need to drive your children everywhere' because 'they cannot go anywhere by themselves', with Finland, where primary school children commute by public transport independently. According to this participant, 'Children need to take responsibility, but we also need to trust them. And we need to trust society that if something goes wrong, somebody would help.' Likewise, Masa remarked on the high degree of autonomy accorded to Finnish schoolchildren (see also Partanen, 2016; Ripley, 2013; T. D. Walker, 2017, Chapter 3). As I show in Section 6.3, this ethos of complementary responsibility shapes teachers' responses to accountability instruments.

Managed meritocracy in Singapore

Turning to Singapore, Table 6.5 summarises the degree to which participants agreed with WVS/EVS and Hofstede survey findings that Singaporeans, on average, are highly confident in public institutions, relatively willing to accept power hierarchies, and fairly tolerant of breaches in civic norms. As in Finland, participants broadly agreed with these descriptions of Singapore's sociocultural context, although some challenged the notion that breaches in norms are justifiable.

Table 6.5 *Interview participants' level of agreement with indicators about Singapore from the sociocultural surveys*

● agree ◑ partially agree ○ disagree

	SINGAPORE										
	Adeline	Jeffrey	Maggie	Joseph	Peter	Sonia	Timothy	Mark	Eleanor	Geok Ling	Jane
High confidence in public institutions	●	●	●	●	●	●	◑	●	●	◑	●
High power distance	●	●	●	●	◑	●	●	●	●	●	●
Justifiable to breach civic norms*	◑	●	○	●	●	○	●	◑	●	●	●

Note. Singaporean participant Andy does not appear in this table because he was the first participant to be interviewed, and I added these questions to the interview guide subsequently.
*This question was phrased as: 'Would you agree that most Singaporeans think it's justifiable to go around the official system for your own benefit if you can get away with it?'

As shown in Table 6.5, all participants agreed that most Singaporeans were highly confident in public institutions. However, Geok Ling and Timothy noted that this confidence may be eroding gradually, which Geok Ling attributed to higher levels of education and which Timothy attributed to a younger generation that tended to be demanding rather than 'appreciative'. Another important nuance to this confidence in institutions came from Eleanor's response: 'Generally, high trust and high expectations. So when expectations are not met, then they start getting upset.' Thus, this confidence is not blind faith. Rather, it is contingent on the efficacy of public institutions. According to Jane, even those Singaporeans who in principle would prefer more political opposition 'don't want to rock the boat' because they are happy that, 'all the while, the country has been doing well'.

Similarly, even though Joseph, Eleanor, and Peter noted that Singapore's power distance is decreasing over time (see also Chua, 2018), only Peter believed that this acceptance of hierarchy was no longer dominant. A more typical response came from Timothy:

> To be respectful of your superiors is definitely a given. But I find there's a tendency towards being overly deferential to them, and almost like they are—maybe not God, but like a king, and they are all-powerful.

Still, this deference may be outward rather than deeply felt. For example, Adeline said that subordinates under 'an incompetent boss' would 'grumble with people at the same level' but would 'still comply with the boss' orders despite complaining behind their back'. Jeffrey, Jane, and Geok Ling echoed her sentiments.

In contrast to the general agreement with the statistics on power distance and confidence in institutions, two Singaporean participants disagreed outright with the WVS/EVS finding that Singaporeans are relatively likely to justify breaches in civic norms for personal benefit. Maggie and Sonia disputed this finding, with the latter remarking that:

> I think even if anyone were to *try* [to circumvent the official system], our government is powerful enough to track them down. […] Plus, people here have been culturally conditioned to just obey. Yeah. 'Obey blindly, don't question, sit in your corner and shut up.'

Among those who agreed that Singaporeans were relatively tolerant of such breaches in civic norms, some participants related personal encounters with Singaporeans who exploit loopholes or shortcuts out of self-interest. For example, Geok Ling mentioned that some parents will falsify their home addresses (perhaps by temporarily renting a second property, or by using a family member's address) in order to help their children gain admission to prestigious schools.

(For a documented instance of this practice, see Chong [2018].) According to Geok Ling, one reason for such behaviour is that Singapore is 'so competitive, and resources are so scarce'. Additionally, a few male participants, like Jeffrey, laughingly recounted 'the secret eighth core value of the army' (which they had encountered during their compulsory two-year National Service stints for Singaporean men): 'You can do anything, just don't get caught.' That said, whether participants agreed or disagreed with the WVS/EVS finding, they emphasised the importance of staying strictly within legal boundaries. Mark believed Singaporeans would do '*any*thing up to—or not even stated by—the law to get an added benefit', but they would not transgress legal limits.

As in Finland, these sociocultural patterns all interact. When asked about the civic norms finding, Eleanor first noted that there are 'so many rules' in Singapore, and that the rigidity of the system can make it 'hard to get things done', before saying:

> I do remember a conversation a long time ago about whether we can or cannot U-turn on the roads, and [former prime minister] Goh Chok Tong said, 'If there is no sign that tells you you cannot U-turn, who's to say you can't?' And that's been something that I tell myself when I'm on the roads, or when I'm at work as well: if there's no specific policy that tells you that you can't, why don't you just try it? […] As long as it's not unethical. […] Singaporeans are always looking for loopholes, because it allows them a certain bit of freedom, a certain thrill that they're able to escape the system.

Thus, Eleanor is experimenting with actions that produce results more efficiently but may contravene tacit norms, although she intends to stay strictly within explicit institutional boundaries—and her basis for this experimentation is a remark from an authoritarian politician.[42]

This combination of hierarchy, confidence in institutions, and a tolerance for strategies that yield a productive advantage is what I call Singapore's 'managed meritocracy'. This view that society is—and should be—competitive and progress-oriented, with a high degree of structure and stratification to ensure maximum productivity, has been actively fostered by Singapore's single-party government throughout the second half of the twentieth century (Barr & Skrbiš, 2008; K. Y. Lee, 2000; K. P. Tan, 2010; Tremewan, 1994; see also Y. Y. Teo, 2018). Singapore's brand of meritocracy not only justifies the government's emphasis on competitiveness, but also helps to reinforce government legitimacy more broadly (Chua, 2018). The government's successful track

[42] I could not find a primary source for the Goh Chok Tong quote that Eleanor cited. However, 'No U-Turn Syndrome', or 'NUTS', is certainly discussed in Singapore as a detrimental local phenomenon. Coined by Singaporean technology entrepreneur Sim Wong Hoo (2003, excerpt accessed at https://singaporeelection.blogspot.com/2006/06/no-u-turn-syndrome-nuts.html on 18 July 2019; see also Seah, 2006), NUTS signifies excessive caution in only taking actions that are expressly permitted by a higher authority, at the expense of autonomous creativity.

record in economic development reinforces the view that it deserves to remain in power, as implied in Jane's comment about not rocking the boat.[43]

Unlike pro-meritocracy arguments that are wedded to libertarian ideals, Singapore's meritocracy is centrally managed (Chua, 2018). The chief arbiter of merit is not the market, but rather the government. Ee (2018) further argues that most Singaporeans accept the government as the arbiter of moral questions, such as abortion, homosexuality, and the death penalty. To illustrate the degree of trust placed in the government, here is the beginning of interview participant Timothy's response when asked whether Singapore's teacher accountability instruments make the education system more effective:

> To use a very civil service phrase, these [accountability-related] expectations are all cascaded down from your ministers and the top policymakers. So what they would have in mind, I believe, is in line with what, hopefully, will help make the country a better one.

This statement is all the more noteworthy since Timothy also expressed considerable doubt about certain government policies. Mark, the American who taught in Singapore, observed:

> Singapore's education system is a model of how the government wants to structure its citizens' time. […] Singapore's government seems to view free time for citizens—where they are thinking on their own and not guided or given some push in a supposed right direction—as a bad thing. […] Students' time is *very* structured. […] That is very different, obviously, from Western society and a lot of other societies where there's a belief that it's in those moments of free time that genius happens.

Mark's observations cohere with Jones' (2019) observations on long school days and scheduled home routines for Singaporean primary school students, and also with Li's (2012) analysis of Confucian (i.e. East Asian) and Western models of learning. Li argues that the Western educational tradition emphasises individual curiosity and mental ability, whereas the Confucian tradition entails a guided, diligent pursuit of self-perfection.

In Singapore, this Confucian ethic has been fused with a hard-nosed competitiveness—colloquially called *kiasu*, a Chinese Hokkien term meaning 'afraid of losing'. As Andy relates, *kiasu* culture centres on 'this idea that if you are not the best at what you do, then you can't actually […] succeed in life'. The Oxford English Dictionary (2006), in turn, defines being *kiasu*

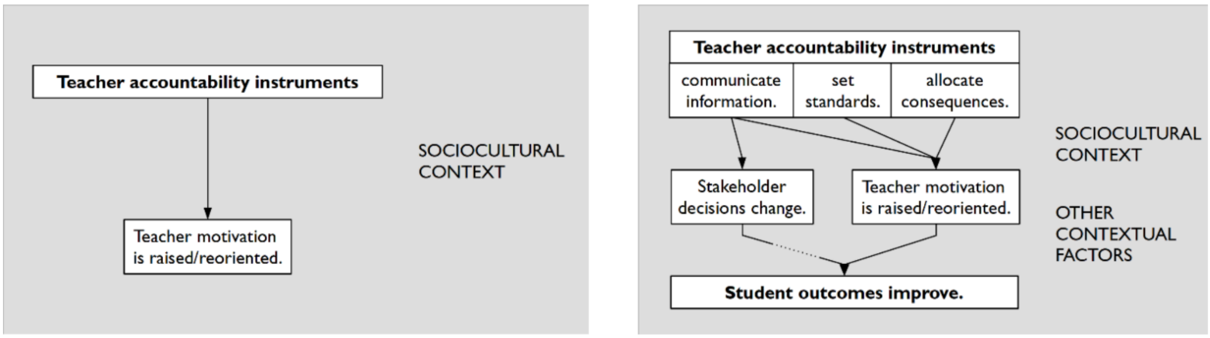
[43] Still, it is important to note that this preoccupation with meritocracy and competition does not encompass all Singaporeans. For example, Maggie spoke critically of how students hear messages from parents, teachers, and society about the importance of 'grades, grades, grades' and 'tuition, tuition, tuition' (i.e. private after-school tutoring), but noted that she 'was in a neighbourhood school' (i.e. a less prestigious, lower-achieving school), where 'students are, thankfully, blissfully unaware and less competitive about things' and consequently 'they actually grow in a more balanced manner'.

as being ‘characterized by a grasping or selfish attitude arising from a fear of missing out on something’. Singaporean *kiasuism* was iconically captured in the hugely popular Mr. Kiasu comic series, by local artist Johnny Lau. In a recent interview, Lau’s remarks on the *kiasu* outlook mirrored my interview participant’s observations on civic norms adherence. According to Lau, being *kiasu* ‘can also mean that you always have a goal, and execute it. Singaporeans do things by the book, but very efficiently’ (Lui, 2018). As I show below, interview participants believe that *kiasuism* is amply evident in how teachers respond to accountability instruments.

6.3 Teacher accountability, sociocultural context, and concepts of motivation

In Chapter 5, I showed how teacher accountability instruments can affect the level and/or direction of teachers’ motivation, and how these motivational effects depend on teachers’ subjective perspectives. In the previous section, I described some dominant patterns in Finland’s and Singapore’s respective sociocultural contexts, which I summarised as complementary responsibility and managed meritocracy, respectively. In this section, I bring the two threads together, exploring how these dominant sociocultural patterns influence teachers’ motivational responses to accountability instruments, as indicated in Figure 6.4.

Figure 6.4 *Relationship between Section 6.3 and the overall conceptual framework*


Teacher accountability instruments
Teacher motivation is raised/reoriented.
SOCIOCULTURAL CONTEXT
Teacher accountability instruments
communicate information.
set standards.
allocate consequences.
Stakeholder decisions change.
Teacher motivation is raised/reoriented.
Student outcomes improve.
SOCIOCULTURAL CONTEXT
OTHER CONTEXTUAL FACTORS

This section

Overall conceptual framework

To outline the argument, I find that a key aspect of sociocultural context is teachers’ implicit concepts of motivation. These implicit concepts of motivation influence teachers’ responses to accountability instruments. Moreover, there were clear differences between the concepts of motivation that were dominant among Finnish and Singaporean participants. This finding emerged inductively. Having identified interview quotations where participants linked teacher motivation, accountability instruments, and sociocultural context, I reviewed a range of psychological theories of motivation with the aim of identifying an empirically validated framework to lend structure to the interview analysis. However, I found that these theories vary substantially in the factors that they view as fundamental to motivation. None of the major theories (as identified by Schunk et al., 2010; and Vroom & Deci, 1992) emphasised motivational factors that were equally salient to the teacher interview data from both countries. However, two distinct theories of motivation did overlap with the concepts of motivation that are implicit in the Finland and Singapore interviews, respectively.

When Finnish participants discussed motivation, they emphasised factors that were also central to Ryan and Deci's (2000b, 2000c) self-determination theory. This theory specifies, among other things, that intrinsic motivation is fostered by feelings of autonomy, competence, and relatedness to others, as described above in Section 2.4. Meanwhile, Singaporean participants emphasised factors that were also central to Vroom's (1964) expectancy theory. In Vroom's theory, motivation is a function of expectancy, instrumentality, and valence. Expectancy refers to the degree to which an actor believes that their effort will lead to successful performance; instrumentality is their degree of belief that such performance will lead to desired outcomes (gaining rewards and avoiding penalties); and valence is the degree to which they value such outcomes. As noted in Section 2.4, these theories differ not only in the explanatory factors they prioritise, but also in the fields where they are applied. While Ryan and Deci's self-determination theory is often cited in education psychology, Vroom's expectancy theory is prominent among management scholars and practitioners.

Empirically, the boundaries between sociocultural context, concepts of motivation, and the level/orientation of teacher motivation are blurry, because all three are mental constructs. However, they are analytically distinct. Sociocultural context, which I defined in Section 2.2 as *dominant patterns of ideas and practices in a given social system that influence people's interactions with their environments*, encompasses the concepts of motivation that may be influential in a given setting. While concepts of motivation are general abstractions that may be affirmed by many people across a given context, the level and direction of a particular teacher's motivation is a person-specific mental state, which in turn manifests in that teacher's actions. In this analysis, I am primarily interested in how concepts of motivation influence teachers' interactions with accountability instruments. (I use 'theories of motivation' to denote formally articulated and psychologically validated frameworks for understanding motivation, such as Ryan and Deci's self-determination theory and Vroom's expectancy theory, while 'concepts of motivation' denotes the mental models that are implicit in participants' accounts of their motivation-related reasoning and experiences.)

Accordingly, for each country I (a) describe the concept of motivation implicit in participants' observations, illustrating its consistency with the larger sociocultural context; (b) show how the teacher accountability approach is compatible with the concept of motivation; and (c) demonstrate that the reported side effects of accountability instruments on teacher motivation are consistent with the implicit concept of motivation. Throughout this analysis, I use Vroom's and Ryan and Deci's theories to lend conceptual clarity.

Concepts of motivation and sociocultural context

When discussing motivation, Finnish participants often linked it to the competence, relatedness, and autonomy that Ryan and Deci (2000c) regard as fundamental to intrinsic motivation. For example, when I asked Juhani how sociocultural context influences teacher accountability, he said:

> There's something called Finnish *sisu* or stubbornness. Teachers are the kind of persons who are stubborn enough to feel the needs that the surrounding society gives them, and they will meet them. And they are also flexible enough to do it in a way that is quite effective.

Sisu is a Finnish word that is notoriously difficult to translate but relates to internal fortitude amid challenges (Lahti, 2019; Nylund, 2018; Partanen, 2016; Strode, 1940). From Juhani's perspective, this 'stubbornness' is directed toward meeting 'the needs that the surrounding society gives them', indicating a sense of relatedness. Moreover, he characterises teachers as being 'flexible enough' to meet these needs 'in a way that is quite effective', indicating competence. Similarly, when I asked Satu whether she received any financial rewards for extra effort in her work, she said that such rewards were negligible, but added that: 'It's inside me that I want to be good at my work. And I can meet parents knowing that I did my best, that I did things I didn't have to do.' Thus, Satu is intrinsically motivated 'to be good at her work', i.e. to be competent, and this motivation is linked to her sense of responsibility toward the parents of her pupils, i.e. a form of relatedness. Another Finnish participant, Antero, likewise said, 'We teach from our personality. From what is inside us. Not from these orders and laws, but from the inside.' He also noted that formal accountability instruments such as orders and laws can 'kill our creativity'. Additionally, Anneli linked motivation to autonomy: 'In Finland, teachers perform better, I think, when we feel that we are trusted, and we can do our work our own way.' Other participants echoed this belief.

Finnish participants' descriptions of unmotivated colleagues were also consonant with Ryan and Deci's theory. For example, Liisa said, 'I promised myself that when I'm tired of doing what I'm doing, then I will quit. I will never become that teacher who's just doing their job and not being interested in it.' Similarly, Helena contrasted 'good teachers' who are 'really innovating, and trying new things, and doing a lot of prep work for the lessons', and 'bad teachers' who are 'just doing their job, just the minimum requirements, or without any passion, or just doing the same thing they've done for years, with minimal effort'. Thus, rather than exploring and extending their capacities, which is a hallmark of intrinsic motivation (Ryan & Deci, 2000a), these unmotivated teachers stick to comfortable routines. All of this suggests that there is considerable overlap between Finnish participants' concept of motivation and Ryan and Deci's theory.

More importantly, Finnish participants' concept of motivation is compatible with the larger sociocultural context. After observing that Finnish teachers are more effective when they can work autonomously, Anneli added, 'But that's probably because our society is based on trust. In *so* many ways. So that's why it works.' This trust is not a carte blanche. Rather, the basis of this trust is expectation that other members of society will likewise fulfil their complementary responsibilities, as discussed in the previous section. In Päivi's words:

> Finnish people know what's wrong, and what's right. And they are very interested if someone near them isn't doing right. So I think teachers know very well if they're doing right or wrong, and if the other teachers are doing right or wrong. It's very much part of our culture.

An interview participant in Müller and Hernández's (2010) study made similar observations, adding that this social awareness is tacit rather than publicly expressed ('we know what type of people they are […] even if we don't talk about it', p. 313). The unspoken nature of such awareness may contribute to the fact that my Finnish interview participants did not feel that it impinged on their actions. In Masa's view, the ethos of 'background responsibility' in Finland includes 'an understanding from other people that, "Oh, this is the job that you're doing. I may not like it a lot, but I'm going to go along with it"'. Correspondingly, Masa attributed teachers' sense of autonomy to the fact that children are deemed responsible for their own schoolwork, such that 'the student's actions don't really reflect on the teacher', which is 'immensely liberating' for teachers. Thus, the expectation that actors at all levels of the education system will fulfil their respective responsibilities—from the Ministry and central government agencies through municipalities and schools to teachers and students, as Helena noted—safeguards teachers' sense of autonomy. In turn, this sense of autonomy enables teachers to 'perform better', as Anneli said.[44]

Singaporean participants emphasised a different set of motivational factors. While saying that 'remuneration […] is never a good way to assess the worth or value of a teacher', Eleanor also noted that 'Singapore's a very expensive country to live in, so [remuneration] does matter to a large number of people'. To use Vroom's (1964) terminology, salaries and bonuses have a large positive valence for many Singaporeans. This valence, together with the belief that good performance is reliably instrumental in reaping such rewards, can influences teachers' outlooks. According to Adeline, 'Singaporean teachers are very typical civil servants, and they like to have their various KPIs [key performance indicators] and know that if they meet them, they might get rewarded.' Faced with this reward structure, teachers may redirect their effort according to their expectancy of where it will yield the best performance, as Jane observed:

> Some teachers may feel that certain areas are less debatable, like exam results, so they will *chiong* [i.e. rush towards, put effort into] that area. Then maybe you look at your CCA [cocurricular activity]: 'Oh dear, it's not possible.' Whatever you do, it will be very hard [to win the inter-school competition], because maybe there's another champion school in your zone already. So you strategize in this way.

The concept of motivation that emerges from these descriptions—where motivation is driven by utility maximisation—matches Vroom's (1964) proposition that motivational force is a function of valence, instrumentality, and expectancy.

Singaporean participants' descriptions of unmotivated teachers also tally with Vroom's theory. As quoted in Sections 5.2 and 5.4 respectively, both Peter and Jane said that some colleagues decide to 'cruise' in school after calculating that it is more lucrative to invest their effort in private tutoring after school hours. In Jane's words, 'they look at the monetary reward': if earnings from private tutoring are greater than bonuses and salary increments from good performance in school, then private tutoring has greater valence. Jane also observed that some colleagues at the middle of the career ladder 'who have been in the system for more than 15 to 20 years' may be 'not so responsive and they drag their feet' when asked to complete tasks,

[44] Finland's complementary responsibility also safeguards the autonomy of those working in other professions as well. Helena related an anecdote of how some foreign exchange students at her school reacted to school cafeteria food that seemed tasteless to them: 'Their solution to it was very simple: "Why don't you just tell the kitchen to make spicier food?" Which makes sense from the point of view of, "The kitchen staff's job is to serve me food," and, "They're lower in the hierarchy." But from our point of view, the kitchen is run by a company, which we buy the service of. And they're just people working in the kitchen. And they don't come and tell me how to teach maths, so I wouldn't feel like it was my job to tell them how to make food.'

because ‘they are the type who say, “Okay, fine. If I haven’t moved up the ladder by now, I will never move.”’ That is, they do not believe that performing additional work will be instrumental in promotions. Additionally, Eleanor and Maggie noted that some teachers may be reluctant to try new pedagogical approaches because, in Eleanor’s words, ‘they’re still getting their ‘A’s, so why rock the boat?’ That is, these new approaches are ambiguous in their expectancy, so channelling effort toward such approaches may not yield successful performance.

Strikingly, echoes of expectancy theory appeared even in the observations of participants who themselves disavowed Singapore’s competitive, progression-oriented system. For example, Andy noted that the *kiasu* mentality described in the previous section does not apply to him and his ‘band of merry colleagues who are just interested in developing the students’. However, he added that ‘we do recognise those who are deserving of credit because [...] something about them enables them to go above and beyond for the students, and we don’t begrudge them if they are rewarded accordingly’ (as quoted in Section 5.2). Thus, despite opting out of the meritocratic race, he endorsed its instrumentality, stating that superior performance is ‘deserving of credit’. This principle underpins the meritocracy that dominates Singapore’s sociocultural context.

Motivation also interacts with the extensive government management of Singapore’s meritocracy. In Jeffrey’s view, Singapore’s dense urban geography enables a high degree of central management, which in turn compels effort:

> You’re kind of fenced in. […] So that very much forces you higher up on the value chain. You basically can’t choose to slack off. […] And that forces everyone to survive in the education system. I mean, you can say publish or perish—but you don’t even have the option (laughter) in the education system. It’s pass or distinction. That’s it. (laughing) Because no matter how badly you do, they’re going to put you somewhere […] that they have specially designated for you. [45]

Similarly, in a study on performativity under the EPMS, Liew (2012) remarks: ‘Teachers that thrive and survive must perform, or else perish’ (p. 300). Thus, in this managed meritocracy, the

[45] Eleanor made an observation that echoes Jeffrey’s remarks on government-directed sorting and stratification in the education system. After describing how a partnership with another school prompted her school to ‘push their students’ to get better grades, she said: ‘Why I think the Ministry thought they needed to [introduce such partnerships] is to provide education options. Because on the ground people were saying that they wanted to have the kids stretched in different ways. […] I think you’re right to say that the socio-political climate affects the way in which education moves. So in the past it was about getting an education for everybody. And that was good enough. But then it moved towards getting a certain *kind* of education. And then they created N(A) [Normal (Academic) stream], N(T) [Normal (Technical) stream], Express [Express stream], GEP [Gifted Education Programme]; later moving towards this IP [Integrated Programme], and then the ITE [Institute of Technical Education]. And so the government really has a very big hand in deciding how the people are educated.’ For an account of these changes in Singapore’s educational offerings over time, see Goh and Gopinathan (2008). For a more detailed discussion of alternative education pathways offered to students who perform below typical grade-level expectations, see Ng (2017).

government has the power to manipulate motivational levers. Besides 'specially designating' roles at suitable levels of the value chain, the government also determines extrinsic reward structures. As I discuss below, the government uses these reward structures to maximise aggregate teacher motivation. Here it will suffice to note that the utility-oriented concept of motivation implicit in participants' statements is fully compatible with the managed meritocracy in which it is embedded.

Thus, Finnish and Singaporean participants had different mental models of how motivation operates. These models map onto different psychological theories of motivation, and cohere with the broader sociocultural contexts. Still, it is important to note that these mental models are not universally shared within each setting. For example, Adeline (Singapore) observed that some teachers 'who have been in an independent school for a really long time' might be 'more Finland-ish', in that 'they're experienced, they don't care about career progression, and they know how to do their job'—thus emphasising a sense of competence and intrinsic motivation. (As noted in Section 3.4, independent schools have more administrative autonomy than the typical school, and tend to be prestigious and high-performing.) In turn, some Finnish participants noted that certain teachers are, to use Satu's words, 'very exact', such that they will reject additional responsibilities on the basis that '[they] have done what [they] have been paid for'—thus following the effort-reward calculations of expectancy theory. However, participants presented both of these categories as exceptions to the norm. In general, Finnish participants regard the motivation to teach as an intrinsic force within the landscape of complementary responsibility. Singaporean participants, enmeshed in their managed meritocracy, regard motivation as an outcome-oriented process, although some teachers prize pecuniary outcomes and others prioritise pupil development for its own sake.

Concepts of motivation and teacher accountability instruments

As shown in Chapter 5, teacher training plays a vital standard-setting role in Finnish teacher accountability. Both the selectivity of the training process and rigour of the training itself ensure that teachers are motivated and competent. Importantly, teachers are not only competent in that they are capable of doing their jobs, but they also *feel* themselves to be competent—in Liisa's words, 'We know what we're supposed to be doing, and then we do it.' Ryan and Deci (2000a) argue that people have a basic need to feel competent, and that the fulfilment of this need supports intrinsic motivation. This sense of competence was also evident in Kristiina's observation that Finnish teachers, when asked to explain particular pedagogical choices, often say, 'Of course, I just know that was the right thing to do.' She added that 'it's very difficult to pinpoint' what is special about Finnish education, because 'so much of it comes from the inside'—another overlap with the notion of intrinsic motivation. Relatedly, Masa observed that:

> Finnish teachers have a master's degree. There's this sort of professional pride [...] and a level of respect that goes with being a teacher. If somebody hears you're a teacher, [...] it's like, 'Okay, that's pretty impressive.' And I think the combination of those factors means that the internal [desire to excel] from the teacher is more important than anything else.[46]

The factors that Masa identifies as crucial to teachers' intrinsic motivation suggest not only a sense of competence, but also a sense of relatedness. 'Professional pride' implies relatedness to colleagues and to collegial standards, while societal respect implies relatedness to society.

This relatedness is strengthened by other aspects of Finland's teacher accountability approach. The national curriculum provides 'a common ground for the students', in Liisa's words. Liisa also emphasised that 'there's no teacher who's not aware of the curriculum'—which was borne out in all of the Finland interviews. Thus, teachers are united in following the same overall guidelines to serve students, reinforcing both collegial and societal relatedness. Collegial relatedness is also maintained by the common pay structure. As Emilia noted:

> If you have a lot of rewards and people are evaluated against each other, then that makes a lot more competition, and that would eat away at collaboration. [...] I wouldn't say no to more money, (laughter) but I'm very happy that we don't have a system of rewards and punishments, actually.

Similarly, Anneli said that teachers 'should get paid equally' because they all gained their qualifications through the same duration of study, so 'the baseline is that every teacher is as competent as everybody else'. Juhani also observed that performance-based compensation would undermine the egalitarianism 'that is inside Finns'. He mentioned that some Finnish studies have

[46] At this point the interview, Masa's actual words were 'internal *stuff* [emphasis added] from the teacher'. However, later in the interview he used the phrase 'internal reason, motivation' which he summarised in his next sentence as 'internal stuff', and which he referred to in the subsequent sentence as 'that internal desire to excel'—hence my substitution of the latter phrase into his quote here. Elsewhere, Masa also spoke of teachers being driven by 'something inside says, "This is what needs to be done"'.

found that 'when you reward people with money, it gives satisfaction for a shorter period of time than when you are valued by the society you work in'.[47]

Besides contributing to teachers' sense of relatedness, Finland's lack of formal monitoring and consequences supports their sense of autonomy. According to Antero,

> When they selected me to study teaching, of course they checked that my personality and who I am fits the job. And after that, I have been on my own. Nobody has come here to say that, 'You must change. And you must do it like this, not like that.' I am in charge here.

This autonomy is premised on the expectation that teachers are willing (i.e. motivated) and able (i.e. competent) to teach well without external inducements, because they '[fit] the job'. As Helena observed, quoted in full in Section 5.2, 'the assumption is that we teachers are professionals who know their job and have the skills and that we are all interested in the same goals'. (See Partanen, 2016, p. 127 for a very similar observation.) Thus, Finland's approach to teacher accountability enables a mutually reinforcing relationship between teachers' competence (knowing their job, having the skills), relatedness (all being interested in the same goals), and their autonomy. Specifically, teacher training establishes competence and relatedness; widespread belief in teachers' competence and relatedness grants them professional autonomy; this autonomy, together with competence and relatedness, supports their intrinsic motivation; and this motivation helps to maintain their competence and the public belief therein.

Singapore's teacher accountability approach follows a different but equally coherent logic. According to Mark:

> On one hand, most teachers view the EPMS […] like any rule that the government puts out, so they have to abide by it. But they do believe that it's another example of the meritocracy in action. […] That the harder they work, the more strategic they are about

[47] As some Nordic commentators note, this emphasis on maintaining the appearance of equality has a dark side. Booth (2014) describes a phenomenon called Jante Law, so named because of a Danish novel lampooning social life in a town called Jante. The principles of Jante Law include, 'You shall never indulge in the conceit of imagining that you are better than we are' and 'You shall not believe that you are more important than we are' (ibid, p. 90). Jante Law may contribute to the preservation of egalitarianism and relatedness, but it is usually discussed critically, as a source of oppressive conformity and petty envy. Despite its Danish origins, Booth (ibid) and Partanen (2016) believe that Jante Law applies to other Nordics, including the Finns. Partanen, herself a Finn, notes that Jante Law is meant as 'a critique of a rather sad aspect of the Nordic character that is often taken too far', adding that 'successful Finns have been to known to feel that other Finns are jealous or disparaging of their achievements', and that 'sometimes Finns betray a distasteful fondness for schadenfreude when a successful person falters' (2016, p. 283). Interview participant Antero's account of financial bonuses for teachers demonstrates a Jante Law-like pattern at work: 'The Finnish mind is something like this: if I get a bonus, for example, I think "Not me. Take it back. Give it to her. Give it to him. Not me." […] We are little bit ashamed, or something like that. And the other side of this coin is if somebody else gets the bonus, then we others think that, 'Hmm, that's very interesting. She's a little bit lazy. Pupils don't like her. Why did the principal give the money to her?' And we are a little bit jealous, of course. But that's the Finnish way of thinking, I think.'

> their work, the higher their performance ranking, and the larger their performance bonus would be. […] And that can be taken positively or negatively.

Thus, if teachers highly value performance bonuses (i.e. if bonuses have a high valence), they will make strategic decisions to get higher performance rankings (i.e. allocate their effort based on expectancy), because they believe that the EPMS compensates hard work (i.e. they believe in its instrumentality). However, the teacher accountability system also influences teachers for whom performance bonuses have little valence. For example, Joseph believed that most teachers were in the profession 'not for extrinsic reasons, but mainly for intrinsic reasons'. Instead, as he said elsewhere in the interview, his teaching experience had 'really made [him] really believe in what MOE wants' because 'over time you realise the value of your work' such that 'it's not so much about the recognition that you get, but you just do it out of your own personal passion'. Thus, his motivation is rooted in 'the value of [his] work', i.e. a different category of outcomes. Financial rewards may have low valence for Joseph, but he is driven by the valence of 'what MOE wants', i.e. a vision of shared productivity for Singapore's survival. Like Joseph, four other Singaporean participants (Peter, Geok Ling, Jane, and Timothy) also accorded little value to EPMS rewards but likewise said that the Ministry's expectations of teachers broadly overlapped with their own. Hence, the EPMS facilitates alignment between the teachers whose goals are primarily altruistic, and those whose goals are primarily material.

(That said, despite this broad agreement with Ministry expectations, teachers were not uncritical of Ministry goals. For example, Geok Ling said she disagreed with the emphasis on committee work, event organisation, and 'certain things that they evaluate teachers on, like organisation awareness and whether you have got helicopter view […], that are really very corporate'. In Singapore government parlance, 'helicopter view' refers to the ability to take a systems-level perspective, while still paying attention to necessary details [J.-M. Ho & Koh, 2018; Quah, 2010]. In turn, even though Jane said she agreed with 'the big picture things that they say, like "mould the future of the nation" and 'every child counts"', she questioned the practical feasibility of some Ministry stipulations.)

The EPMS is also designed to foster alignment between teacher appraisal standards nationwide, in order to maintain the perception that EPMS gradings are fair. As noted in Section 5.4, Singaporean participants observed that unfair appraisals can be highly demotivating. In Maggie's words, such unfairness makes you 'feel that whatever effort you put in is not *worth it* [emphasis added]'. To use Vroom's (1964) terminology, teachers will only be motivated if they believe that the EPMS allows good performance to be instrumental for desirable rewards. The Ministry invests considerable resources into maintaining this instrumentality. For example, Jane also noted that 'they always call [performance grades] a collective decision', and that an external moderator joins ranking meetings to ensure consistency across schools. Similarly, Andy said he 'personally' felt that EPMS rankings were 'a fair and consistent kind of measurement', since his school principal had 'gone to great lengths to actually explain that it's done before a ranking panel', rather than depending on one supervisor's or school leader's vagaries. However, he added that some colleagues believe that EPMS rankings depend on 'the willingness of your superior to actually fight for you at such ranking discussions and whether you have offended certain people'. On the whole, most participants agreed that the grading system was flawed, but not unacceptably so. In Sonia's view, 'I would say it's 80 percent accurate, although sometimes it does not match what you think you should have gotten.' Hence, the EPMS is viewed as an imperfect but adequate arbiter of teachers' merit.[48]

Some participants implied that the EPMS is intended not only to raise and reorient the motivation of Singapore's teachers, but also to influence other actors within Singapore's outcomes-oriented ecosystem. When asked whether Singapore's teacher accountability approach made it easier or harder to be a good teacher, Timothy said that he '[did]n't want to give it credit by saying it makes it easier', and he noted that he had 'many reservations' about the way the EPMS was carried out, but conceded that:

> It's good to have some expectations set in stone for me to base myself against, and to know what exactly you can potentially reach. […] We always tell students, 'You must aim for something high. You cannot aim for just mediocrity.' So, I guess, in the spirit of what we do, it's just putting into practice what we preach.

That is, if children are to learn how to motivate themselves in a meritocracy, they should see it modelled by their teachers. Zooming out from classrooms to national politics, Eleanor suggested that the EPMS indirectly factors into voters' calculations of the value of supporting the government. When asked what would happen if Singapore replaced the EPMS with Finland's accountability approach, she said:

> My immediate response is, it's not going to work. […] It will be very teacher-dependent, and so it will be the luck of the draw if your child gets into this classroom where this teacher is a bit more progressive or has more initiative, then you benefit. Whereas in other classrooms where the teacher is just happy with what she or he is

[48] One reason why teachers regard EPMS rankings as being reasonably fair is because they believe that egregiously unfair rankings would be reflected in an undesirable grade for the school leadership team in the biannual School Climate Survey, as mentioned by Jane and Joseph. Conducted nationwide every other year, the School Climate Survey allows teachers to anonymously evaluate their schools (J.-M. Ho & Koh, 2018; Singapore Teachers' Union, 2011)—thus, maintaining meritocracy among schools and school leaders, just as the EPMS maintains meritocracy among teachers.

> doing, then the child is not going to learn as much. So that inconsistency is something the government will not want to risk, because the feedback from the people will be quite strong.

That is, the EPMS maintains a sufficient baseline of teacher motivation, which in turn maintains consistent educational quality, which maintains public belief in the government's merit. Not only is Singapore's approach to teacher accountability fully compatible with interview participants' concept of motivation, it is also fully integrated into the managed meritocracy that is evident in its social milieu.

Concepts of motivation and side effects of teacher accountability instruments

Thus far, I have shown that Finland's and Singapore's approaches to teacher accountability are compatible with participants' implicit concepts of motivation, which are themselves compatible with the wider sociocultural contexts. However, this does not mean that the accountability instruments have uniformly desirable effects on teacher motivation. Given the complexity of motivational processes and the heterogeneity within any group of teachers, accountability instruments can have less-than-desirable side effects. In Chapter 5, I outlined some general categories of side effects and discussed their underlying mechanisms. Here I offer a different perspective. For each country, I describe specific unintended effects that participants reported, and explain these side effects using the respective concept of motivation.

To begin with Finland, I noted in Section 5.3 that some Finnish teachers are demotivated not because feedback is burdensome or biased, but because it is absent. According to Päivi,

> Our headmaster is really so busy that he doesn't have much time for us. So he's happy when we are doing our work and he doesn't hear anything about it. (laughter) […] But I know that there are teachers who would like to have more help. And who would like to have more thanks. They think that the headmaster doesn't *care* about what they are doing. And they would like to have, 'Oh, that went well. That was nice.'

As discussed above, this lack of monitoring and feedback is based on the assumption that most teachers are competent and intrinsically motivated. It is also compatible with the preservation of teacher autonomy, which itself is central to teachers' intrinsic motivation. However, if the lack of feedback diminishes teachers' sense of relatedness (e.g. if they 'think that the headmaster doesn't care about what they're doing'), intrinsic motivation may be compromised. Deci and Ryan (2000) also propose that good feedback can reinforce feelings of competence, which in turn boosts motivation. Even among those who did not say that the relative lack of feedback reduced their motivation, several Finnish participants (Anneli, Antero, Helena, Liisa, Maarit, and Satu) said, unprompted, that they would welcome lesson observations in order to get feedback on their teaching practice, and they lamented the lack of lesson observations in Finland—which aligns with the proposition that intrinsically motivated people seek to improve their capabilities.

Another side effect of the Finnish approach to teacher accountability is that certain curricular changes may have no impact on teacher practice. Despite teachers' general compliance with the curriculum, their autonomy and sense of competence may be such that they disregard curricular changes that clash with their beliefs about how best to fulfil their responsibilities. For example, Päivi mentioned that even though the Ministry currently prefers teachers to pose open-ended questions without supplying answers to students, she nonetheless preferred to 'tell them what's important and what's not important' because they were 'still so small'. She added, laughing, 'Yes, I should give more space to the children, and not teach so much. But I like teaching, and I think I know how things work.' Similarly, Antero said he was 'happy that the principal trusts [him] so much' that he could channel his effort away from the curricular goals that are not congruent with his conception of his responsibilities. When asked whether accountability instruments make it easier or harder for him to be a good teacher, he replied, evoking a national icon:

> I'm a slightly old-fashioned teacher. I like the Finnish Formula 1 driver, Kimi Räikkönen, when he shouted in his team radio, 'Shut up, I know what I'm doing.' So I think in here [i.e. his classroom], too, I know what I'm doing. My focus is on the pupils. And I am on the right path.

Liisa and Satu also said that curricular changes may not have any effect on the practice of veteran teachers. The fact that they both mentioned hypothetical teachers who have been teaching for more than 'thirty-five years' suggests that there may be a local trope of the stubborn teacher approaching retirement. Still, at different points in their interviews, they each also mentioned such autonomy in relation to themselves. In Satu's words, 'This curriculum is not so important, if you have worked for a long time. I know what I have to do.' Likewise, Liisa noted that the current curriculum has many expectations, some of which were impractical: 'So, I don't want to punish myself for not doing everything, when nobody else is. […] It's just me and my conscience, and if I feel that I'm doing most of it, and I cannot do more, then I'm happy.' (For another example of Finnish teachers criticising the curriculum, see Y. Li & Dervin, 2018, p. 49.)

However, in situations where Finnish teachers lack the capacity to resist external controls, their motivation may suffer. Juhani observed that if he is ordered to do something at a particular time and in a particular way, he instinctively resists, because 'that takes out all the possibilities you can create in it' and 'it gives the signal that "I think you're not quite good enough"'; whereas 'when it's based on the trust that everybody tries to learn and educate oneself and become a better teacher, it works'. (Liisa made a similar observation.) A specific instance of motivational harm was given in Masa's experience of Finland's recent *Kilpailukykysopimus* (KIKY, i.e. competitiveness pact) policy. One KIKY stipulation is that all full-time employees in Finland must work an extra 24 hours annually without additional pay (SAK, 2016; see also Y. Li & Dervin, 2018, p. 108). According to Masa, this was an 'extreme *dis*incentive':

> Teachers started complaining about being a teacher, which was new, for me. People were starting to think about changing jobs. Because all of a sudden, we've got this bureaucracy that's keeping track of this time. They're realising that we already do more than that anyway. And […] now they're doing it for the time, which equates to money, rather than because something inside says, 'This is what needs to be done.' So the teachers' personal standards have dropped.

This decline in 'teachers' personal standards' and their internal senses of 'what needs to be done' is consistent with Ryan and Deci (2000a) argument that unmet needs for autonomy may jeopardise intrinsic motivation.

In contrast, several Singaporean participants noted instances where teachers' motivation decreased because of lowered expectancy or instrumentality. Sonia gave an example of the former:

> There are only so many extra duties available to go around, to help you stand out from the crowd. And if everyone's going to be grabbing them, those teachers who are in it generally for teaching end up taking the hard step of leaving the service, because this accountability procedure demotivates them so much. It is a huge bone that people have to pick with the system.

Thus, if teachers believe that they will not be able 'to stand out from the crowd' (i.e. their effort will not lead to the desired performance), this diminished expectancy may be demotivating. Eleanor made a similar observation. As for instrumentality, Peter noted that reporting officers who are unfair or who do not convey their expectations clearly 'can be quite demoralising for a teacher who's put in a lot of effort, and then maybe is told that it's not good enough, or just does not get any validation for it'. Thus, when effort and performance do not lead to valued outcomes, this diminished instrumentality may be demotivating. (Similar observations about perceived unfairness from Sonia and Maggie were quoted in Sections 5.3 and 5.4, respectively.)

Another side effect of such competition for promotions and bonuses in Singapore is that some teachers may channel their effort away from fulfilling the spirit of their responsibilities, and toward prominent portfolios instead. Such strategies were mentioned by several participants. For example, Timothy said that:

> The *kiasuism* of teachers versus teachers, to put it bluntly, may colour how some teachers choose to showcase what they're doing, in a very targeted and purposeful manner. And I mean this in a negative way. I mean that they […] are doing it for the sake of being more visible, because they already have in mind what they want […] and then they are working towards it.

Thus, teachers' calculations of how to optimise positive-valence outcomes may prompt grandstanding that diverts motivation and resources away from student development. Grandstanding strategies run the gamut, from a colleague who only bothered to prepare presentation slides when their lesson was being observed (as Joseph reported), to 'climbers' aiming for promotions who '*force* their subordinates to embark on projects' (as Andy reported). Additionally, Liew (2012) found that some teachers complete their EPMS work review forms using strategies to present themselves in disproportionately favourable light.

Another unintended reorientation of motivation occurs when teachers channel substantial effort toward preventing negative-valence outcomes. This stems from the emphasis in Singapore's accountability system not only on rewards, but also on punishments. While several participants talked about the *kiasu* (literally, 'afraid of losing') mentality, as discussed in the previous section, Adeline and Maggie also noted that teachers are *kiasi*—literally, 'afraid of death', or excessively cautious. In Maggie's experience,

> It permeates the whole system. We have to cover our butts. […] You could speak about it in more positive words, such as, 'Let's ensure that we do our very best to protect the child, or to ensure their welfare.' But very frequently, in informal conversation, you may meet colleagues speaking about it in terms of, 'Oh, we have to protect ourselves.'

She added that she doesn't 'want to do things for that purpose alone', but she had been in a situation where 'if you leave a bit of weakness visible, then certain people can be like hyenas coming after you—even though it's not your fault'. Mark made a similar comment about accountability and risk-aversion, saying that, 'from a cynical perspective', one advantage of the EPMS in 'the ass-covering society that is Singapore' is that it 'allows for some paperwork to show why someone got a grade, or to prevent any supposed doubt'. To give a specific example, Jane mentioned that this cautiousness meant that 'if you meet up with a parent, you make sure you have documentation, and then you make them sign it', thus adding to the already weighty paperwork burden.

All the aforementioned side effects of Singapore's teacher accountability approach may somewhat dampen overall educational outcomes. More worryingly, one other unintended effect of the EPMS is that its orientation toward rewards and penalties (and Singapore's competitive orientation more broadly) has hampered a major policy platform that the Ministry has been pushing since the mid-2000s: a shift toward a more holistic, less exam-oriented view of education (H. L. Lee, 2004; Singapore MOE, 2013). In Peter's words: 'MOE has taken steps towards shifting the focus away from grades. […] All of that is great in terms of what it seeks to achieve. But, honestly, it hasn't changed that competitive culture in Singapore.' The inertia hampering this policy change was also reported by Sonia, Andy, and Eleanor (see also Ab Kadir, 2017; Hogan et al., 2013; Lam, 2014). Similarly, Adeline observed that, despite the Ministry's push to move away from the preoccupation with grades,

> it's still very entrenched in the exam-based mindset that you need to do well in order to get good grades, to go everywhere. And even parents buy in to that mindset. […] A lot of teachers want to buy in to the shift away from exam-based education. […] We're quite torn between, 'Yes, we believe in this more holistic education'—but yet we know that, in the end, the students will be judged for their exam grades.

To use Vroom's (1964) terminology, this ingrained 'exam-based mindset' means that 'good grades' have higher valence than the outcomes of 'holistic education'. Even if teachers themselves may value the holistic alternative, the awareness that 'students will be judged for their exam grades' means that grades retain their high positive valence. The 'shift away from exam-based education' not only has lower valence, but also lower instrumentality, since getting good grades is a familiar path to success, but the benefits of holistic education are less certain. Hence teachers do not invest much effort into this attempted shift, consistent with what expectancy theory would predict. Notably, Ryan and Deci's (2000c) self-determination theory would instead predict that the attempted shift away from grades would raise teachers' intrinsic motivation: if teachers personally concur with this shift (as in Adeline's account), it would increase their sense of autonomy, thus boosting their motivation. However, participants reported low adherence rather than mobilised effort. Put differently, the way in which Singaporean teachers have responded to this accountability-related policy platform does not fit the concept of motivation articulated by Finnish participants. Rather, it fits the concept of motivation articulated by their Singaporean counterparts—as well as the broader sociocultural context in Singapore.

Summary

In this section, I have shown that Finnish and Singaporean participants understand motivation in fundamentally different ways. Finnish participants associated motivation with an internal drive to fulfil responsibilities to those around them. This drive is supported by teachers' confidence in their abilities to fulfil their responsibilities, and it is sustained by the autonomy that teachers enjoy. Singaporean participants associated motivation with the pursuit of outcomes, whether these outcomes relate to student growth or personal gain. The force with which these outcomes are pursued depends on both their desirability and their attainability.

Each of these concepts of motivation is consonant with the sociocultural context in which it is embedded. Each concept is also compatible with the respective teacher accountability approach. In Finland, strong standard-setting instruments safeguard teachers' competence levels and maintain their relatedness to colleagues and students, and the sparseness of formal instruments for monitoring and rewarding performance ensures that teachers have the autonomy that they regard as crucial to intrinsic motivation. In Singapore, an interlocking system of standards, information-gathering tools, and performance-based consequences helps to align teachers' actions toward system-level plans, whether the teachers personally prioritise social or economic goals.

It is worth noting that, among the six sociocultural constructs that I included in the statistical analysis, civic norms comes the closest to capturing these concepts of motivation that influence teachers' responses to accountability instruments. When asked about the civic norms statistics, participants' comments reflected many of the qualities that relate to their understandings of motivation: Finnish participants said that it was important to follow civic norms, i.e. fulfil their responsibilities, unless they could fulfil those responsibilities more competently through some alternative route; whereas Singaporean participants said their compatriots were likely to subvert norms for personal benefit, i.e. for positive-valence outcomes, but they would not risk being caught breaking the law, i.e. they would avoid negative-valence outcomes. These mental models of motivation were also evident throughout participants' descriptions of how teachers respond to accountability instruments. The consonance between beliefs about norms and concepts of motivation 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 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, like participants' concepts of motivation, relates to why people do what they do. Based on the evidence in the field interviews, I believe that this 'why' is socioculturally contingent and crucial to teacher accountability.

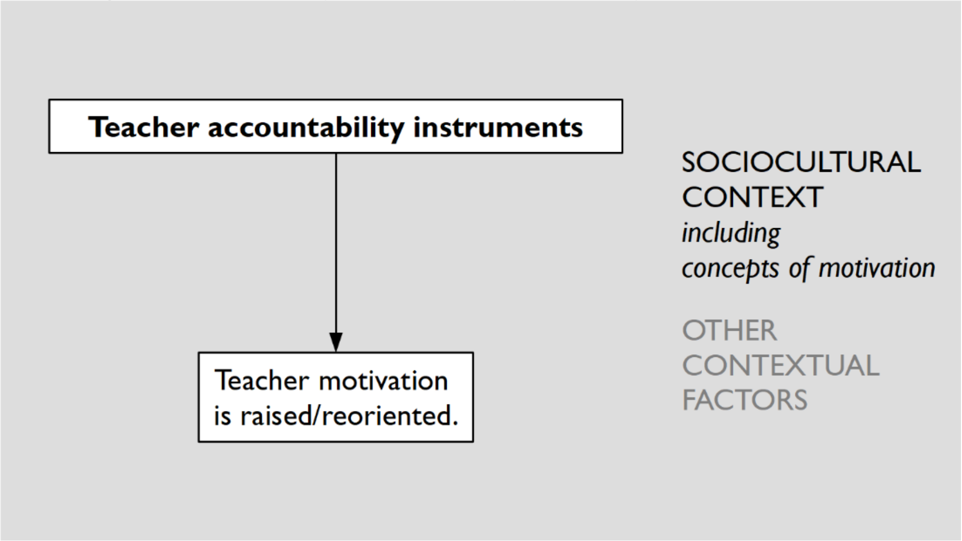
6.4 Discussion

This chapter has investigated RQ3, *to what extent does the influence of teacher accountability instruments on teacher motivation depend on sociocultural context?* As in Chapter 5, the ILSA data and the teacher interviews imply different answers. While the TALIS 2013 data suggested that higher levels of social trust and of confidence in institutions may intensify the negative effects of teacher accountability instruments on teacher job satisfaction, these results were not consistent across the TIMSS 2015 and PISA 2012 data. Results for the main (i.e. non-interacted) effect of accountability instruments on teacher motivation were similarly inconsistent, as were the results for moderation by other contextual variables. Hence, the ILSA data suggests that sociocultural context affects the relationship between teacher accountability and teacher motivation to a very limited extent, if at all.

However, the teacher interviews suggest that sociocultural context strongly shapes the relationship between accountability instruments and teacher motivation. Specifically, participants' mental models for understanding motivation are closely tied to their sociocultural contexts. These concepts of motivation, in turn, influence their responses to accountability instruments, whether consciously or unconsciously. For example, Singaporean participants' motivational responses to accountability instruments can be explained using the valence-instrumentality-expectancy model (Vroom, 1964) that matches Singaporean participants' implicit concept of motivation—but at least some of their responses cannot be explained by Ryan and Deci's (2000b) model of intrinsic motivation, which matches Finnish participants' implicit concept of motivation.

Figure 6.5 summarises the evidence presented in this chapter for the conceptual framework. While the ILSA analysis did not find robust evidence of a moderating role for either sociocultural context or the other contextual variables, the interview analysis found considerable evidence that sociocultural context does play such a role. Given that teacher motivation entails mental processes that are better captured in individual interviews than in statistical aggregates, I believe that the answer to RQ3 is that the influence of teacher accountability instruments on teacher motivation depends on sociocultural context to a large extent, at least in Finland and Singapore. (I discuss the cross-country generalisability of these interview findings in Section 7.3.)

Figure 6.5 *Evidence for the conceptual framework from the RQ3 analysis*


Teacher accountability instruments
Teacher motivation is raised/reoriented.
SOCIOCULTURAL CONTEXT *including concepts of motivation*
OTHER CONTEXTUAL FACTORS

In arguing that different sociocultural contexts may entail different concepts of motivation, I echo psychological studies that posit context-dependent models of motivation (e.g. Klassen et al., 2010; J. Li, 2012; Plaut & Markus, 2005). For example, S.S. Iyengar and Lepper's (1999) experimental work found that Anglo-American children had higher intrinsic motivation when they were free to choose their own tasks, whereas Asian-American children had higher intrinsic motivation when the task was ostensibly chosen by their mothers or classmates. In my fieldwork, Finnish participants' emphasis on autonomy mirrors these Anglo-American children's demonstrated preferences. Others have found that Chinese and Japanese mothers were more likely than Caucasian-American mothers to attribute unsatisfactory mathematics performance to their child's lack of effort (Hess, Chang, & McDevitt, 1987; S. D. Holloway, Kashiwagi, Hess, & Azuma, 1986). This emphasis on effort was also evident in Singaporean participants' descriptions of motivation.

That said, concepts of motivation are only one facet of sociocultural context—and some other facets are also salient to teacher accountability. In another analysis of these Finland and Singapore interviews, I explain Finnish and Singaporean teachers' comparably positive responses to their disparate teacher accountability systems by showing that each system accords authority to those who are most trusted to deliver desirable educational outcomes.[49] In Finland, trust is distributed among stakeholders at each level of the education system, such that the most trusted actor at the level of the classroom is the teacher—hence their great latitude in classroom practice

[49] This analysis was also conducted during the course of my PhD, for a chapter in an edited volume that is currently in preparation for publication. For details, see footnote 32.

(see also Aho, Pitkänen, & Sahlberg, 2006). In Singapore, trust is instead concentrated on the government, hence the government-led performance management system (see also S. K. Lee et al., 2008, 'Conclusion'). It would also be possible to use these interviews to argue that the effectiveness of these contrasting teacher accountability systems depends on perceptions of efficacy (i.e. what works) or perceptions of fairness (and, by extension, legitimacy, as discussed in Section 2.4), although these arguments would have somewhat thinner substantiation from the interviews than the concepts of motivation or the distribution of trust arguments.

The fact that this argument can be constructed along different narratives is unsurprising, given the complexity of sociocultural context and human behaviour. Still, concepts of motivation offer a useful organising device for the purposes of this thesis, which investigates the validity of a motivation-centric framework for teacher accountability. Concepts of motivation may likewise be a useful area to consider for teacher accountability policymaking that seeks to influence the level and/or direction of teacher motivation. Moreover, as discussed above, there are established psychological research programmes supporting not only the context-dependence of concepts of motivation, but also each of the concepts of motivation that I outline.

More importantly, in arguing that sociocultural context influences teachers' motivational responses to accountability instruments, this chapter aligns with other studies underscoring the importance of tailoring educational management to the local context (Pritchett, 2015; UNESCO, 2017; A. Walker & Dimmock, 2002). As discussed in Section 2.1, there is abundant evidence of teacher accountability instruments working in one context but failing in another. Based on the evidence in this chapter, I conclude that one way to forestall such failures is to include sociocultural context among the elements considered in the design of teacher accountability policy.

One challenge to my attempt to claim a causal pathway from accountability instruments to teacher motivation, conditioned by sociocultural context, is that policy instruments and sociocultural context are mutually influencing. That is, sociocultural context influences the choice of policy instruments, and chosen policy instruments influence sociocultural context down the line. This feedback loop makes it difficult to identify cause and effect.

This difficulty is heightened by the fact that policy instruments can be designed expressly to influence sociocultural context. For example, the Finnish government has a unit devoted to building Finland's international image. Its strategies include coordinated messaging that emphasises Finns' belief in equality, social responsibility, and education, among other things (Finland Promotion Board, 2017a, 2017b). In Singapore, the National Education campaign meant that from 1997 to 2017, students regularly encountered six specific statements about Singapore—including 'We must uphold meritocracy and incorruptibility' and 'No one owes Singapore a living'—whether on posters, at annual school events, or in social studies lessons (Chia, 2018; Sim & Print, 2005; Singapore MOE, 2018a; J. Tan, 2010). The efficacy of Singapore's national messaging is apparent in my field interviews: many participants used the government's favoured terminology conversationally. For example, Joseph said that: 'The Singapore teacher accountability system is based on the context where, to ensure survival in Singapore, our manpower and the economy need to be productivity-driven.' The emphases on ensuring survival, building manpower, and being productivity-driven are all common government refrains.[50] While these culture-influencing instruments may not be explicitly connected to teacher accountability, it would be difficult to deny the possibility of such a link. For example, Low (2016) notes that both the Singaporean and Finnish governments consistently put forth advertising campaigns about teachers' valuable role in nation-building (see also Crehan, 2016 on Singapore; and Ripley, 2013 on Finland)—which may well influence standard-setting in teacher accountability.

Given these feedback loops between policy and culture, one alternative explanation for the sociocultural compatibility of Finland's and Singapore's respective teacher accountability approaches, as experienced by interview participants, is that their sociocultural contexts affect the teacher accountability causal pathway only at the point of policy design—rather than also affecting teachers' motivational responses to accountability instruments, as I posit in Figure 6.5. This alternative explanation could proceed as follows:

Teachers' motivational responses depend solely on the accountability instruments themselves. Holding all else equal, two groups of teachers facing the same set of accountability instruments would have similar aggregate responses even if their sociocultural contexts diverged. Any

[50] To give another example, Eleanor mentioned that Singapore had 'fifty years to progress from third world to first world', mirroring the title of a book by national patriarch Lee Kuan Yew's (2000), *From Third World to First: The Singapore Story: 1965-2000*. In her study of the Singapore government's influence over complex moral issues, Ee (2018) also found that her interview participants' language frequently mirrored government discourse. Similarly, when a group of Massachusetts educators visited Singapore in 2015, several observed that there was remarkable overlap in how actors at different levels of the education system used identical phrases to describe education plans. Beyond education, see K.P. Tan (2018) on Singapore's approach to building a national narrative and international brand more generally.

apparent resonance between the sociocultural context and teachers' responses to the accountability instruments (or their descriptions of their responses) is coincidental. Perhaps sociocultural context influenced policymakers to articulate accountability standards using terminology that resonated with dominant cultural values, so teachers' concepts of motivation simply reflect these accountability standards or politicians' rhetoric about them. Whatever the origins of this coincidence, the alternative argument being advanced is that sociocultural context is irrelevant to how teachers respond to accountability instruments.

Stated so baldly, this argument may seem implausible. Yet it is implicit in some public discussions of educational 'best practices', as described in Section 1.1, with news headlines such as 'The Japanese Education System May Solve the Problems of US Public Education' (Easterday, 2017) or 'The Secret to Finland's Success With Schools, Moms, Kids—and Everything' (Khazan, 2013).

However, the field interviews suggest that sociocultural context does influence not only teachers' responses to accountability instruments, but also other aspects of their daily work. For example, when I asked participants what they were most proud of in their work, there was a between-country difference in the orientation their responses, as shown in Table 6.6. These differences correspond to their sociocultural contexts. (As shown in the interview guide in Table 3.6, I asked this towards the beginning of the interview, and the questions about sociocultural context came much later. Thus, there is no possibility that the process of formulating descriptions of sociocultural context inadvertently influenced their framing of professional accomplishments. Instead, any resonances between the two must be due to a third element, i.e. sociocultural context itself.)

Most Finnish participants said that they were most proud of something in their own actions or abilities. For example, Juhani said:

> For some pupils, mathematics can be actually revolting. And I understand that. So I just try to give all an equal opportunity to learn it. But I don't *force* anyone to learn it. I think that's what I'm most proud of: that I can understand the views of different pupils.

This focus on one's actions and skills fits Finnish participants' concept of motivation as intrinsic, autonomous, and competence-based, within the wider sociocultural context of complementary responsibility. Singaporean participants were more likely to articulate their achievements in terms of students' responses, whether in the form of expressed gratitude or observable growth. For example, Peter said:

> I think it's those moments when I realise that, the process of walking with a student, I was able to help them see significant breakthroughs, be it in the academic subject that I'm teaching […] or, in the process of being a teacher in charge of the students' council.

This focus on student outcomes fits Singaporean participants' concept of motivation as the pursuit of desirable outcomes, within a broader emphasis on outcomes-based meritocracy.

Table 6.6 *Summary of interview participants' responses when asked what they are most proud of in their work as a teacher*

	FINLAND												SINGAPORE									
	Anneli	Liisa	Emilia	Kristiina	Masa	Satu	Maarit	Hannele	Helena	Antero	Päivi	Juhani	Adeline	Jeffrey	Maggie	Joseph	Peter	Andy	Timothy	Mark	Eleanor	Jane
Orientation																						
Own actions/abilities	●	●	●		●	●	●	●	●			●	●		●					●		
Students' responses				●						●	●			●		●	●	●	●	●	●	●
Area																						
Social/emotional/relational				●			●	●	●	●	●	●		●		●	●			●		●
Academic/pedagogical	●	●	●	●	●	●			●			●			●		●	●			●	
Organisational					●								●							●		

Note. Singaporean participants Sonia and Geok Ling do not appear in this table because time constraints prevented me from asking them this question.

(This between-country difference in how participants conceptualised their accomplishments was not accompanied by between-country differences in the areas of accomplishment that participants valued. In each country, roughly half the participants cited social, emotional, or relational accomplishments; approximately half also cited academic or pedagogical accomplishments; and one or two cited organisational accomplishments. In terms of within-country variation, there appears to have been a shift over time in Finland in that teachers who entered the profession earlier all mentioned accomplishments related to pupils' socioemotional development, whereas those who entered the profession more recently all mentioned academic or pedagogical accomplishments. This may be related to changes over time in how teachers' roles are viewed. For example, as several participants mentioned, it used to be taboo for Finnish teachers to be seen drinking alcohol in public, but this has since become acceptable; although participants noted that they still try to set a good example socially. However, since the discussion here is focusing on sociocultural differences between countries, I do not explore this within-country variation in depth.)

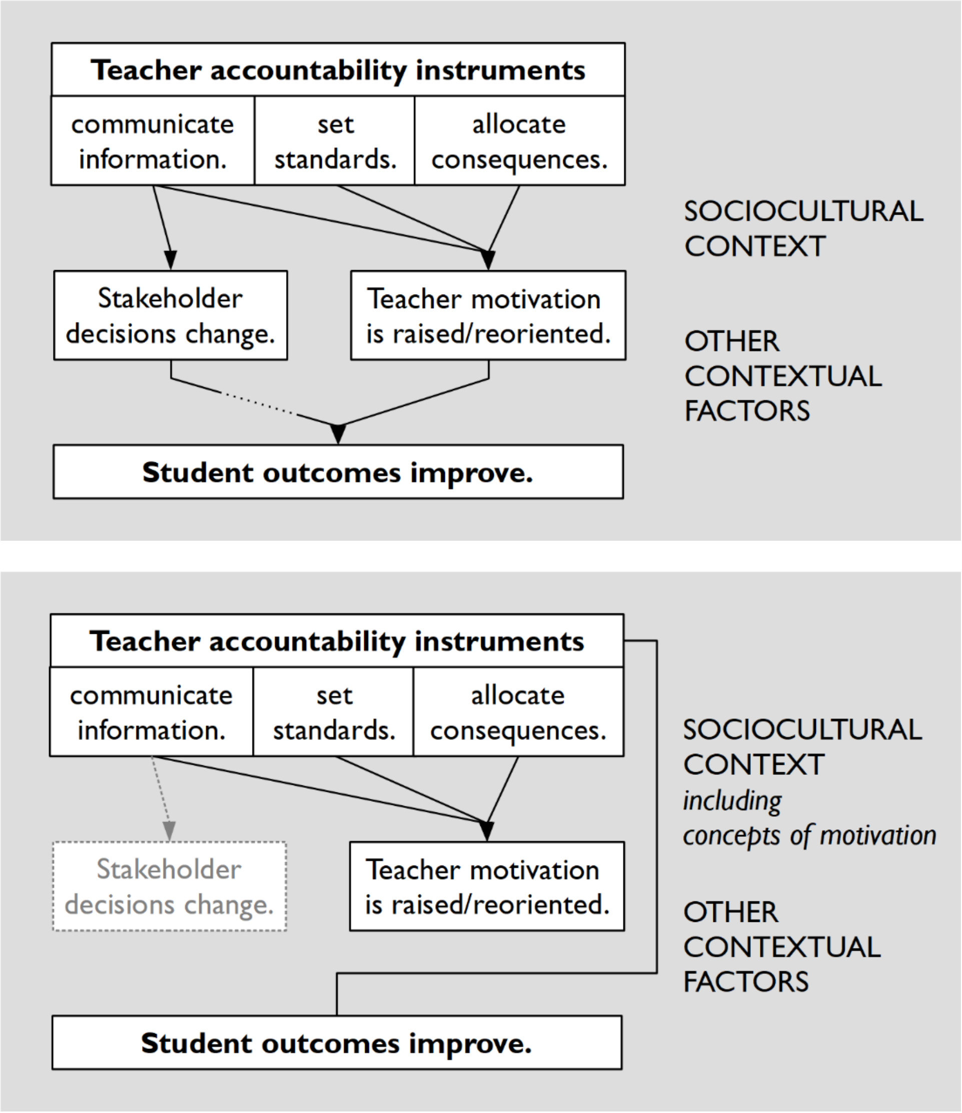
Beyond the snapshot in Table 6.6, others have found evidence that the Singaporean sociocultural context has hampered attempts to promote critical thinking (C. Tan, 2017) and differentiated instruction (Heng & Song, 2020) in schools. Additionally, the Finnish interview participant who currently works in international education noted that the main challenge in recruiting teachers for Finland-inspired schools overseas did not relate to pedagogical skills, which were ‘the easy part’. Rather, the challenge was finding qualified local teachers who were predisposed to the Finnish approach to teaching and were open to ‘changing their mindset’ even further. This emphasis on compatible mindsets adds weight to the argument that sociocultural contexts—dominant values, practices, and decision-making patterns—affect teacher practice. Given this suggestive evidence that sociocultural context influences a few different aspects of teacher practice, it is reasonable to expect that sociocultural context would also influence how teachers respond to accountability instruments, as I have argued in this chapter.

Chapter 7: Conclusion

7.1 To what extent does the evidence support the proposed conceptual framework?

In this thesis, I have explored the relationship between teacher accountability and sociocultural context. I began by developing a theory-driven framework for mapping the intended outcomes of teacher accountability instruments. I then attempted to validate this framework using cross-country survey data alongside interviews with Finnish and Singaporean teachers. Overall, the empirical data offered evidence for some aspects of the framework, but gave no evidence for others, as summarised in Figure 7.1.

Figure 7.1 *Proposed and validated conceptual frameworks for teacher accountability instruments*


Teacher accountability instruments
communicate information.
set standards.
allocate consequences.
Stakeholder decisions change.
Teacher motivation is raised/reoriented.
Student outcomes improve.
SOCIOCULTURAL CONTEXT
OTHER CONTEXTUAL FACTORS
Teacher accountability instruments
communicate information.
set standards.
allocate consequences.
Stakeholder decisions change.
Teacher motivation is raised/reoriented.
Student outcomes improve.
SOCIOCULTURAL CONTEXT
including
concepts of motivation
OTHER CONTEXTUAL FACTORS

(a) Proposed conceptual framework

(b) Elements of the framework that were validated

Most importantly, both the cross-country surveys and the teacher interviews suggest that sociocultural context affects at least some of the processes underlying teacher accountability. The ILSA analysis in Chapter 4 showed that the relationship between teacher accountability instruments and student outcomes is significantly moderated by the strength of country-level civic norms. The interview analysis in Chapter 6 found socioculturally aligned differences in how interview participants' motivation is influenced by teacher accountability instruments. These differences correspond to cross-context differences in how they understand motivation and in the factors that they regard as beneficial or detrimental to motivation. As discussed in Sections 6.3 and 6.4, I characterise these understandings as 'concepts of motivation', which are a component of *the dominant* (but not necessarily universal) *patterns of ideas and practices* that constitute the sociocultural context. Together, these sources add suggestive evidence to the case for designing accountability policies that are compatible with sociocultural context (see Section 2.3 for evidence from other studies).

Additionally, in Chapter 5, I demonstrated that teacher motivation can be a useful construct for working through the various intended and unintended teacher-level effects of accountability instruments. As emphasised throughout the thesis, raising or reorienting a teacher's motivation is only an intermediate step along the pathway to raising student outcomes. Still, it is a vital step. Teachers are an important (arguably, the most important) interface between education policy and students. I have also identified three different mechanisms by which accountability instruments can influence teacher motivation: setting standards, communicating information, and allocating consequences. As discussed theoretically in Section 2.5 and empirically in Section 5.4, these mechanisms are distinct not only in form but also in psychological function. Accordingly, any discussion of 'what works for whom in what circumstances' for teacher accountability policy would do well to consider the respective affordances and constraints of these mechanisms.

However, neither the cross-country surveys nor the teacher interviews offered evidence to substantiate the next step along the intended causal pathway: from teacher motivation to student outcomes, via changes in teaching and learning processes. To use Pawson and Tilley's (1997, p. xv) terminology, I have shown evidence for the importance of mechanisms (Chapter 5), the link between mechanisms and contexts (Chapter 6), and the link between contexts and outcomes (Chapter 4), but I do not have evidence to complete the 'context + mechanism = outcome' equation. Although I analysed cross-country survey data that had the potential to support (or weaken) the pathway from accountability instruments through teacher motivation to student outcomes (Section 5.1), these data did not give any conclusive support. Evidence for the pathway via changes in stakeholder decisions is likewise incomplete, since neither the statistical analyses nor the field interviews were designed to examine the relationship between stakeholders' decisions and student outcomes.

7.2 What I am arguing

While my empirical sources did not comprehensively validate the conceptual framework, the evidence presented in this thesis constitutes a meaningful contribution to the polarised debate on teacher accountability. (See Sections 1.1 and 2.1 for an overview of this debate.) I argue that teacher accountability policy is not a matter of whether performance bonuses are inherently good or bad, or whether test-based accountability is anathema or panacea. Every approach has advantages and disadvantages. Teacher accountability is a realm of trade-offs, which stakeholders must choose between and perhaps attempt to mitigate. (See Section 2.1 for a summary of some side effects of accountability instruments.) For example, one trade-off in Singapore's approach is that the strong alignment of goals across teachers comes at the price of high stress levels, as described in Section 5.3. One risk in the Finnish model is its dependence on intrinsic motivation, such that there are few safeguards if teacher motivation is eroded over time, a risk which some participants expressed concerns about (see, for example, Masa's account of the KIKY policy, Section 6.3).[51]

An essential part of deciding between such trade-offs is to look closely at the implementation context, as in the context-sensitive models for policy analysis discussed in Section 2.2. I argue that such analyses must also consider sociocultural patterns. Sociocultural context is complex, and it can be intractable to the standardisation and quantification that dominate some types of education discourse and economically oriented policy planning. But it can be a powerful enabler or inhibitor of teacher accountability nonetheless.

[51] More generally, relying on teachers' intrinsic motivation does little to help the students whose teachers happen to lack such motivation at a given point in time. In the words of the Finnish interview participant who is currently working in an international education organisation: 'As a teacher, […] it didn't really matter whether there were restrictions here or rewards there. Because everything was working smoothly. And I'm sure this is the case in most schools in Finland. […] But if there's a problem—for instance, as a parent, if we notice from our children that things are not right in the classroom—there's no official protocol for how this is being handled. Because we need to talk to the principal, and it's her or his responsibility to talk with the teacher. But if the principal thinks that the teacher is fine, what can we do?'

Accordingly, context in general and sociocultural context in particular should be incorporated into frameworks and theories of change for analysing and designing teacher accountability policy. Some existing frameworks for educational accountability do leave room for sociocultural context, such as in Pritchett's (2015) framework of accountability actors and relationships (outlined in Section 2.5), where the 'motivation' element is sufficiently open-ended to accommodate sociocultural variation in motivation. However, frameworks that do not explicitly bring sociocultural context into their theories of change run the risk of sidelining it.

Similarly risky are those frameworks that explicitly include context but assign it an erroneously limited role. For example, the McKinsey report *How the World's Most Improved School Systems Keep Getting Better* (Mourshed et al., 2010) does emphasise context, but mainly for its role in informing how policymakers should frame and introduce interventions in order to gain stakeholder support—and such contextual concerns are deemed 'secondary to getting the fundamentals right' (p. 11), where the fundamentals entail choosing from a menu of best practices to suit the current level of student outcomes in the education system. Contrary to the assumptions in this McKinsey report, such best practices are not context-neutral. For example, some of the interventions in their menu, such as teacher career ladders and performance-based rewards, would likely have more negative side effects than desirable effects in Finland.

In making a case for the importance of sociocultural context in teacher accountability policy, I have drawn on evidence from both cross-country surveys and teacher interviews. Neither of these empirical sources could make the case adequately on its own. Instead, they play complementary roles. The statistical analysis in Chapter 4 constitutes a highly reduced representation of the elements in my conceptual framework: teacher accountability instruments are represented by a limited selection of instruments compressed to a unidimensional scale, student outcomes are reduced to scores on a small set of cognitive proficiency tests, teacher motivation is proxied fuzzily by job satisfaction or principal-reported morale, and sociocultural context is replaced by a handful of survey-based constructs. Nonetheless, the argumentative weight of these analyses derives from showing that the same ILSA datasets that typically underpin 'best practice' arguments (as noted in Section 1.1) can equally underpin the argument that it is worthwhile to look at sociocultural context in designing teacher accountability policy. However, the ILSA data on their own only indicate vague associations: more extensive teacher accountability seems on average to work well in some sociocultural contexts, but less effectively in others. These associations do not provide a clear basis for a theory of change or for policy reform, nor for designing particular teacher accountability instruments for particular contexts.

To start understanding how and why sociocultural context may matter, I turned to the teacher interviews in Finland and Singapore. Despite drawing on a relatively small number of interviews, this analysis shows that sociocultural context can profoundly affect teachers' motivational responses to accountability instruments, at least among the interview participants and the colleagues whom they have encountered over the years. The interviews also demonstrate that what is best for teacher accountability in one context may be ineffective in another, as suggested by some studies discussed in Section 2.1. In the words of Helena, a Finnish interview participant:

> We have been getting a lot of visitors in the past years, because of the PISA results. And you can see that they come in here thinking, 'Okay, can we copy this?' And they usually leave, I think, with, 'No, we can't.'

Given the realist approach to validity in this thesis, as discussed in Section 3.1, I do not aim to claim that the observations from interviews with a non-random sample of 12 participants in each country has external validity for all Finnish lower secondary school teachers or for all Singaporean secondary school teachers. Rather, these interviews strengthen the overall validity of the argument that the effects of teacher accountability instruments are shaped by sociocultural context (among other factors). They contribute to the validity of this argument both by demonstrating, in granular firsthand narratives that are corroborated by prior studies, the influence of sociocultural context on teachers' responses to accountability instruments (in Chapters 5 and 6), and by offering suggestive evidence that weighs against certain counterarguments—that is, validity threats (Maxwell, 2017; Pawson, 2013)—to the overall argument of this thesis (in Sections 4.4, 5.5, and 6.4).

Thus, whether a given policymaker is more inclined to be persuaded by large-scale input-output analyses, or small-scale narratives of particular accountability processes, I have shown evidence suggesting that it would be prudent for them to think seriously about sociocultural context when designing teacher accountability policy.

7.3 Caveats

That said, the extent to which my empirical findings are generalisable remains an open question. In addition to the limitations of the cross-sectional ILSA data, as discussed in Section 3.6, there are nontrivial considerations about the field interviews. Firstly, even though my interviews in both countries reached saturation, they were based on relatively small non-systematic samples.

Perhaps more importantly—although I have less of a methodological or empirical basis for this concern—I suspect that Finland and Singapore may have unusually distinctive sociocultural patterns at the national level. This is not to imply that all Finns trust all other Finns, or that all Singaporeans affirm the government's brand of meritocracy. In fact, I presented evidence to the contrary in Chapter 6. Neither am I implying that Finland and Singapore are as demographically homogeneous as their media images may suggest. As noted in Section 3.4, both countries have nontrivial ethnolinguistic diversity. What I am suggesting is that the overwhelming majority of Finns would recognise the presence and, to some extent, the legitimacy of sociocultural patterns related to complementary responsibility; and that the same is true of Singaporeans and managed meritocracy (however the terminology may differ in practice). These dominant sociocultural patterns provide clear orientations for teacher accountability policymaking.

This degree of clarity is lacking in the other countries where I have lived (i.e. Malaysia, the United States, and the United Kingdom), at least in my estimation. Had I conducted fieldwork in countries other than Finland and Singapore, it may have been more difficult to trace the convergence between the sociocultural patterns affecting teachers' responses to accountability instruments and the larger sociocultural patterns affecting the general population. Nonetheless, I anticipate that in most policy jurisdictions it would be both possible and valuable to identify some recurring, socioculturally embedded patterns in teachers' responses to accountability instruments. Even where there is a multiplicity of sociocultural patterns among different subgroups of teachers, identifying and forestalling these patterns could potentially strengthen teacher accountability policymaking. For example, observational studies have found that teacher appraisals can be ineffective as accountability instruments when the appraisals are designed conducted by actors with insufficiently high social status, i.e. less-educated villagers in India (Narwana, 2015) and fellow teachers (rather than supervisors or head teachers) in Indonesia (Broekman, 2016), as noted in Section 2.3. A similar pattern was observed in a recent randomised-control trial of a social accountability mechanism in Indonesia, where committees of local residents struggled to resist pressure to award high appraisal scores to teachers when the scores were associated with financial incentives (Gaduh et al., 2020). Even though India and Indonesia are large, socioculturally plural contexts with diverse value orientations, the importance of social status is an identifiable sociocultural pattern that can and should be factored into the design of teacher accountability policy. More generally, even in a setting where it is difficult to identify shared sociocultural priorities that can positively raise and reorient teacher motivation, it may well be more feasible to identify sociocultural pitfalls that may hamper the intended influence of an off-the-shelf accountability instrument on teacher motivation in that context, at least for teachers whose concept of motivation fits the modal pattern.

Apart from these generalisability issues, there are several other caveats to my argument. Firstly, although the conceptual framework ends with 'Student outcomes improve', I am not arguing that there are specific sets of student outcomes that matter most. The desirability of different student outcomes is defined by stakeholders in a given context, often based on both sociocultural priorities and expectations of future needs. Determining which student outcomes are most important (and what constitutes valid evidence of these outcomes) should be fundamental to the design of teacher accountability policy in any education system—not least because accountability instruments that optimise for student achievement may have negative side effects for socio-emotional development (Mausethagen, 2013; Walsh, 2006; Zhao, 2017), well-being (Heller-Sahlgren, 2018), and equity (Harris & Herrington, 2006). Moreover, empirical evidence suggests that the teachers and classroom practices that are associated with higher test scores do not always overlap with the teachers and practices that develop students' socio-emotional skills, attitudes, and happiness (Blazar & Kraft, 2016; Kraft & Grace, 2016). These tensions were evident in the field interviews, wherein some Singaporean participants lamented the lack of creativity and risk-taking among their high-scoring students, and some Finnish participants worried about recent survey findings that Finnish students are not happy in school.

Additionally, I am not arguing that there is only one possible mode of effective teacher accountability for a given sociocultural context. Even in the small sample of Finnish teacher interviews, there was considerable diversity in accountability practices, such as in the degree to which colleagues and principals informally monitored teacher practice via professional collaborations. Yet these various degrees were all part of an effective system.

Neither am I arguing that there are specific, definable aspects of sociocultural context that definitively matter for teacher accountability across countries. Classroom teaching involves numerous entangled demands and interactions which may be difficult for external observers to identify (as noted in Section 2.1; see also Brown & McIntyre, 1993; Scott, 1998)—and, consequently, the sociocultural patterns that affect classroom teachi/ng may not be apparent from afar and may not be easily standardised. This was evident in the contrast between the aggregate cultural statistics from the cross-country surveys and the nuances and ambivalences mentioned by interview participants. Although the analysis in Chapter 6 indicated that teachers' implicit concepts of motivation are central to accountability, this is partly an artefact of focusing the analysis on teacher motivation to begin with. There are other ways of conceptualising the salient sociocultural patterns that my interview participants identified, as noted in Section 6.4.

Despite the focus on national-level sociocultural context in this project, other levels of context—such as schools, local communities, ethnolinguistic subgroups, and districts—also matter tremendously to the implementation of teacher accountability policy, as noted in Section 2.5. Similarly, Singaporean interview participants emphasised that teachers' experiences of accountability vary considerably depending on their line managers and school leaders. It is also important to note that cultural patterns do not divide neatly along national borders (Anderson-Levitt, 2012). I focused on national-level culture simply because I was interested in national-level differences in teacher accountability policy, and did not have the time or resources to examine other levels of sociocultural context.

Moreover, I am certainly not arguing that sociocultural compatibility is the only contextual factor influencing the efficacy of teacher accountability instruments. Numerous other contextual factors—including student academic preparation, teacher knowledge, and material resources—also matter, as noted in Section 2.1. This is indicated in the conceptual framework diagram, and suggested by the significant interaction term between teacher accountability and GDP in the student outcome regressions in Section 4.3.

I also do not mean to imply that compatibility between teacher accountability policy and contexts yields faithful policy implementation across the board. In both Finland and Singapore, there are gaps between stated accountability instruments and heterogeneously implemented reality—and between each country's idealised international image and on-the-ground experiences. As noted in Section 6.3, notwithstanding the centrality of Finland's national curriculum to its school system and teacher accountability approach, some teachers have few qualms about diverging from the curriculum when they believe they know better. Also, despite Finland's emphasis on egalitarianism and shared standards (Vainikainen et al., 2017), a few participants noted that there is significant inconsistency in student assessment grades across schools. (The teachers' union recently called for better assessment guidelines to reduce such inconsistency; Yle Uutiset, 2019.) In Singapore, notwithstanding the highly centralised and codified EPMS, participants mentioned that managers in some schools do not rely on EPMS paperwork when assigning grades to teachers, instead referring to school-specific documentation tools. Also, despite the extensive teacher performance standards, some Singaporean participants said that most of these standards were negotiable, as long as teachers dealt effectively with finances, exams, and student safety (which Jeffrey dubbed 'the Holy Trinity' and Maggie deemed to be 'cardinal rules').

Finally, I am not suggesting that teacher accountability policymaking is a straightforward, technical process. There are far too many entangled factors and processes for this to be the case. Teacher accountability is a small subset of the policy landscape and the historical trajectory of any country. Moreover, education and culture not only change over time—but they also influence each other in complex ways. This is a particular concern given the temporal flatness of both the cross-country surveys and my fieldwork. Still, interview participants noted several ongoing changes in Finland's and Singapore's educational and cultural contexts; including deep education budget cuts in Finland, increasing questioning of authority in Singapore, and growing pressure from parents and difficulties in engaging technologically preoccupied students in both countries. These trends, and their possible implications for current teacher accountability models, are reminders that teacher accountability policymaking is an open-ended, iterative process.

7.4 Directions for future research

These caveats suggest a few directions for future research on teacher accountability and sociocultural context. Firstly, the cross-country statistical analysis could be extended by looking at a wider range of student outcomes, including equity within and between schools, as well as socioemotional outcomes. Recent work using survey item response rates as a non-self-reported proxy for students' (Hitt, Trivitt, & Cheng, 2016; Zamarro, Hitt, & Mendez, 2016) may be a germane approach—although such analyses would need to take into account cross-country differences in survey response styles (e.g. He et al., 2014). Another way to extend the statistical analysis would be to examine other sociocultural constructs that would be expected to influence accountability processes, such as survey-based measures of cultural tightness and looseness (Gelfand et al., 2004, 2011) or observational measures of civic honesty from 'lost wallet' experiments (Cohn et al., 2019). Furthermore, it would be worthwhile to test the mediation model using other measures of teacher motivation—although, as noted in Section 2.4, all such standardised measures will be imperfect proxies. The statistical analysis may also be improved with more sophisticated modelling techniques, such as using Bayesian-inspired approaches to incorporate the measurement uncertainty of the survey-based aggregates for accountability and sociocultural context into the calculations.

This line of research into teacher accountability and sociocultural context would also be greatly strengthened by detailed analyses of other country contexts, including a similar interview analysis to examine teachers' perceptions and motivational responses in these other contexts. In particular, the conceptual framework could be weakened or further validated by investigations of teacher accountability in (a) education systems that have comparably strong track records but lack Finland's and Singapore's clearly defined sociocultural priorities; (b) education systems where the teacher accountability approach is similar to Finland's or Singapore's but without comparably high performance; and (c) socioculturally similar education systems with different teacher accountability approaches and student outcome configurations, such as Finland and its Nordic neighbours. For example, Camphuijsen, Møller, and Skedsmo (2019, 2020) argue that even though Norway has moved away from its former light-touch educational accountability (which was similar to Finland's current approach) toward test-based accountability, the influence of Nordic-style social democracy remains amply evident. A key impetus for Norway's adoption of test-based accountability was inequitable student outcomes, rather than the desire to compete at the top of the league tables. Also, despite the test-based approach, Norwegian teachers do not face performance-based consequences because the assumption is that the mere publication of school result comparisons will activate teachers' sense of responsibility toward the trusting public. Among these possible case selection approaches for extending the interview study, I anticipate that the greatest value for policy applications would be gained from (a), because it is unlikely that policymakers in most countries can depend on easily identifiable sociocultural orientations to guide their teacher accountability policymaking, as noted in Section 7.3, and much could be learned from how teacher accountability policy can successfully accommodate more variegated sociocultural contexts.

Finally, it may be worthwhile to conduct observational and/or longitudinal studies on the pathway from teacher accountability instruments to student outcomes via teacher motivation and classroom practices, and on how these relationships are influenced by sociocultural context. As noted above, my empirical analysis could only validate the first step along this pathway, from accountability instruments to teacher motivation, but did not find evidence for completing the pathway to student outcomes. Beyond this thesis, the educational research base includes some empirical evidence for a connection between teacher motivation and student outcomes, but this evidence base is relatively weak, as noted in Section 2.5. Nonetheless, the teacher motivation pathway is the implicit theory of change in much of teacher accountability policy. As such, this pathway warrants further study to examine both the mechanisms and pitfalls along the way.

7.5 Final reflections

In *Seeing Like a State*, James Scott (1998) describes how the scale and complexity of the modern state can compel government officials to view people, property, and resources in simplified, standardised form. While this bird's-eye view may be necessary for organising large-scale interventions, it can lead to disaster when governments forget that the real picture is more variegated and then impose policies that are dangerously reductive, as Scott documents. Such a high-level, distant view is evident in some versions of the push for best practices and evidence-based education policy. These versions assume that effective policies can always be transferred from one educational setting to another, with equal efficacy, regardless of context and history (as noted in Section 1.1; see also Feniger & Lefstein, 2014; Wiseman, 2010).

In contrast, others argue that teachers' subjective perspectives are central to any education policy reforms. For example, Waller (1932) argued that:

> The common-sense understanding which teachers have of their problems bites deeper into reality than do the maunderings of most theorists. Teachers will do well to insist that any program of educational reform shall start with them, that it shall be based upon, and shall include, their common-sense insight. (p. 457)

Without making any similarly trenchant judgements of theorists, I agree that teachers are pivotal to school improvement and that their firsthand perspectives are an important source of policy insight—as is evident in the interviews presented in Chapters 5 and 6. The centrality and complexity of such firsthand articulations is one reason why I do not attempt to determine which aspects of sociocultural context are most salient to the design of teacher accountability policy, since salience is likely to vary from context to context.

Over the course of this research project, I have come to believe that the most likely path to effective teacher accountability policy would entail policymakers (a) learning as much as possible from the experiences of other education systems, in order to identify not only the teacher accountability instruments that may work well in their context, but also the sorts of contextual factors that may enable or inhibit such instruments; and (b) speaking at length with a large number of teachers, school leaders, and frontline administrators in their context, in order to better understand their motivations, priorities, and constraints. Ideally, (a) and (b) should occur concurrently and iteratively. However, my beliefs here are hardly novel, and policymakers face limitations of time, resources, and political acceptability that are far from ideal. Be that as it may, if the discourse around teacher accountability can be shifted such that the question of 'What is our sociocultural context?' becomes as routine as 'What are high-performing countries doing?' and 'How much money do we have?', then policymakers may be more likely to design accountability instruments that prompt teachers to work harder and more effectively, with fewer negative side effects and more lasting benefits for students.

Appendix A: Minimising multicollinearity in TIMSS 2015 and TALIS 2013 regressions

As noted in under 'Modelling' in Section 3.3, the relatively small country sample sizes in the TIMSS 2015 and TALIS 2013 datasets mean that there is a risk of multicollinearity if too many country-level variables are included in a regression model. In this appendix, I demonstrate this risk using the main TIMSS 2015 dataset, and explain how I addressed it in the reported analysis. Table A.1 shows TIMSS 2015 results for model 1, which looks at the degree to which sociocultural context moderates the relationship between teacher accountability instruments and student outcomes.

In Table A.1, column (a) shows the empty model for partitioning variance across levels of analysis, and column (b) shows results for a model with all non-sociocultural predictors and their associated interaction terms. Next, column (c) has all six sociocultural predictors as well as the full complement of interaction terms, thus is analogous to model 1 for PISA 2015, as shown in column (e) of Table 4.9. This model has 15 country-level predictors: (i) the weighted mean of teacher accountability instruments ($\overline{\text{Accountability}}$), (ii) GDP per capita, (iii) six sociocultural variables, and (iv) seven interactions between teacher accountability instruments and GDP as well as the sociocultural variables. Despite the large number of cases at the pupil, teacher, and school levels, the inclusion of so many country-level predictors for a sample of only 23 countries is highly likely to constitute overfitting.

One indication that the column (c) model was overfitted—and, consequently, is affected by multicollinearity—is that standard errors of the parameter estimates for both the GDP term and $\overline{\text{Accountability}}$*GDP interaction term more than doubled in moving from column (b) to column (c). Such increases in standard error suggest collinearity issues. Additionally, the coefficient of the $\overline{\text{Accountability}}$*GDP term in column (b) was large and positive (+25.47), suggesting that wealthier countries benefit more from additional teacher accountability instruments than their less wealthy counterparts—which was consistent with direction of the corresponding interaction terms in all of the PISA 2015 and 2012 models. However, in column (c) this coefficient is large and negative (-38.10). Hence, the model with all six sociocultural constructs is likely to be unreliable for the TIMSS 2015 data. (Although Table A.1 only shows results for the main TIMSS 2015 dataset that is matched with accountability data from PISA 2015, similar changes occurred to the standard errors and coefficients of the GDP-related terms in moving between the analogous models for the TIMSS 2015 dataset matched with accountability data from PISA 2012.)

Table A.1 *Model 1: results for multilevel regressions for TIMSS 2015 for different combinations of sociocultural constructs*

Y_{ptsc} = mathematics proficiency	(a) *Variance components*	(b) *Other predictors*	(c) *All predictors*	(d) *Confidence*	(e) *Networks*	(f) *Norms*	(g) *Trust*	(h) *Power distance*	(i) *Uncertainty avoidance*
Constant	505.62**	492.29**	494.78**	485.07**	493.21**	483.63**	489.26**	498.55**	486.10**
	(12.04)	(10.70)	(10.71)	(8.58)	(10.81)	(7.70)	(9.41)	(13.44)	(8.04)
Country level									
$\overline{\text{Accountability}}_c$		7.42	-20.64	15.93	3.66	26.59	-0.34	-4.38	17.96
		(26.09)	(18.88)	(25.23)	(25.08)	(15.98)	(24.88)	(25.27)	(13.77)
GDP_c		14.04*	51.50**	22.36**	14.20*	12.74*	31.88**	19.04	23.32**
		(6.49)	(15.52)	(6.83)	(6.80)	(6.13)	(9.65)	(11.13)	(8.29)
Confidence in institutions_c			-10.49	-18.66*					
			(9.91)	(9.41)					
Civic networks_c			5.48		1.77				
			(9.28)		(9.22)				
Civic norms_c			-5.93			14.06			
			(17.72)			(8.22)			
Social trust_c			-21.61				-27.25		
			(18.85)				(14.87)		
Power distance_c			25.22					17.61	
			(23.92)					(16.64)	
Uncertainty avoidance_c			9.91						20.68*
			(24.12)						(9.78)
Interactions with $\overline{\text{Accountability}}_c$									
*GDP_c		25.47*	-38.10	8.74	28.68*	21.57*	-11.33	22.11	31.47*
		(12.70)	(29.00)	(15.51)	(11.62)	(10.55)	(17.78)	(20.31)	(15.38)
*Confidence in institutions_c			78.25**	50.15					
			(28.43)	(30.04)					
*Civic networks_c			-64.47**		-21.97				
			(17.00)		(24.95)				
*Civic norms_c			-34.87			-71.20**			
			(49.17)			(16.48)			
*Social trust_c			120.46**				102.19*		
			(34.25)				(44.93)		
*Power distance_c			-72.17					15.53	
			(63.03)					(25.53)	
*Uncertainty avoidance_c			31.81						9.75
			(40.72)						(19.69)
*Home ed resources_{ptsc}		-12.59**	-12.59**	-12.58**	-12.59**	-12.59**	-12.59**	-12.59**	-12.60**
		(3.67)	(3.66)	(3.66)	(3.67)	(3.67)	(3.67)	(3.66)	(3.66)
*Teaching experience_{tsc}		0.14	0.13	0.15	0.14	0.13	0.14	0.14	0.13
		(0.11)	(0.12)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.12)
*School resources_{sc}		-2.31	-2.30	-2.27	-2.32	-2.20	-2.30	-2.26	-2.33
		(4.24)	(4.28)	(4.25)	(4.22)	(4.23)	(4.26)	(4.23)	(4.27)

Table A.1 (continued)

		(a) *Variance components*	(b) *Other predictors*	(c) *All predictors*	(d) *Confidence*	(e) *Networks*	(f) *Norms*	(g) *Trust*	(h) *Power distance*	(i) *Uncertainty avoidance*
Pupil level										
Home educational resources$_{ptsc}$			23.48** (1.75)	23.48** (1.75)	23.48** (1.75)	23.48** (1.75)	23.48** (1.75)	23.48** (1.75)	23.48** (1.75)	23.48** (1.75)
Teacher level										
Teaching experience$_{tsc}$			0.21** (0.08)	0.20** (0.08)	0.20** (0.08)	0.21** (0.08)	0.21** (0.07)	0.20** (0.07)	0.21** (0.08)	0.21** (0.08)
School level										
School resources$_{sc}$			5.58** (1.59)	5.51** (1.60)	5.57** (1.59)	5.58** (1.59)	5.53** (1.59)	5.56** (1.59)	5.59** (1.59)	5.55** (1.60)
Variance parameters	Pupil	4 169.71	3 968.20	3 968.21	3 968.21	3 968.20	3 968.19	3 968.22	3 968.20	3 968.19
	Teacher	2 256.89	1 899.06	1 897.97	1 899.37	1 899.09	1 898.60	1 898.58	1 899.14	1 898.71
	School	701.60	416.33	417.43	415.85	416.20	416.99	416.57	416.27	416.89
	Country	3 313.15	1 266.25	415.52	1 148.80	1 234.28	821.18	1 088.52	1 017.04	913.13
-2*loglikelihood		1374531	1367627	1367602	1367625	1367627	1367617	1367624	1367622	1367620

Note. $\overline{\text{Accountability}}_c$ is taken from PISA 2015 data. N(pupils)=118 363; N(teachers)=6 147; N(schools)=3 761; N(countries)=23.
**p* < 0.05. ***p* < 0.01.

Given the likelihood that the TIMSS models with all six sociocultural constructs were overfitted, I focus instead on TIMSS 2015 models that only have one sociocultural context, entered singly in turn, as shown in columns (d) through (i) in Table A.1. (These columns correspond to the results summarised in Table 4.10 in the column for TIMSS 2015 data matched with $\overline{\text{Accountability}}$ from PISA 2015.) Each of these single-sociocultural-construct models have 5 country-level predictors, rather than 15. Looking the standard errors of the GDP main effect and the GDP interaction across columns (b) through (i), it is clear that the single-sociocultural-construct show far less indication of multicollinearity than the model with all six sociocultural constructs. This is especially true for civic norms in column (f), where the precision of the GDP-related parameter estimates improved marginally relative to column (b), as indicated by the slightly smaller standard errors. Similarly, the standard errors of these GDP-related parameters for the civic networks model in column (e) are also comparable to those in column (b). While the standard errors of these two GDP-related parameters show increasing deviation for confidence in institutions in column (d), uncertainty avoidance in column (i), social trust in column (g), and power distance in column (h), they are still far closer to the values in column (b) than is the case for the model with all six sociocultural constructs. This is especially noticeable for the $\overline{\text{Accountability}}$*GDP interaction term: while its standard error in column (b) is 12.70, and the standard error ranges between 10.55 and 20.31 for the single-sociocultural-construct models, it reaches 29.00 in the model with all six sociocultural constructs.

To summarise: firstly, the civic norms and civic networks models show no more collinearity than the model with minimal country-level predictors in column (b). Secondly, although the confidence in institutions, social trust, power distance, and uncertainty avoidance models do appear to have more collinearity issues than the model in column (b), these collinearity issues are considerably smaller than those in the model with all six sociocultural constructs. Thus, in using the TIMSS 2015 models with single sociocultural constructs as sensitivity checks for RQ1 in Chapter 4 and RQ3 in Chapter 6, I am not implying full confidence in the robustness of these estimates. Rather, I use these models because they are the best available way of addressing these research questions using the TIMSS 2015 data. (Although TIMSS 2015 is the main dataset for RQ2 in Chapter 5, this issue of country-level collinearity is less important since the main parameters of interest for RQ2 are at the pupil and school levels rather than the country level. See Section 5.1 for more details.)

Furthermore, to check whether the single-sociocultural-construct models were already overfitted, I tested a model that was similar to the civic norms model in column (f) of Table A.1, but without the main GDP parameter and the GDP interaction parameter. This reduced the number of country-level predictors from five to three. However, dropping these two parameters instead raised the standard errors of the other country-level predictors: the standard error for civic norms increased from 8.22 to 12.16; the standard error for $\overline{\text{Accountability}}$ increased from 15.98 to 31.37; and the standard error for the $\overline{\text{Accountability}}$*civic norms interaction ballooned from 16.48 to 38.10. Hence, based on these standard errors, the models with $\overline{\text{Accountability}}$, GDP, one sociocultural construct, and the interactions between $\overline{\text{Accountability}}$ and the other two country-level predictors, as shown in columns (d) through (i) of Table A.1, appear to generate the most precise country-level estimates for this dataset with this multilevel statistical approach.

In addition to the small country-level sample size, another contributing factor to the multicollinearity in the overfitted TIMSS 2015 model with all six sociocultural constructs is that GDP is correlated with some of the sociocultural constructs, as shown in Table A.2. These correlations are already evident across the larger combined set of countries, where n=70 for the WVS/EVS social capital scales, and n=66 for the Hofstede power distance and uncertainty avoidance indices. However, they are especially large in the TIMSS 2015 subset. Among the 23 countries in the TIMSS 2015 main dataset, three sociocultural variables—confidence in institutions, uncertainty avoidance, and social trust—have correlations with GDP exceeding an absolute value of 0.5. That is, the small country sample size in the TIMSS 2015 dataset not only reduces the available statistical power, but also appears to reduce the variability of the country-level predictors, as indicated by the larger correlations as compared to the set of all available observations and the PISA 2015 main dataset.

Table A.2 *Pairwise correlations (and number of countries) between sociocultural constructs and 2014 GDP per capita in different subsamples of the data*

	Confidence in institutions	Civic networks	Civic norms	Social trust	Power distance	Uncertainty avoidance
All available observations	.330** (70)	.286* (70)	.066 (70)	.430** (70)	-.452** (66)	-.307* (66)
PISA 2015 main dataset	.337* (57)	.180 (57)	.005 (57)	.447** (57)	-.546** (57)	-.349** (57)
TIMSS 2015 main dataset	.528** (23)	.389 (23)	-.117 (23)	.708** (23)	-.424* (23)	-.634** (23)

Note. Values for sociocultural constructs are from 2014 or earlier.
**p* < 0.05. ***p* < 0.01. (two-tailed)

From Table A.2, it is also worth noting that the correlation between per capita GDP and civic norms is consistently negligible. This may partly account for the stability of the standard errors for the GDP-related parameters between columns (b) and (f) of Table A.1. In addition, this may contribute to the fact that civic norms is the only robust sociocultural moderator of the association between teacher accountability and student outcomes, as shown in Table 4.10 of Section 4.3. However, correlation between country-level predictors is not the only factor underlying the degree of collinearity in the multilevel models: the power distance model in column (h) of Table A.1 appeared to have more collinearity issues than any of the other models with a single sociocultural construct, despite the correlation between GDP and power distance in the TIMSS 2015 dataset being smaller than the corresponding correlations for confidence in institutions, uncertainty avoidance, and social trust.

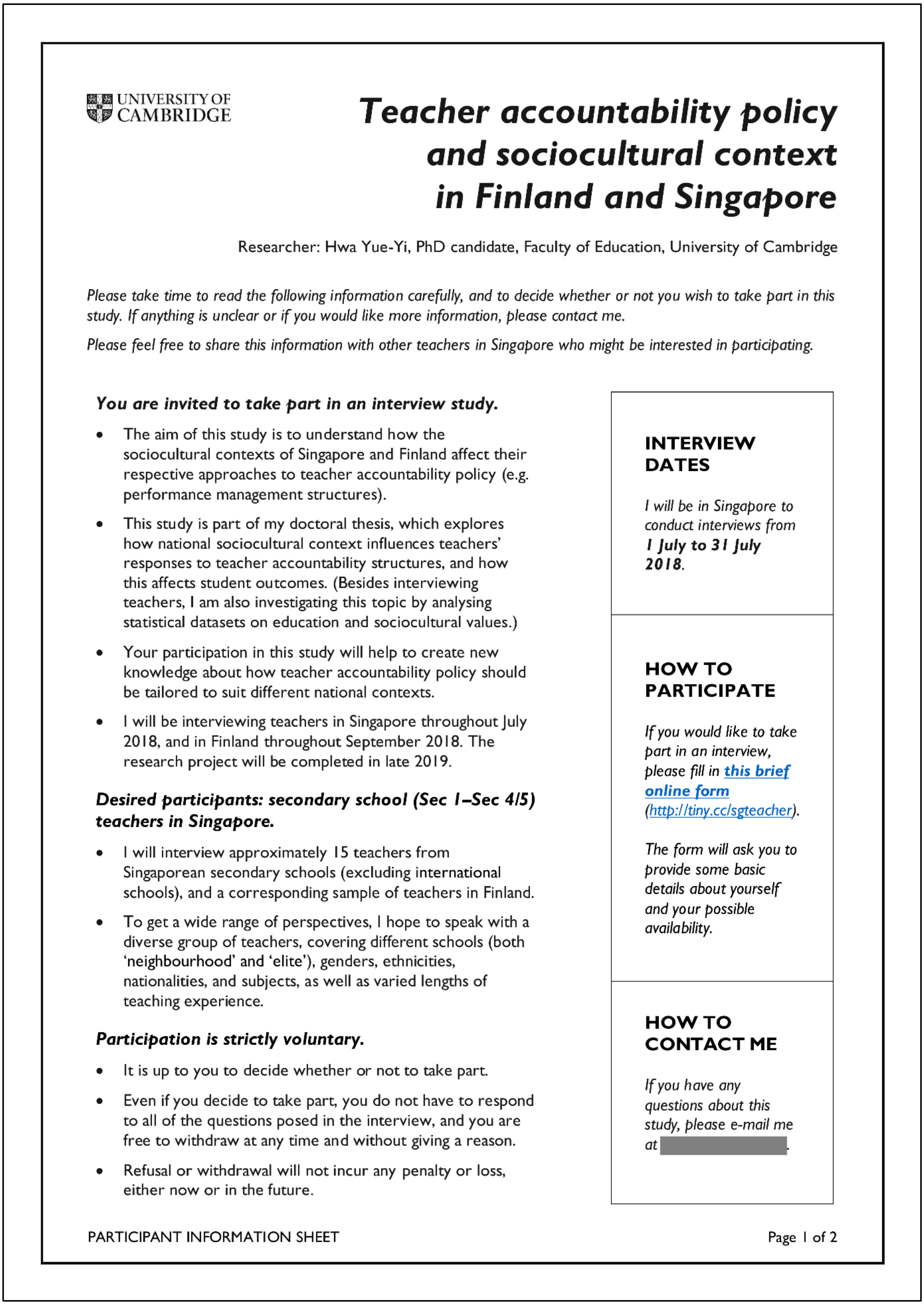
While the PISA 2015 and PISA 2012 models with all six sociocultural constructs did not appear to have similar issues with multicollinearity, such issues were evident in the TALIS 2013 analysis, which had 29 countries in the main dataset. In the TALIS 2013 model for teacher motivation without any sociocultural predictors (shown in column (b) of Table 6.1, analogous to column (b) of Table A.1 in the TIMSS 2015 discussion here), the standard errors on the GDP parameter and $\overline{\text{Accountability}}$*GDP parameter were 0.038 and 0.044, respectively. After adding in all six sociocultural constructs and their associated interaction terms, the standard error for the GDP parameter was largely unchanged at 0.040, but the standard error on the interaction term more than doubled to 0.097. (Full results from this model are not presented in the thesis text, but are available upon request.) Thus, as with the TIMSS 2015 student outcome regressions, the TALIS 2013 teacher motivation regressions appear to be overfitted when all six sociocultural constructs are included simultaneously. In the models with each sociocultural construct entered singly, the standard errors on the $\overline{\text{Accountability}}$*GDP parameter range from 0.036 (for civic norms, in column (e) of Table 6.1) to 0.083 (for social trust, in column (f) of Table 6.1). This is similar to the TIMSS 2015 single-sociocultural-construct models, where some of the models show no more country-level collinearity than the model without any sociocultural constructs; whereas some appear to have considerably more collinearity, but less so than the model with all six sociocultural constructs. Hence, under the circumstances, presenting results from the models that each have a single sociocultural construct is most balanced analytic approach.

Appendix B: Interview documentation

In the pages that follow, I present copies of the documentation that I used in conducting the field interviews. Documents are presented in the following order:

- participant information sheet for teachers in Singapore (2 pages)
- consent form for Singapore participants (1 page)
- participant information sheet for teachers in Finland (2 pages)
- consent form for Finland participants (1 page)
- interview guide printout for Singapore, used for my reference during interviews (2 pages)
- interview guide printout for Finland (2 pages)
- teacher accountability instruments definition sheet for Finland, used to facilitate conceptual clarity for any participants who may otherwise have struggled with an extended English-language definition delivered orally (1 page)

The versions of the documents included here are those that I circulated and/or printed out prior to each leg of the fieldwork. As such, they do not reflect amendments that were made along the way. In particular, as noted in the 'Designing the interview guide' subsection of Section 3.4, after the first interview in Singapore I added questions about my working hypothesis and about the WVS/EVS and Hofstede aggregate sociocultural measures. Also, after the second Singapore interview, I stopped asking the first of the two hypothetical questions (i.e. #10 in the Singapore interview guide below) because it was neither informative nor theoretically sound. These changes are reflected in the Finland interview guide below, which I updated and printed after returning to Cambridge from Singapore.



UNIVERSITY OF CAMBRIDGE

Teacher accountability policy and sociocultural context in Finland and Singapore

Researcher: Hwa Yue-Yi, PhD candidate, Faculty of Education, University of Cambridge

Please take time to read the following information carefully, and to decide whether or not you wish to take part in this study. If anything is unclear or if you would like more information, please contact me.

Please feel free to share this information with other teachers in Singapore who might be interested in participating.

You are invited to take part in an interview study.

- The aim of this study is to understand how the sociocultural contexts of Singapore and Finland affect their respective approaches to teacher accountability policy (e.g. performance management structures).
- This study is part of my doctoral thesis, which explores how national sociocultural context influences teachers' responses to teacher accountability structures, and how this affects student outcomes. (Besides interviewing teachers, I am also investigating this topic by analysing statistical datasets on education and sociocultural values.)
- Your participation in this study will help to create new knowledge about how teacher accountability policy should be tailored to suit different national contexts.
- I will be interviewing teachers in Singapore throughout July 2018, and in Finland throughout September 2018. The research project will be completed in late 2019.

Desired participants: secondary school (Sec 1–Sec 4/5) teachers in Singapore.

- I will interview approximately 15 teachers from Singaporean secondary schools (excluding international schools), and a corresponding sample of teachers in Finland.
- To get a wide range of perspectives, I hope to speak with a diverse group of teachers, covering different schools (both 'neighbourhood' and 'elite'), genders, ethnicities, nationalities, and subjects, as well as varied lengths of teaching experience.

Participation is strictly voluntary.

- It is up to you to decide whether or not to take part.
- Even if you decide to take part, you do not have to respond to all of the questions posed in the interview, and you are free to withdraw at any time and without giving a reason.
- Refusal or withdrawal will not incur any penalty or loss, either now or in the future.

INTERVIEW DATES

I will be in Singapore to conduct interviews from ***1 July to 31 July 2018****.*

HOW TO PARTICIPATE

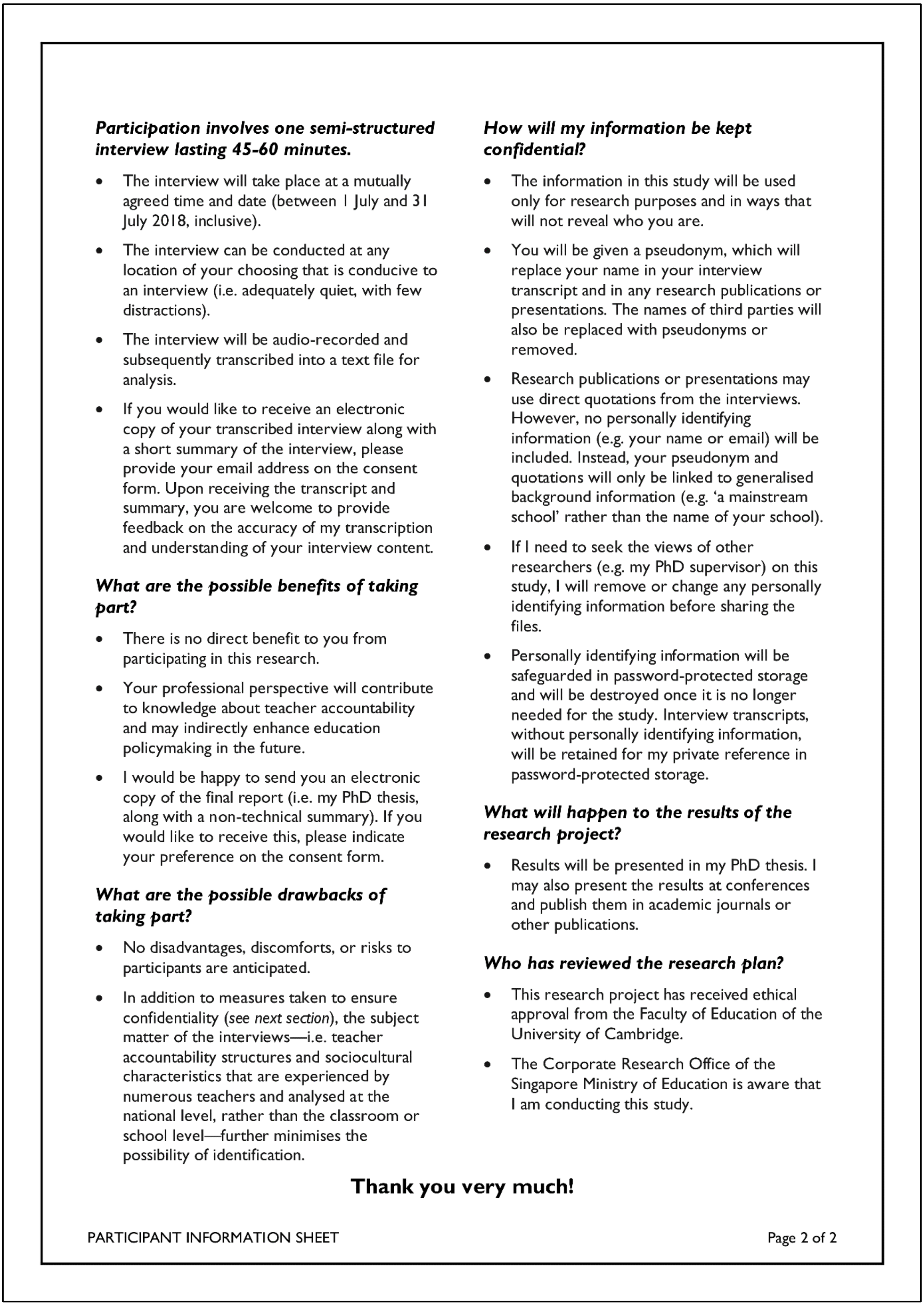
If you would like to take part in an interview, please fill in ***this brief online form*** *(http://tiny.cc/sgteacher).*

The form will ask you to provide some basic details about yourself and your possible availability.

HOW TO CONTACT ME

If you have any questions about this study, please e-mail me at [redacted].

PARTICIPANT INFORMATION SHEET Page 1 of 2



Participation involves one semi-structured interview lasting 45-60 minutes.

- The interview will take place at a mutually agreed time and date (between 1 July and 31 July 2018, inclusive).
- The interview can be conducted at any location of your choosing that is conducive to an interview (i.e. adequately quiet, with few distractions).
- The interview will be audio-recorded and subsequently transcribed into a text file for analysis.
- If you would like to receive an electronic copy of your transcribed interview along with a short summary of the interview, please provide your email address on the consent form. Upon receiving the transcript and summary, you are welcome to provide feedback on the accuracy of my transcription and understanding of your interview content.

What are the possible benefits of taking part?

- There is no direct benefit to you from participating in this research.
- Your professional perspective will contribute to knowledge about teacher accountability and may indirectly enhance education policymaking in the future.
- I would be happy to send you an electronic copy of the final report (i.e. my PhD thesis, along with a non-technical summary). If you would like to receive this, please indicate your preference on the consent form.

What are the possible drawbacks of taking part?

- No disadvantages, discomforts, or risks to participants are anticipated.
- In addition to measures taken to ensure confidentiality (*see next section*), the subject matter of the interviews—i.e. teacher accountability structures and sociocultural characteristics that are experienced by numerous teachers and analysed at the national level, rather than the classroom or school level—further minimises the possibility of identification.

How will my information be kept confidential?

- The information in this study will be used only for research purposes and in ways that will not reveal who you are.
- You will be given a pseudonym, which will replace your name in your interview transcript and in any research publications or presentations. The names of third parties will also be replaced with pseudonyms or removed.
- Research publications or presentations may use direct quotations from the interviews. However, no personally identifying information (e.g. your name or email) will be included. Instead, your pseudonym and quotations will only be linked to generalised background information (e.g. 'a mainstream school' rather than the name of your school).
- If I need to seek the views of other researchers (e.g. my PhD supervisor) on this study, I will remove or change any personally identifying information before sharing the files.
- Personally identifying information will be safeguarded in password-protected storage and will be destroyed once it is no longer needed for the study. Interview transcripts, without personally identifying information, will be retained for my private reference in password-protected storage.

What will happen to the results of the research project?

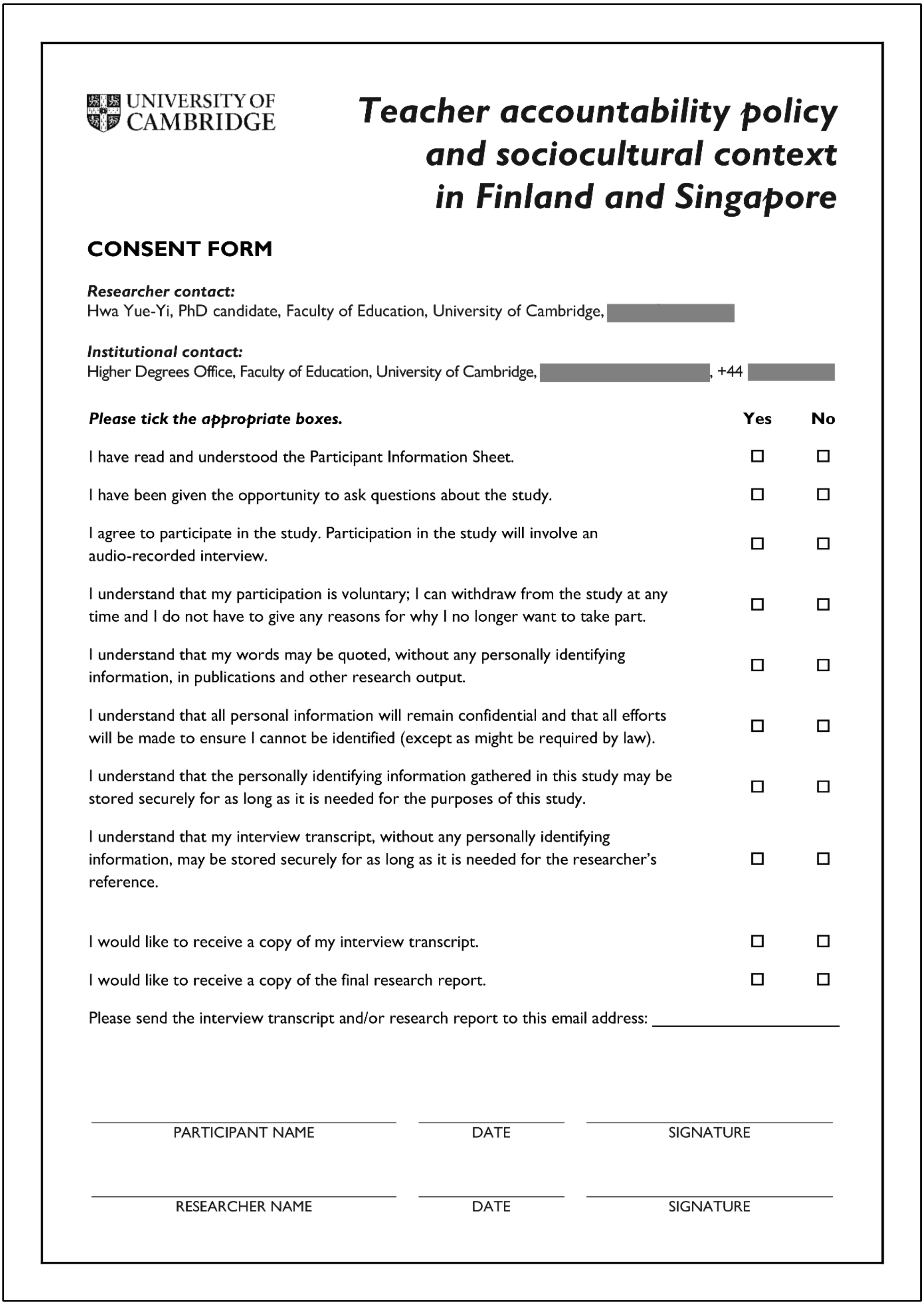
- Results will be presented in my PhD thesis. I may also present the results at conferences and publish them in academic journals or other publications.

Who has reviewed the research plan?

- This research project has received ethical approval from the Faculty of Education of the University of Cambridge.
- The Corporate Research Office of the Singapore Ministry of Education is aware that I am conducting this study.

Thank you very much!

PARTICIPANT INFORMATION SHEET Page 2 of 2



UNIVERSITY OF CAMBRIDGE

Teacher accountability policy and sociocultural context in Finland and Singapore

CONSENT FORM

Researcher contact:
Hwa Yue-Yi, PhD candidate, Faculty of Education, University of Cambridge, [redacted]

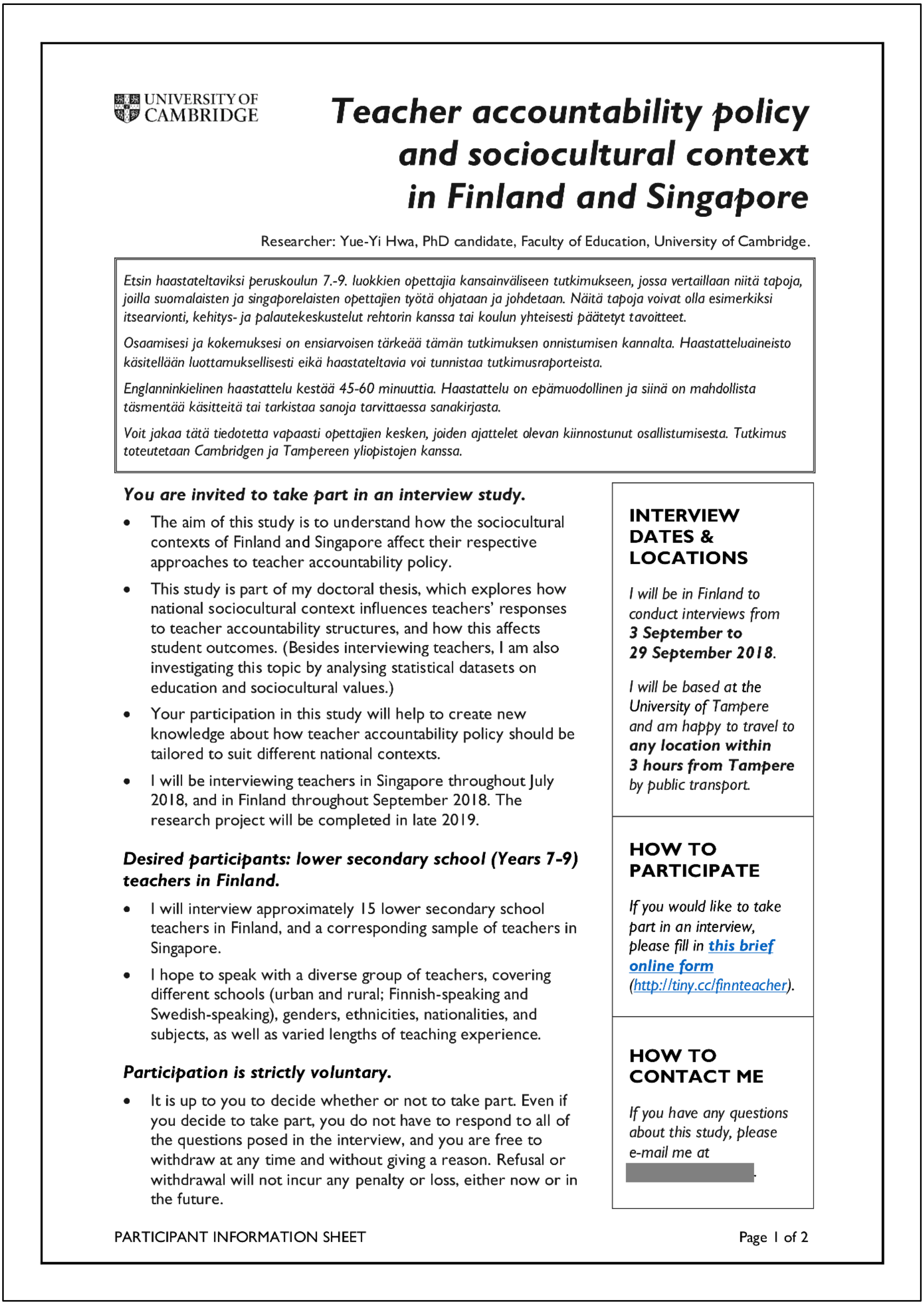
Institutional contact:
Higher Degrees Office, Faculty of Education, University of Cambridge, [redacted], +44 [redacted]

Please tick the appropriate boxes.	**Yes**	**No**
I have read and understood the Participant Information Sheet.	□	□
I have been given the opportunity to ask questions about the study.	□	□
I agree to participate in the study. Participation in the study will involve an audio-recorded interview.	□	□
I understand that my participation is voluntary; I can withdraw from the study at any time and I do not have to give any reasons for why I no longer want to take part.	□	□
I understand that my words may be quoted, without any personally identifying information, in publications and other research output.	□	□
I understand that all personal information will remain confidential and that all efforts will be made to ensure I cannot be identified (except as might be required by law).	□	□
I understand that the personally identifying information gathered in this study may be stored securely for as long as it is needed for the purposes of this study.	□	□
I understand that my interview transcript, without any personally identifying information, may be stored securely for as long as it is needed for the researcher's reference.	□	□
I would like to receive a copy of my interview transcript.	□	□
I would like to receive a copy of the final research report.	□	□

Please send the interview transcript and/or research report to this email address: ______________________

______________________ ____________ ______________________
PARTICIPANT NAME DATE SIGNATURE

______________________ ____________ ______________________
RESEARCHER NAME DATE SIGNATURE



UNIVERSITY OF CAMBRIDGE

Teacher accountability policy and sociocultural context in Finland and Singapore

Researcher: Yue-Yi Hwa, PhD candidate, Faculty of Education, University of Cambridge.

> *Etsin haastateltaviksi peruskoulun 7.-9. luokkien opettajia kansainväliseen tutkimukseen, jossa vertaillaan niitä tapoja, joilla suomalaisten ja singaporelaisten opettajien työtä ohjataan ja johdetaan. Näitä tapoja voivat olla esimerkiksi itsearvionti, kehitys- ja palautekeskustelut rehtorin kanssa tai koulun yhteisesti päätetyt tavoitteet.*
>
> *Osaamisesi ja kokemuksesi on ensiarvoisen tärkeää tämän tutkimuksen onnistumisen kannalta. Haastatteluaineisto käsitellään luottamuksellisesti eikä haastateltavia voi tunnistaa tutkimusraporteista.*
>
> *Englanninkielinen haastattelu kestää 45-60 minuuttia. Haastattelu on epämuodollinen ja siinä on mahdollista täsmentää käsitteitä tai tarkistaa sanoja tarvittaessa sanakirjasta.*
>
> *Voit jakaa tätä tiedotetta vapaasti opettajien kesken, joiden ajattelet olevan kiinnostunut osallistumisesta. Tutkimus toteutetaan Cambridgen ja Tampereen yliopistojen kanssa.*

You are invited to take part in an interview study.

- The aim of this study is to understand how the sociocultural contexts of Finland and Singapore affect their respective approaches to teacher accountability policy.
- This study is part of my doctoral thesis, which explores how national sociocultural context influences teachers' responses to teacher accountability structures, and how this affects student outcomes. (Besides interviewing teachers, I am also investigating this topic by analysing statistical datasets on education and sociocultural values.)
- Your participation in this study will help to create new knowledge about how teacher accountability policy should be tailored to suit different national contexts.
- I will be interviewing teachers in Singapore throughout July 2018, and in Finland throughout September 2018. The research project will be completed in late 2019.

Desired participants: lower secondary school (Years 7-9) teachers in Finland.

- I will interview approximately 15 lower secondary school teachers in Finland, and a corresponding sample of teachers in Singapore.
- I hope to speak with a diverse group of teachers, covering different schools (urban and rural; Finnish-speaking and Swedish-speaking), genders, ethnicities, nationalities, and subjects, as well as varied lengths of teaching experience.

Participation is strictly voluntary.

- It is up to you to decide whether or not to take part. Even if you decide to take part, you do not have to respond to all of the questions posed in the interview, and you are free to withdraw at any time and without giving a reason. Refusal or withdrawal will not incur any penalty or loss, either now or in the future.

INTERVIEW DATES & LOCATIONS

I will be in Finland to conduct interviews from ***3 September to 29 September 2018****.*

I will be based at the University of Tampere and am happy to travel to ***any location within 3 hours from Tampere*** *by public transport.*

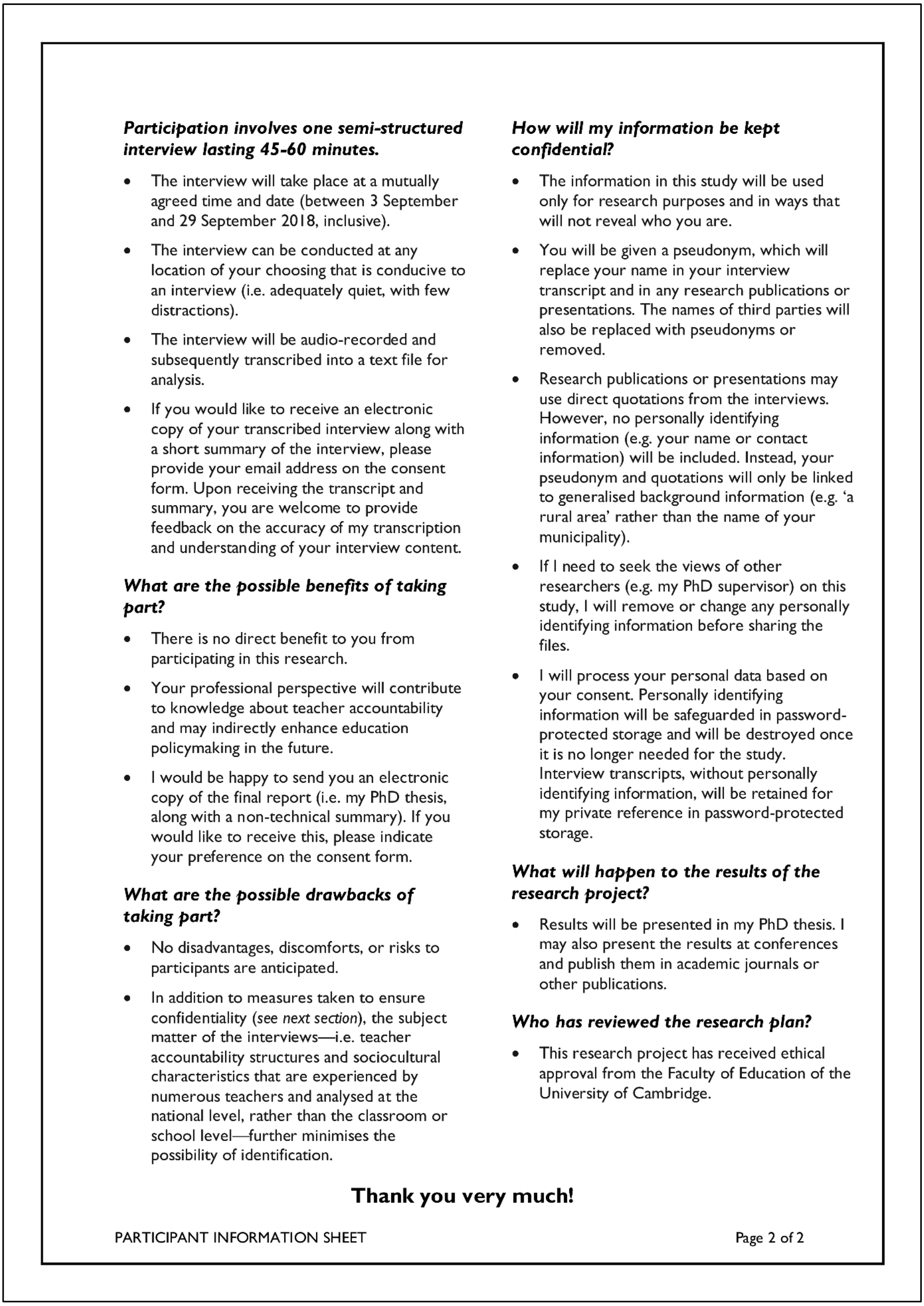
HOW TO PARTICIPATE

If you would like to take part in an interview, please fill in ***this brief online form*** *(http://tiny.cc/finnteacher).*

HOW TO CONTACT ME

If you have any questions about this study, please e-mail me at [redacted].

PARTICIPANT INFORMATION SHEET Page 1 of 2



Participation involves one semi-structured interview lasting 45-60 minutes.

- The interview will take place at a mutually agreed time and date (between 3 September and 29 September 2018, inclusive).
- The interview can be conducted at any location of your choosing that is conducive to an interview (i.e. adequately quiet, with few distractions).
- The interview will be audio-recorded and subsequently transcribed into a text file for analysis.
- If you would like to receive an electronic copy of your transcribed interview along with a short summary of the interview, please provide your email address on the consent form. Upon receiving the transcript and summary, you are welcome to provide feedback on the accuracy of my transcription and understanding of your interview content.

What are the possible benefits of taking part?

- There is no direct benefit to you from participating in this research.
- Your professional perspective will contribute to knowledge about teacher accountability and may indirectly enhance education policymaking in the future.
- I would be happy to send you an electronic copy of the final report (i.e. my PhD thesis, along with a non-technical summary). If you would like to receive this, please indicate your preference on the consent form.

What are the possible drawbacks of taking part?

- No disadvantages, discomforts, or risks to participants are anticipated.
- In addition to measures taken to ensure confidentiality (*see next section*), the subject matter of the interviews—i.e. teacher accountability structures and sociocultural characteristics that are experienced by numerous teachers and analysed at the national level, rather than the classroom or school level—further minimises the possibility of identification.

How will my information be kept confidential?

- The information in this study will be used only for research purposes and in ways that will not reveal who you are.
- You will be given a pseudonym, which will replace your name in your interview transcript and in any research publications or presentations. The names of third parties will also be replaced with pseudonyms or removed.
- Research publications or presentations may use direct quotations from the interviews. However, no personally identifying information (e.g. your name or contact information) will be included. Instead, your pseudonym and quotations will only be linked to generalised background information (e.g. 'a rural area' rather than the name of your municipality).
- If I need to seek the views of other researchers (e.g. my PhD supervisor) on this study, I will remove or change any personally identifying information before sharing the files.
- I will process your personal data based on your consent. Personally identifying information will be safeguarded in password-protected storage and will be destroyed once it is no longer needed for the study. Interview transcripts, without personally identifying information, will be retained for my private reference in password-protected storage.

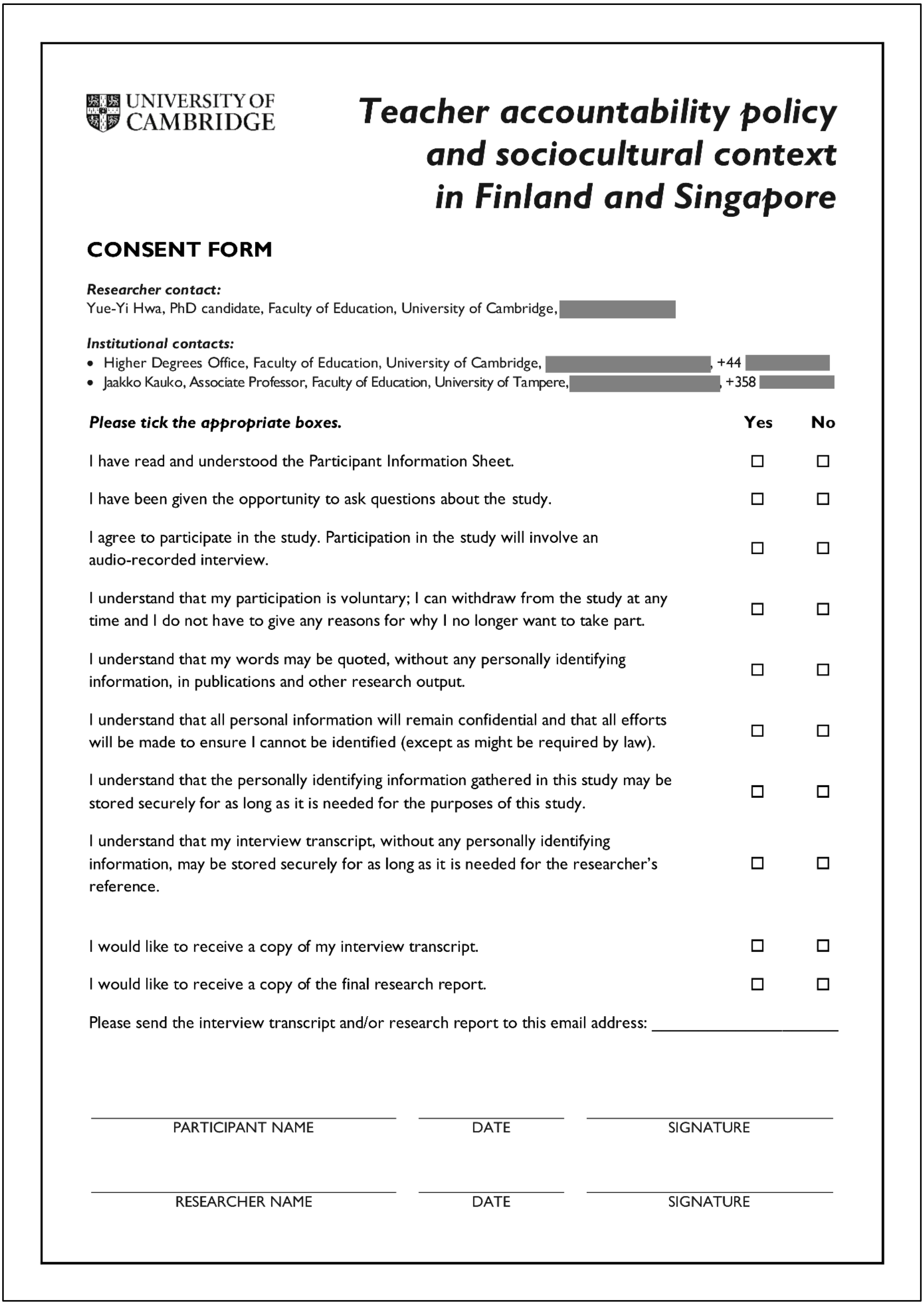
What will happen to the results of the research project?

- Results will be presented in my PhD thesis. I may also present the results at conferences and publish them in academic journals or other publications.

Who has reviewed the research plan?

- This research project has received ethical approval from the Faculty of Education of the University of Cambridge.

Thank you very much!



UNIVERSITY OF CAMBRIDGE

Teacher accountability policy and sociocultural context in Finland and Singapore

CONSENT FORM

Researcher contact:
Yue-Yi Hwa, PhD candidate, Faculty of Education, University of Cambridge, [redacted]

Institutional contacts:

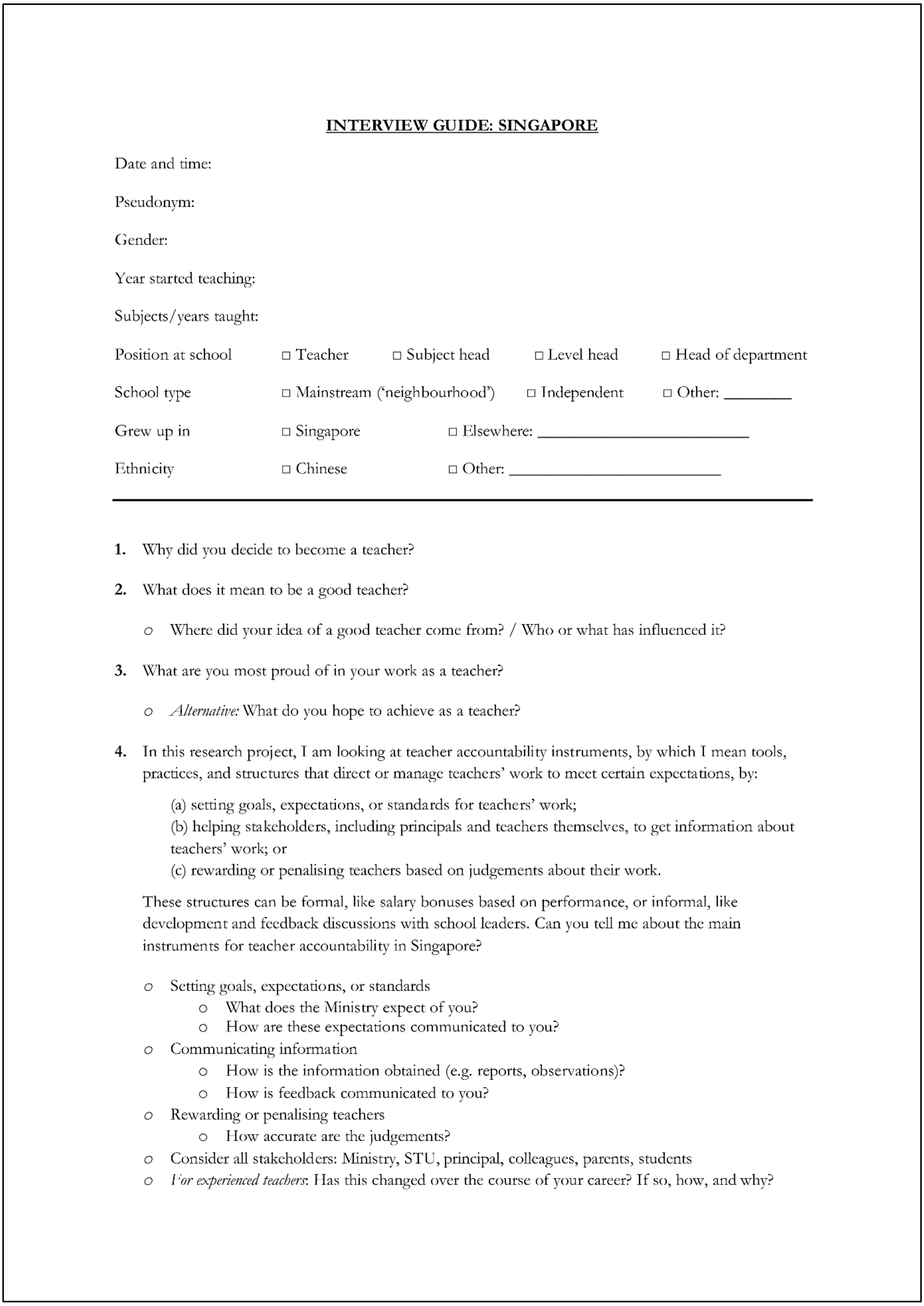
- Higher Degrees Office, Faculty of Education, University of Cambridge, [redacted], +44 [redacted]
- Jaakko Kauko, Associate Professor, Faculty of Education, University of Tampere, [redacted], +358 [redacted]

Please tick the appropriate boxes.	**Yes**	**No**
I have read and understood the Participant Information Sheet.	□	□
I have been given the opportunity to ask questions about the study.	□	□
I agree to participate in the study. Participation in the study will involve an audio-recorded interview.	□	□
I understand that my participation is voluntary; I can withdraw from the study at any time and I do not have to give any reasons for why I no longer want to take part.	□	□
I understand that my words may be quoted, without any personally identifying information, in publications and other research output.	□	□
I understand that all personal information will remain confidential and that all efforts will be made to ensure I cannot be identified (except as might be required by law).	□	□
I understand that the personally identifying information gathered in this study may be stored securely for as long as it is needed for the purposes of this study.	□	□
I understand that my interview transcript, without any personally identifying information, may be stored securely for as long as it is needed for the researcher's reference.	□	□
I would like to receive a copy of my interview transcript.	□	□
I would like to receive a copy of the final research report.	□	□

Please send the interview transcript and/or research report to this email address: ______________________

______________________ ____________ ____________________
PARTICIPANT NAME DATE SIGNATURE

______________________ ____________ ____________________
RESEARCHER NAME DATE SIGNATURE



INTERVIEW GUIDE: SINGAPORE

Date and time:

Pseudonym:

Gender:

Year started teaching:

Subjects/years taught:

Position at school □ Teacher □ Subject head □ Level head □ Head of department

School type □ Mainstream ('neighbourhood') □ Independent □ Other: ________

Grew up in □ Singapore □ Elsewhere: __________________________

Ethnicity □ Chinese □ Other: __________________________

---

1. Why did you decide to become a teacher?

2. What does it mean to be a good teacher?

- Where did your idea of a good teacher come from? / Who or what has influenced it?

3. What are you most proud of in your work as a teacher?

- *Alternative:* What do you hope to achieve as a teacher?

4. In this research project, I am looking at teacher accountability instruments, by which I mean tools, practices, and structures that direct or manage teachers' work to meet certain expectations, by:

(a) setting goals, expectations, or standards for teachers' work;
(b) helping stakeholders, including principals and teachers themselves, to get information about teachers' work; or
(c) rewarding or penalising teachers based on judgements about their work.

These structures can be formal, like salary bonuses based on performance, or informal, like development and feedback discussions with school leaders. Can you tell me about the main instruments for teacher accountability in Singapore?

- Setting goals, expectations, or standards
  - What does the Ministry expect of you?
  - How are these expectations communicated to you?
- Communicating information
  - How is the information obtained (e.g. reports, observations)?
  - How is feedback communicated to you?
- Rewarding or penalising teachers
  - How accurate are the judgements?
- Consider all stakeholders: Ministry, STU, principal, colleagues, parents, students
- *For experienced teachers*: Has this changed over the course of your career? If so, how, and why?

Reiterate: For the rest of this interview, when I say 'accountability instruments', I'm referring to these tools, practices, and structures that you have mentioned.

5. How do these accountability instruments affect your work?

- how you teach and interact with students
- which aspects of your work you prioritise
- your workload
- your autonomy
- how effectively you can meet your own and others' expectations of you
- your motivation and well-being.

6. In what ways do these accountability instruments make it easier or harder for you to be a good teacher?

- *Prompt: for both positives and negatives*

7. In what ways do these accountability instruments help Singapore's education system be more effective in developing students' capabilities?

8. In this research project, I am interested in how society and culture influence education. Can you tell me about aspects of Singaporean culture that affect Singapore's education system?

- *Example: emphasis in Malaysian culture on respecting elders and people in authority, and on 'saving face'*

9. In what ways does Singaporean culture affect how teachers respond to accountability instruments?

- *Example: Malaysian teachers exaggerating their performance in official reports, to protect the school's good name*

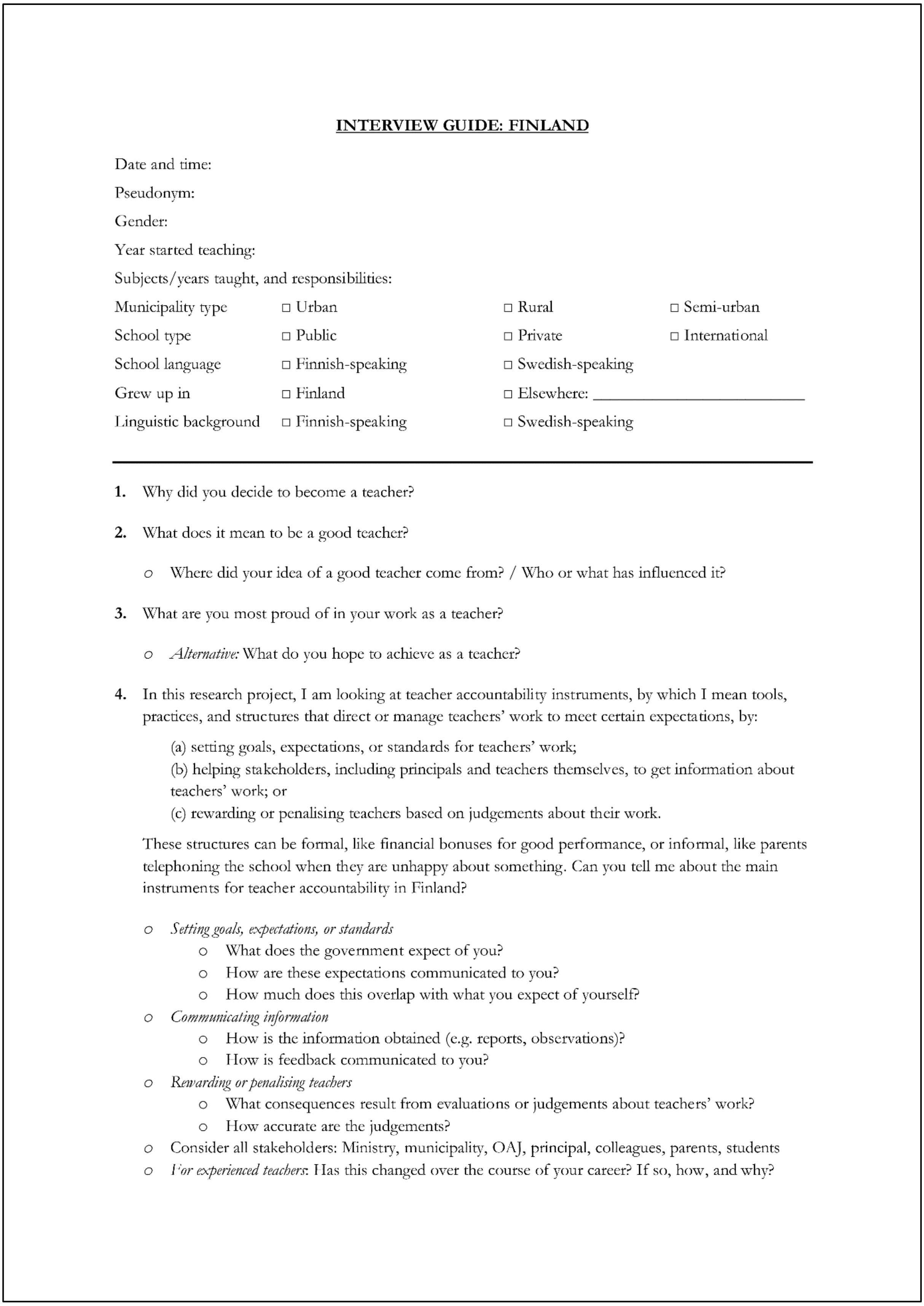
10. Imagine if Singapore took away all teacher accountability instruments. How would you react, and how would this affect your work?

- How would most other Singaporean teachers respond?
- How do you think this would affect the effectiveness of the education system?

11. Suppose Singapore were to adopt Finland's teacher accountability instruments. In Finland, teachers do not undergo formal evaluation. However, because there is no formal evaluation, and because the teaching career structure is flat, high-performing teachers are not rewarded with promotion. How would you react, and how would this affect your work?

- How would most other Singaporean teachers respond?
- How do you think this would affect the effectiveness of the education system?

12. Is there anything else that you would like to mention? In particular, are there any other important aspects of teacher accountability or sociocultural context in Singapore—or any connections between these things—that we have not discussed yet?



INTERVIEW GUIDE: FINLAND

Date and time:

Pseudonym:

Gender:

Year started teaching:

Subjects/years taught, and responsibilities:

Municipality type	□ Urban	□ Rural	□ Semi-urban
School type	□ Public	□ Private	□ International
School language	□ Finnish-speaking	□ Swedish-speaking	
Grew up in	□ Finland	□ Elsewhere: ___________________________	
Linguistic background	□ Finnish-speaking	□ Swedish-speaking	

---

1. Why did you decide to become a teacher?

2. What does it mean to be a good teacher?

- Where did your idea of a good teacher come from? / Who or what has influenced it?

3. What are you most proud of in your work as a teacher?

- *Alternative:* What do you hope to achieve as a teacher?

4. In this research project, I am looking at teacher accountability instruments, by which I mean tools, practices, and structures that direct or manage teachers' work to meet certain expectations, by:

(a) setting goals, expectations, or standards for teachers' work;
(b) helping stakeholders, including principals and teachers themselves, to get information about teachers' work; or
(c) rewarding or penalising teachers based on judgements about their work.

These structures can be formal, like financial bonuses for good performance, or informal, like parents telephoning the school when they are unhappy about something. Can you tell me about the main instruments for teacher accountability in Finland?

- *Setting goals, expectations, or standards*
    - What does the government expect of you?
    - How are these expectations communicated to you?
    - How much does this overlap with what you expect of yourself?
- *Communicating information*
    - How is the information obtained (e.g. reports, observations)?
    - How is feedback communicated to you?
- *Rewarding or penalising teachers*
    - What consequences result from evaluations or judgements about teachers' work?
    - How accurate are the judgements?
- Consider all stakeholders: Ministry, municipality, OAJ, principal, colleagues, parents, students
- *For experienced teachers*: Has this changed over the course of your career? If so, how, and why?

5. How do these accountability instruments affect your work?

   - how you teach and interact with students; which aspects of your work you prioritise; your workload; your autonomy; your motivation

6. On the whole, do these accountability instruments make it easier or harder for you to be a good teacher?

   - *Prompt: for both positives and negatives*

7. In what ways do these accountability instruments help Finland's education system be more effective in developing students' capabilities?

8. In this research project, I am interested in how society and culture influence education. Can you tell me about aspects of Finnish culture that affect Finland's education system?

9. In what ways does Finnish culture affect how teachers respond to accountability instruments?

10. Imagine if Finland were to adopt Singapore's teacher accountability instruments. In Singapore, every teacher is formally evaluated twice a year by a supervising teacher, based on many different aspects of their professional work. Every teacher has a specific level of responsibility on the career ladder, and teachers are given grades based on how their performance compares to other teachers of the same level. These grades determine teachers' annual bonuses and affect their promotion through the teacher career ladder. How would you react, and how would this affect your work?

    - How would most other Finnish teachers respond?
    - How do you think this would affect the effectiveness of the education system?

11. Next, I would like to ask for your thoughts about some aspects of my research project. In the wake of international rankings from tests like PISA and TIMSS, some people say, "Finland has such good schools; let's copy their policies," or, "Singapore does so well; let's copy their policies." Based on my research so far, it seems unlikely that one country could copy another country's policies wholesale and expect the same outcomes. My current hypothesis is that effective education systems have teacher accountability instruments that are compatible with their sociocultural context. From your perspective, to what extent is this hypothesis plausible?

12. In my statistical analysis, besides using educational datasets like PISA and TIMSS, I also used sociocultural datasets like the World Values Survey. I would like to ask your opinion on how accurately these datasets describe Finnish people. Would you agree that most Finns you know:

    (a) trust other people to act fairly?
    (b) prefer equal distributions of power rather than unequal hierarchies?
    (c) think that you should follow the official system rather than trying to find loopholes or take shortcuts for your own benefit?

13. Is there anything else that you would like to mention? In particular, are there any other important aspects of teacher accountability or sociocultural context in Finland—or any connections between these things—that we have not discussed yet?

<u>Teacher accountability instruments</u>

- **niitä tapoja, joilla suomalaisten ja singaporelaisten opettajien työtä ohjataan ja johdetaan. Näitä tapoja voivat olla esimerkiksi itsearvionti, kehitys- ja palautekeskustelut rehtorin kanssa tai koulun yhteisesti päätetyt tavoitteet.**
- are tools, practices, and structures that direct or manage teachers' work to meet certain expectations, by:
  - i. *setting goals, expectations, or standards for teachers' work;*
  - ii. *collecting information about teachers' work and communicating that information to stakeholders (including the Ministry of Education and Culture, kunta/kaupunki, principal, other teachers, parents, students, OAJ); or*
  - iii. *rewarding or penalising teachers based on evaluations of their work.*
- These structures can be formal, like financial bonuses for good performance, or informal, like parents telephoning the school when they are unhappy about something.

References

Ab Kadir, M. A. (2017). Engendering a culture of thinking in a culture of performativity: The challenge of mediating tensions in the Singaporean educational system. *Cambridge Journal of Education*, *47*(2), 227–246. https://doi.org/10.1080/0305764X.2016.1148115

Abazaoglu, I., & Aztekin, S. (2016). The Role of Teacher Morale and Motivation on Students' Science and Math Achievement: Findings from Singapore, Japan, Finland and Turkey. *Universal Journal of Educational Research*, *4*(11), 2606–2617.

Abelmann, C., Elmore, R. F., Even, J., Kenyon, S., & Marshall, J. (1999). *When Accountability Knocks, Will Anyone Answer?* (CPRE Research Report Series No. RR-42). Retrieved from Consortium for Policy Research in Education website: http://www.cpre.org/sites/default/files/researchreport/782_rr42.pdf

Abramson, C. M., & Dohan, D. (2015). Beyond Text: Using Arrays to Represent and Analyze Ethnographic Data. *Sociological Methodology*, *45*(1), 272–319. https://doi.org/10.1177/0081175015578740

Adams, G., & Markus, H. R. (2001). Culture As Patterns: An Alternative Approach to the Problem of Reification. *Culture & Psychology*, *7*(3), 283–296. https://doi.org/10.1177/1354067X0173002

Adkins, G. K. (2004). *Teacher performance pay: The perceptions of certified school-based personnel* (PhD Thesis). University of Central Florida.

Adnot, M., Dee, T. S., Katz, V., & Wyckoff, J. (2017). Teacher Turnover, Teacher Quality, and Student Achievement in DCPS. *Educational Evaluation and Policy Analysis*, *39*(1), 54–76. https://doi.org/10.3102/0162373716663646

Ahmad, S. L. bin. (2016). 'The looking glass self'—Locating my professional identity from interactions with students and colleagues. In Y. Fang (Ed.), *Singapore teachers: Narratives of care, hope and commitment* (pp. 101–116). New Jersey: WS Education.

Aho, E., Pitkänen, K., & Sahlberg, P. (2006). *Policy development and reform principles of basic and secondary education in Finland since 1968* (No. 36871). Retrieved from World Bank website: http://documents.worldbank.org/curated/en/124381468038093074/Policy-development-and-reform-principles-of-basic-and-secondary-education-in-Finland-since-1968

Akbar, Y. H., & Vujić, V. (2014). Explaining corruption: The role of national culture and its implications for international management. *Cross Cultural Management: An International Journal*, *21*(2), 191–218. https://doi.org/10.1108/CCM-03-2013-0050

Alesina, A., & Giuliano, P. (2015). Culture and Institutions. *Journal of Economic Literature*, *53*(4), 898–944. https://doi.org/10.1257/jel.53.4.898

Alexander, R. J. (2001). *Culture and pedagogy: International comparisons in primary education*. Malden, Massachusetts: Blackwell Pub.

Allen, D. (2016). *Education and Equality*. University of Chicago Press.

Altrichter, H., & Kemethofer, D. (2015). Does accountability pressure through school inspections promote school improvement? *School Effectiveness and School Improvement*, *26*(1), 32–56. https://doi.org/10.1080/09243453.2014.927369

Altschuler, D. (2013). How patronage politics undermines parental participation and accountability: Community-managed schools in Honduras and Guatemala. *Comparative Education Review*, *57*(1), 117–144. https://doi.org/10.1086/667963

American Psychological Association. (n.d.). Morale. In *APA Dictionary of Psychology*. Retrieved 15 April 2020, from https://dictionary.apa.org/morale

Andere, E. (2014). *Teachers' Perspectives on Finnish School Education: Creating Learning Environments*. Springer International Publishing.

Anderson-Levitt, K. M. (2012). Complicating the concept of culture. *Comparative Education*, *48*(4), 441–454. https://doi.org/10.1080/03050068.2011.634285

Andrabi, T., Das, J., & Khwaja, A. I. (2017). Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets. *American Economic Review*, *107*(6), 1535–1563. https://doi.org/10.1257/aer.20140774

Andrews, M., Pritchett, L., & Woolcock, M. (2017). *Building State Capability: Evidence, Analysis, Action*. Oxford, New York: Oxford University Press.

Ang, J. C. C. (2016). Leaders that make or break your career. In Y. Fang (Ed.), *Singapore teachers: Narratives of care, hope and commitment* (pp. 207–218). New Jersey: WS Education.

Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review*, *64*(6, Pt.1), 359–372. https://doi.org/10.1037/h0043445

Atuhurra, J., & Kaffenberger, M. (2019, May 23). System (In)Coherence Seen through a Curriculum Lens: Ugandan Teachers Face Conflicting Demands from Curriculum and Examination Bodies [blog]. Retrieved from https://www.riseprogramme.org/blog/system_incoherence_curriculum

Ball, S. J., Maguire, M., & Braun, A. (2012). *How Schools Do Policy: Policy Enactments in Secondary Schools*. Abingdon, Oxfordshire; New York: Routledge.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Barber, M., & Mourshed, M. (2007). *How the world's best-performing schools come out on top*. Retrieved from McKinsey & Company website: http://mckinseyonsociety.com/how-the-worlds-best-performing-schools-come-out-on-top/

Barber, M., Rodriguez, N., & Artis, E. (2016). *Deliverology in practice: How education leaders are improving student outcomes*. Thousand Oaks, California: Corwin.

Barr, M. D., & Skrbiš, Z. (2008). *Constructing Singapore: Elitism, ethnicity and the Nation-Building Project*. Copenhagen: NIAS-Press.

Barrett, A. M. (2005). Teacher accountability in context: Tanzanian primary school teachers' perceptions of local community and education administration. *Compare*, *35*(1), 43–61. https://doi.org/10.1080/03057920500033530

Bassok, D., Latham, S., & Rorem, A. (2016). Is Kindergarten the New First Grade? *AERA Open*, *2*(1), 2332858415616358. https://doi.org/10.1177/2332858415616358

Bates, M. A., & Glennerster, R. (2017, Summer). The Generalizability Puzzle (SSIR). *Stanford Social Innovation Review*, *15*(3). Retrieved from https://ssir.org/articles/entry/the_generalizability_puzzle

Bell, A., & Jones, K. (2015). Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data. *Political Science Research and Methods*, *3*(1), 133–153. https://doi.org/10.1017/psrm.2014.7

Benoliel, P., & Berkovich, I. (2018). A Cross-National Examination of the Effect of the Schwartz Cultural Dimensions on PISA Performance Assessments. *Social Indicators Research*, *139*(2), 825–845. https://doi.org/10.1007/s11205-017-1732-z

BERA. (2011). *Ethical Guidelines for Educational Research*. London: British Educational Research Association.

Bergbauer, A. B., Hanushek, E. A., & Woessmann, L. (2018). *Testing* (NBER Working Paper No. 24836). https://doi.org/10.3386/w24836

Berggren, H., & Trägårdh, L. (2010). Social trust and radical individualism: The paradox at the heart of Nordic capitalism. In *Shared Norms for the New Reality: The Nordic Way* (pp. 13–27). Retrieved from https://www.globalutmaning.se/wp-content/uploads/sites/8/2011/01/Davos-The-nordic-way-final.pdf

Berry, J. W. (Ed.). (2011). *Cross-cultural psychology: Research and applications* (3rd ed). Cambridge; New York: Cambridge University Press.

Besley, T. (2007). *Principled Agents? The Political Economy of Good Government*. Oxford: Oxford University Press.

Bevir, M. (2007). *Encyclopedia of governance*. Thousand Oaks: Sage Publications.

Bhaskar, R. (1993). *Dialectic: The pulse of freedom*. London ; New York: Verso.

Biasi, B. (2018). *The Labor Market for Teachers Under Different Pay Schemes* (Working Paper No. 24813). https://doi.org/10.3386/w24813

Biesta, G. J. J. (2011). *Good Education in an Age of Measurement: Ethics, Politics, Democracy*. Paradigm Publishers.

Bjork, C. (2016). *High-stakes schooling: What we can learn from Japan's experiences with testing, accountability, and education reform*. Chicago: The University of Chicago Press.

Blazar, D., & Kraft, M. A. (2016). Teacher and Teaching Effects on Students' Attitudes and Behaviors. *Educational Evaluation and Policy Analysis*, 0162373716670260. https://doi.org/10.3102/0162373716670260

Bleich, E., & Pekkanen, R. (2013). How to report interview data. In L. Mosley (Ed.), *Interview research in political science* (pp. 84–105). Ithaca: Cornell University Press.

Booher-Jennings, J. (2005). Below the Bubble: "Educational Triage" and the Texas Accountability System. *American Educational Research Journal*, *42*(2), 231–268. https://doi.org/10.3102/00028312042002231

Booth, M. (2014). *The almost nearly perfect people: The truth about the Nordic miracle*. London: Jonathan Cape.

Bourdieu, P. (1986). The forms of capital. In J. G. Richardson (Ed.), *Handbook of Theory and Research for the Sociology of Education* (pp. 241–258). New York: Greenwood.

Bovens, M. (2007). Analysing and Assessing Accountability: A Conceptual Framework. *European Law Journal*, *13*(4), 447–468. https://doi.org/10.1111/j.1468-0386.2007.00378.x

Bovens, M., Schillemans, T., & Goodin, R. E. (2014). Public accountability. In M. Bovens, R. E. Goodin, & T. Schillemans (Eds.), *The Oxford Handbook of Public Accountability* (pp. 1–20). Retrieved from https://global.oup.com/academic/product/the-oxford-handbook-of-public-accountability-9780199641253

Bradler, C., Dur, R., Neckermann, S., & Non, A. (2016). Employee Recognition and Performance: A Field Experiment. *Management Science*, *62*(11), 3085–3099. https://doi.org/10.1287/mnsc.2015.2291

Breakspear, S. (2012). *The Policy Impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance* (OECD Education Working Papers No. 71). https://doi.org/10.1787/5k9fdfqffr28-en

Briole, S., & Maurin, E. (2019). *Does Evaluating Teachers Make a Difference?* (SSRN Scholarly Paper No. ID 3390297). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=3390297

Broadfoot, P., & Osborn, M. (1993). *Perceptions of teaching: Primary school teachers in England and France*. London; New York: Cassell.

Broekman, A. (2016). The effects of accountability: A case study from Indonesia. In J. Evers & R. Kneyber (Eds.), *Flip the system: Changing education from the ground up* (Online edition, pp. 72–96). https://doi.org/10.4324/9781315678573

Brown, S. A., & McIntyre, D. (1993). *Making sense of teaching*. Buckingham; Philadelphia: Open University Press.

Bruns, B., & Luque, J. (2014). *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean*. https://doi.org/10.1596/978-1-4648-0151-8

Bryan, M. L., & Jenkins, S. P. (2013). *Regression Analysis of Country Effects Using Multilevel Data: A Cautionary Tale* (No. No. 7583). Bonn: IZA.

Bryk, A. S., & Schneider, B. (2002). *Trust in Schools: A Core Resource for Improvement. A Volume in the American Sociological Association's Rose Series in Sociology.* New York: Russell Sage Foundation.

Butler, R. (2007). Teachers' achievement goal orientations and associations with teachers' help seeking: Examination of a novel approach to teacher motivation. *Journal of Educational Psychology*, *99*(2), 241–252. https://doi.org/10.1037/0022-0663.99.2.241

Butler, R. (2012). Striving to Connect: Extending an Achievement Goal Approach to Teacher Motivation to Include Relational Goals for Teaching. *Journal of Educational Psychology*, *104*(3), 726–742. https://doi.org/10.1037/a0028613

Butrymowicz, S. (2014, February 10). Lessons from Abroad: Singapore's secrets to training world-class teachers. Retrieved 5 April 2019, from The Hechinger Report website: https://hechingerreport.org/lessons-from-abroad-singapores-secrets-to-training-world-class-teachers/

Cairney, P. (2012). *Understanding public policy: Theories and issues*. Houndmills, Basingstoke, Hampshire; New York: Palgrave Macmillan.

Cambridge Assessment. (2018). *A Cambridge Approach to Improving Education*. Retrieved from University of Cambridge Local Examinations Syndicate website: https://www.cambridgeassessment.org.uk/Images/cambridge-approach-to-improving-education.pdf

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, *2*(1), 67–90. https://doi.org/10.1016/0149-7189(79)90048-X

Camphuijsen, M. K., Møller, J., & Skedsmo, G. (2019, April 15). *The adoption, evolution and sedimentation of test-based accountability in Norwegian education: Narratives, expectations and strategies*. Presented at the Comparative and International Education Society Conference, San Francisco.

Camphuijsen, M. K., Møller, J., & Skedsmo, G. (2020). Test-based accountability in the Norwegian context: Exploring drivers, expectations and strategies. *Journal of Education Policy*. https://doi.org/10.1080/02680939.2020.1739337

Candido, H. H. D., & Eriksson, S. (2019, April 16). *Finland's engagement in global education export*. Presented at the Comparative and International Education Society Annual Conference, San Francisco.

Caprara, G. V., Barbaranelli, C., Steca, P., & Malone, P. S. (2006). Teachers' self-efficacy beliefs as determinants of job satisfaction and students' academic achievement: A study at the school level. *Journal of School Psychology*, *44*(6), 473–490. https://doi.org/10.1016/j.jsp.2006.09.001

Carrasco Ogaz, D. A. (2016). *Multivariate approaches to school climate factors and school outcomes* (Doctoral thesis, University of Sussex). Retrieved from http://sro.sussex.ac.uk/id/eprint/61527/

Cartwright, N., & Hardie, J. (2012). *Evidence-Based Policy: A Practical Guide to Doing It Better*. Retrieved from https://www.oxfordscholarship.com/view/10.1093/acprof:osobl/9780199841608.001.0001/acprof-9780199841608

Cerna, L. (2014). *Trust: What it is and Why it Matters for Governance and Education* [OECD Education Working Papers]. Retrieved from OECD website: http://www.oecd-ilibrary.org/content/workingpaper/5jxswcg0t6wl-en

Cheema, J. (2014). Some General Guidelines for Choosing Missing Data Handling Methods in Educational Research. *Journal of Modern Applied Statistical Methods*, *13*(2). https://doi.org/10.22237/jmasm/1414814520

Chen, W. (2007). The Structure of Secondary School Teacher Job Satisfaction and Its Relationship with Attrition and Work Enthusiasm. *Chinese Education & Society*, *40*(5), 17–31. https://doi.org/10.2753/CED1061-1932400503

Chia, L. (2018, March 5). National Education in schools to be refreshed: Janil Puthucheary. *CNA*. Retrieved from https://www.channelnewsasia.com/news/singapore/national-education-in-schools-to-be-refreshed-janil-puthucheary-10014546

Chiang, H., Wellington, A., Hallgren, K., Speroni, C., Herrmann, M., Glazerman, S., & Constantine, J. (2015). *Evaluation of the Teacher Incentive Fund: Implementation and Impacts of Pay-for-Performance after Two Years*. NCEE 2015-4020. Retrieved from National Center for

Education Evaluation and Regional Assistance website: https://eric.ed.gov/?id=ED559723

Chiong, C., Menzies, L., & Parameshwaran, M. (2017). Why do long-serving teachers stay in the teaching profession? Analysing the motivations of teachers with 10 or more years' experience in England. *British Educational Research Journal*, *43*(6), 1083–1110. https://doi.org/10.1002/berj.3302

Chistolini, S. (2010). International Survey in Eight Countries about Teachers and Teaching Profession. *Journal of Pedagogy*, *1*(2). https://doi.org/10.2478/v10159-010-0010-9

Chiu, M. M., & Klassen, R. M. (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. https://doi.org/10.1016/j.learninstruc.2008.11.002

Chong, A., & Gradstein, M. (2015). On Education and Democratic Preferences. *Economics & Politics*, *27*(3), 362–388. https://doi.org/10.1111/ecpo.12061

Chong, E. (2018, January 29). Couple fined for lying about home address to get child enrolled in prestigious primary school. *The Straits Times*. Retrieved from https://www.straitstimes.com/singapore/courts-crime/couple-fined-for-lying-about-home-address-to-get-child-admission-in

Chua, B. H. (2018). *Liberalism Disavowed: Communitarianism and State Capitalism in Singapore*. Ithaca, NY: Cornell University Press, 2018, ©2017.

Chung, J. (2009). *An investigation of reasons for Finland's success in PISA* (PhD thesis, University of Oxford). Retrieved from https://ora.ox.ac.uk/objects/uuid:62b7a22f-d930-4eb0-893d-d703fd9d182d

Clapham, A., & Vickers, R. (2018). Neither a borrower nor a lender be: Exploring 'teaching for mastery' policy borrowing. *Oxford Review of Education*, *44*(6), 787–805. https://doi.org/10.1080/03054985.2018.1450745

Clarke, P., Crawford, C., Steele, F., & Vignoles, A. F. (2010). *The Choice Between Fixed and Random Effects Models: Some Considerations for Educational Research* (IZA Discussion Paper No. 5287). Retrieved from IZA website: https://papers.ssrn.com/abstract=1700456

Coco, G., & Lagravinese, R. (2014). Cronyism and education performance. *Economic Modelling*, *38*(Supplement C), 443–450. https://doi.org/10.1016/j.econmod.2014.01.027

Cohn, A., Maréchal, M. A., Tannenbaum, D., & Zünd, C. L. (2019). Civic honesty around the globe. *Science*, eaau8712. https://doi.org/10.1126/science.aau8712

Coleman, J. S. (1988). Social Capital in the Creation of Human Capital. *American Journal of Sociology*, *94*, S95–S120.

Collier, P. (2017). Culture, Politics, and Economic Development. *Annual Review of Political Science*, *20*(1), 111–125. https://doi.org/10.1146/annurev-polisci-051215-024720

Condron, D. J. (2011). Egalitarianism and Educational Excellence Compatible Goals for Affluent Societies? *Educational Researcher*, *40*(2), 47–55. https://doi.org/10.3102/0013189X11401021

Conway, P. F., & Murphy, R. (2013). A rising tide meets a perfect storm: New accountabilities in teaching and teacher education in Ireland. *Irish Educational Studies*, *32*(1), 11–36. https://doi.org/10.1080/03323315.2013.773227

Craig, S. G., Imberman, S. A., & Perdue, A. (2015). Do administrators respond to their accountability ratings? The response of school budgets to accountability grades. *Economics of Education Review*, *49*, 55–68. https://doi.org/10.1016/j.econedurev.2015.07.005

Creese, B., Gonzalez, A., & Isaacs, T. (2016). Comparing international curriculum systems: The international instructional systems study. *The Curriculum Journal*, *27*(1), 5–23. https://doi.org/10.1080/09585176.2015.1128346

Crehan, L. (2016). *Cleverlands: The secrets behind the success of the world's education superpowers*. London: Random House.

Crouch, D. (2015, June 17). Highly trained, respected and free: Why Finland's teachers are different. *The Guardian*. Retrieved from https://www.theguardian.com/education/2015/jun/17/highly-trained-respected-and-free-why-finlands-teachers-are-different

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. https://doi.org/10.1080/02619768.2011.587114

Davis, J., & Wilson, S. M. (2000). Principals' Efforts to Empower Teachers: Effects on Teacher Motivation and Job Satisfaction and Stress. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *73*(6), 349–353. https://doi.org/10.1080/00098650009599442

Day, P., & Klein, R. (1987). *Accountabilities: Five public services*. London: Tavistock Publishing.

De Grauwe, A., & Lugaz, C. (2007). District Education Offices in French-Speaking West Africa: Autonomy, Professionalism and Accountability. *Prospects: Quarterly Review of Comparative Education*, *37*(1), 113–125.

de Jesus, S. N., & Lens, W. (2005). An Integrated Model for the Study of Teacher Motivation. *Applied Psychology*, *54*(1), 119–134. https://doi.org/10.1111/j.1464-0597.2005.00199.x

de Lancer Julnes, P., & Holzer, M. (2001). Promoting the Utilization of Performance Measures in Public Organizations: An Empirical Study of Factors Affecting Adoption and Implementation. *Public Administration Review*, *61*(6), 693–708. https://doi.org/10.1111/0033-3352.00140

de Roock, R. S., & Espeña, D. M. (2018). Constructing underachievement: The discursive life of Singapore in US federal education policy. *Asia Pacific Journal of Education*, *38*(3), 303–318. https://doi.org/10.1080/02188791.2018.1505600

Deci, E. L. (1992). Introduction: The History of Motivation in Psychology and Its Relevance for Management. In V. H. Vroom & E. L. Deci (Eds.), *Management and motivation: Selected readings* (2nd ed., pp. 9–29). Harmondsworth: Penguin Books.

Deci, E. L., & Ryan, R. M. (2000). The 'What' and 'Why' of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry*, *11*(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

Dee, T. S., & Dizon-Ross, E. (2019). School Performance, Accountability, and Waiver Reforms: Evidence from Louisiana. *Educational Evaluation and Policy Analysis*, *41*(3), 316–349. https://doi.org/10.3102/0162373719849944

Demetriou, D., & Kyriakides, L. (2012). The impact of school self-evaluation upon student achievement: A group randomisation study. *Oxford Review of Education*, *38*(2), 149–170. https://doi.org/10.1080/03054985.2012.666032

Deming, W. E. (1993). *The new economics for industry, government, education*. Cambridge, Massachusetts: Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Diamond, J. B., & Spillane, J. P. (2004). High-stakes accountability in urban elementary schools: Challenging or reproducing inequality. *Teachers College Record*, *106*(6), 1145–1176. https://doi.org/10.1111/j.1467-9620.2004.00375.x

Dimmock, C., & Tan, C. Y. (2013). Educational leadership in Singapore: Tight coupling, sustainability, scalability, and succession. *Journal of Educational Administration*, *51*(3), 320–340. https://doi.org/10.1108/09578231311311492

Dimmock, C., & Tan, C. Y. (2016). Explaining the Success of the World's Leading Education Systems: The Case of Singapore. *British Journal of Educational Studies*, *64*(2), 161–184. https://doi.org/10.1080/00071005.2015.1116682

Dixit, A. (2002). Incentives and Organizations in the Public Sector: An Interpretative Review. *The Journal of Human Resources*, *37*(4), 696–727. https://doi.org/10.2307/3069614

Dizon-Ross, R. (2018). *How Does School Accountability Affect Teachers? Evidence from New York City* (Working Paper No. 24658). https://doi.org/10.3386/w24658

Dobbins, M., & Martens, K. (2012). Towards an education approach à la finlandaise? French education policy after PISA. *Journal of Education Policy*, *27*(1), 23–43. https://doi.org/10.1080/02680939.2011.622413

Douglas, H. (2014). Democratization, education reform, and the Mexican teachers' union. *Latin American Research Review*, *49*(1), 62–82. https://doi.org/10.1353/lar.2014.0008

Dubnick, M. J. (2003). Accountability and ethics: Reconsidering the relationships. *International Journal of Organization Theory and Behavior; Boca Raton*, *6*(3), 405–441.

Dubnick, M. J. (2005). Accountability and the Promise of Performance: In Search of the Mechanisms. *Public Performance & Management Review*, *28*(3), 376–417.

Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives Work: Getting Teachers to Come to School. *American Economic Review*, *102*(4), 1241–1278.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, *95*(2), 256–273. https://doi.org/10.1037/0033-295X.95.2.256

Easley, J. I., & Tulowitzki, P. (Eds.). (2016). *Educational Accountability: International perspectives on challenges and possibilities for school leadership*. London; New York: Routledge.

Easterday, M. (2017, February 20). The Japanese education system may solve the problems of US public education. Retrieved 23 August 2017, from The Hill website: http://thehill.com/blogs/pundits-blog/education/320350-the-japanese-education-system-may-solve-the-problems-of-us

Eddy-Spicer, D., Ehren, M. C. M., Bangpan, M., Khatwa, M., & Perrone, F. (2016). *Under what conditions do inspection, monitoring and assessment improve system efficiency, service delivery and learning outcomes for the poorest and most marginalised? A realist synthesis of school accountability in low- and middle-income countries*. London: EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University College London.

Edwards, D. B. Jr., & DeMatthews, D. E. (2014). Historical trends in educational decentralization in the United States and developing countries: A periodization and comparison in the post-WWII context. *Education Policy Analysis Archives*, *22*. https://doi.org/10.14507/epaa.v22n40.2014

Ee, D. (2018). *A Monopoly over Morality: How Moral Issues Are Publicly Resolved in Singapore* (Unpublished B.A. thesis). Yale University, New Haven, Connecticut.

Ehren, M. C. M., Eddy-Spicer, D., Bangpan, M., & Reid, A. (2016). School inspections in low- and middle-income countries: Explaining impact and mechanisms of impact. *Compare: A Journal of Comparative and International Education*, 1–15. https://doi.org/10.1080/03057925.2016.1239188

Elliott, E. S., & Dweck, C. S. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology*, *54*(1), 5.

Elsevier. (2016, January). Scopus Content Coverage Guide. Retrieved from https://www.elsevier.com/__data/assets/pdf_file/0007/69451/scopus_content_coverage_guide.pdf

Elster, J. (1998). A plea for mechanisms. In P. Hedstrom & R. Swedberg (Eds.), *Social Mechanisms* (pp. 45–73). https://doi.org/10.1017/CBO9780511663901.003

Elster, J. (2015). *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences* (2nd edition). https://doi.org/10.1017/CBO9781107763111

Engel, L. C., & Rutkowski, D. (2018). Pay to play: What does PISA participation cost in the US? *Discourse: Studies in the Cultural Politics of Education*. https://doi.org/10.1080/01596306.2018.1503591

European Commission/EACEA/Eurydice (Ed.). (2018). *Teaching careers in Europe: Access, progression and support*. Luxembourg: Publications Office of the European Union.

Evans, L. (1997). Addressing problems of conceptualization and construct validity in researching teachers' job satisfaction. *Educational Research*, *39*(3), 319–331. https://doi.org/10.1080/0013188970390307

EVS. (2011). *European Values Study 2008: Integrated Dataset (EVS 2008)* (ZA4800 Data file version 3.0.0, doi:10.4232/1.11004). Cologne: GESIS Data Archive.

EVS. (2015, October 30). Integrated values surveys 1981-2014. Retrieved 15 March 2017, from European Values Study website: http://www.europeanvaluesstudy.eu

EVS. (2016a). *EVS 2008 Method Report*. Cologne: GESIS Data Archive.

EVS. (2016b). *EVS 2008—Variable Report—Integrated Dataset* (GESIS-Variable Reports No. 2016 | 2). European Values Study and GESIS Data Archive for the Social Sciences.

Farrand, J. (1988). Mexican Primary School Teachers' Sense of Professional Responsibility. *Comparative Education*, *24*(1), 103–124.

Feenstra, R. C., Inklaar, R., & Timmer, M. P. (2016). *Penn World Table 9.0*. https://doi.org/10.15141/S5J01T

Feniger, Y., & Lefstein, A. (2014). How not to reason with PISA data: An ironic investigation. *Journal of Education Policy*, *29*(6), 845–855. https://doi.org/10.1080/02680939.2014.892156

Finland Promotion Board. (2017a). *Building the image of Finland – review of the country image work 2015–2016*. Retrieved from Finland Promotion Board website: https://toolbox.finland.fi/strategy-research/building-image-finland-review-country-image-work-2015-2016/

Finland Promotion Board. (2017b). *Finland's Country Branding Strategy*. Retrieved from https://toolbox.finland.fi/strategy-research/finlands-country-branding-strategy/

Finnish National Board of Education. (2013). *Teachers in Finland– trusted professionals*. Retrieved from http://www.oph.fi/download/148962_Teachers_in_Finland.pdf

Finnish National Board of Education. (2014). *National Core Curriculum for Basic Education*. Helsinki: Finnish National Board of Education.

Firestone, W. A., & Pennell, J. R. (1993). Teacher Commitment, Working Conditions, and Differential Incentive Policies. *Review of Educational Research*, *63*(4), 489–525. https://doi.org/10.3102/00346543063004489

Fryer, R. G. (2013). Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. *Journal of Labor Economics*, *31*(2), 373–407. https://doi.org/10.1086/667757

Fryer, R. G., & Levitt, S. D. (2010). An Empirical Analysis of the Gender Gap in Mathematics. *American Economic Journal: Applied Economics*, *2*(2), 210–240. https://doi.org/10.1257/app.2.2.210

Fullan, M., Rincón-Gallardo, S., & Hargreaves, A. (2015). Professional Capital as Accountability. *Education Policy Analysis Archives*, *23*(0), 15. https://doi.org/10.14507/epaa.v23.1998

Fukuyama, F. (2013). What Is Governance? *Governance*, *26*(3), 347–368. https://doi.org/10.1111/gove.12035

Gaduh, A., Pradhan, M., Priebe, J., & Susanti, D. (2020). Scores, Camera, Action? Incentivizing Teachers in Remote Areas (RISE Working Paper Series 20/035). Research on Improving Systems of Education (RISE). https://doi.org/10.35489/BSG-RISE-WP_2020/035

Gailmard, S. (2014). Accountability and Principal–Agent Theory. In M. Bovens, R. E. Goodin, & T. Schillemans (Eds.), *The Oxford Handbook of Public Accountability* (pp. 90–105). Oxford: Oxford University Press. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199641253.001.0001/oxfordhb-9780199641253-e-002

Ganimian, A. J., & Murnane, R. J. (2016). Improving Education in Developing Countries: Lessons From Rigorous Impact Evaluations. *Review of Educational Research*, *86*(3), 719–755. https://doi.org/10.3102/0034654315627499

Garet, M. S., Wayne, A. J., Brown, S., Rickles, J., Song, M., & Manzeske, D. (2017). *The Impact of Providing Performance Feedback to Teachers and Principals* (No. NCEE 2018-4001). Retrieved from National Center for Education Evaluation and Regional Assistance, U.S. Department of Education website: https://ies.ed.gov/ncee/pubs/20184001/pdf/20184001.pdf

Gawadi, N. (1996). A Fresh Look at Quality Assurance: Is Integrated Research Quality Assurance Possible? *The Quality Assurance Journal*, *1*(1), 11–15. https://doi.org/10.1002/(SICI)1099-1786(199609)1:1<11::AID-QAJ6>3.0.CO;2-B

Gebhardt, E. (2009, September). *Multiple Regression and Multi-Level Modelling using PISA data*. Slide deck from a workshop presented at the PISA Research Conference, Kiel, Germany. Retrieved from https://www.slideshare.net/egebhardt72/multi-level-modellingampweights-workshop-kiel09

Gelfand, M. J., Lim, B.-C., & Raver, J. L. (2004). Culture and accountability in organizations: Variations in forms of social control across cultures. *Human Resource Management Review*, *14*(1), 135–160. https://doi.org/10.1016/j.hrmr.2004.02.007

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B.-C., … Yamaguchi, S. (2011). Differences Between Tight and Loose Cultures: A 33-Nation Study. *Science*, *332*(6033), 1100–1104. https://doi.org/10.1126/science.1197754

Gerhart, B., & Fang, M. (2017). Competence and pay for performance. In *Handbook of competence and motivation: Theory and application, 2nd ed* (pp. 232–250). New York: The Guilford Press.

Gill, B. P., Lerner, J. S., & Meosky, P. (2016). Reimagining accountability in K–12 education. *Behavioral Science & Policy*, *2*(1), 57–70. https://doi.org/10.1353/bsp.2016.0007

Gilligan, D., Karachiwalla, N., Kasirye, I., Lucas, A., & Neal, D. (2018). *Educator Incentives and Educational Triage in Rural Primary Schools* (No. S-89237-UGA-1). Retrieved from International Growth Centre website: https://www.theigc.org/wp-content/uploads/2018/06/Gilligan-et-al-2018-Working-paper.pdf

Glazerman, S., & Seifullah, A. (2012). *An Evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after Four Years. Final Report.* Retrieved from Mathematica Policy Research, Inc website: https://www.mathematica-mpr.com/our-publications-and-findings/publications/an-evaluation-of-the-chicago-teacher-advancement-program-chicago-tap-after-four-years

Glewwe, P., Ilias, N., & Kremer, M. (2010). Teacher Incentives. *American Economic Journal: Applied Economics*, *2*(3), 205–227. https://doi.org/10.1257/app.2.3.205

Goh, C. B., & Gopinathan, S. (2008). The development of education in Singapore since 1965. In S.-K. Lee, C. B. Goh, B. Fredriksen, & J. P. Tan (Eds.), *Toward a better future: Education and training for economic development in Singapore since 1965* (pp. 12–38). Washington, D.C.; Singapore: World Bank; National Institute of Education.

Goh, C. B., & Lee, S. K. (2008). Making teacher education responsive and relevant. In S. K. Lee, C. B. Goh, B. Fredriksen, & J. P. Tan (Eds.), *Toward a better future: Education and training for economic development in Singapore since 1965* (pp. 96–113). Washington, D.C.; Singapore: World Bank; National Institute of Education.

Gopinathan, S. (2015). *Education*. Singapore: Institute of Policy Studies; Straits Times Press.

Gorski, P. S. (2013). "What is Critical Realism? And Why Should You Care?" *Contemporary Sociology*, *42*(5), 658–670. https://doi.org/10.1177/0094306113499533

Greany, T., & Higham, R. (2018). *Hierarchy, Markets and Networks: Analysing the 'self-improving school-led system' agenda in England and the implications for schools*. London: UCL IOE Press.

Green, A., Janmaat, J. G., & Han, C. (2009). *Regimes of social cohesion* (LLAKES Research Paper No. 1). Retrieved from Centre for Learning and Life Chances in Knowledge Economies and Societies website: https://dera.ioe.ac.uk/10486/1/Z.-Regimes-of-Social-Cohesion.pdf

Grek, S. (2009). Governing by numbers: The PISA 'effect' in Europe. *Journal of Education Policy*, *24*(1), 23–37. https://doi.org/10.1080/02680930802412669

Grek, S., Lawn, M., Lingard, B., & Varjo, J. (2009). North by northwest: Quality assurance and evaluation processes in European education. *Journal of Education Policy*, *24*(2), 121–133. https://doi.org/10.1080/02680930902733022

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, Gender, and Math. *Science*, *320*(5880), 1164–1165. https://doi.org/10.1126/science.1154094

Hammersley, M. (1998). *Reading ethnographic research: A critical guide* (2nd edition). New York; London: Longman.

Han, J., & Yin, H. (2016). Teacher motivation: Definition, research development and implications for teachers. *Cogent Education*, *3*(1), 1217819. https://doi.org/10.1080/2331186X.2016.1217819

Han, S. W. (2018). Who expects to become a teacher? The role of educational accountability policies in international perspective. *Teaching and Teacher Education*, *75*, 141–152. https://doi.org/10.1016/j.tate.2018.06.012

Han, S. W., Borgonovi, F., & Guerriero, S. (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. https://doi.org/10.3102/0002831217729875

Hanifi, R. (2013). Voluntary Work, Informal Help and Trust: Changes in Finland. *Procedia - Social and Behavioral Sciences*, *72*, 32–46. https://doi.org/10.1016/j.sbspro.2013.02.004

Hanushek, E. A. (2019). Testing, Accountability, and the American Economy. *The ANNALS of the American Academy of Political and Social Science*, *683*(1), 110–128. https://doi.org/10.1177/0002716219841299

Hanushek, E. A., Link, S., & Woessmann, L. (2013). Does school autonomy make sense everywhere? Panel estimates from PISA. *Journal of Development Economics*, *104*, 212–232. https://doi.org/10.1016/j.jdeveco.2012.08.002

Hanushek, E. A., & Rivkin, S. G. (2006). Chapter 18 Teacher Quality. In E. Hanushek & F. Welch (Eds.), *Handbook of the Economics of Education* (Vol. 2, pp. 1051–1078). https://doi.org/10.1016/S1574-0692(06)02018-6

Hanushek, E. A., & Woessmann, L. (2011). The Economics of International Differences in Educational Achievement. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the Economics of Education* (Vol. 3, pp. 89–200). https://doi.org/10.1016/B978-0-444-53429-3.00002-8

Harris, D. N., & Herrington, C. D. (2006). Accountability, Standards, and the Growing Achievement Gap: Lessons from the Past Half-Century. *American Journal of Education*, *112*(2), 209–238. https://doi.org/10.1086/498995

Harvey, L., & Green, D. (1993). Defining Quality. *Assessment & Evaluation in Higher Education*, *18*(1), 9–34. https://doi.org/10.1080/0260293930180102

Hastings, J. S., & Weinstein, J. M. (2008). Information, School Choice, and Academic Achievement: Evidence from Two Experiments. *The Quarterly Journal of Economics*, *123*(4), 1373–1414. https://doi.org/10.1162/qjec.2008.123.4.1373

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York: The Guilford Press.

He, J., van de Vijver, F. J. R., Dominguez Espinosa, A., & Mui, P. (2014). Toward a unification of acquiescent, extreme, and midpoint response styles: A multilevel study. *International Journal of Cross Cultural Management*, *14*, 1741–2838. https://doi.org/10.1177/1470595814541424

He, J., van de Vijver, F. J. R., & Kulikova, A. (2017). Country-level correlates of educational achievement: Evidence from large-scale surveys. *Educational Research and Evaluation*, *23*(5–6), 163–179. https://doi.org/10.1080/13803611.2017.1455288

Hedström, P., & Swedberg, R. (1998). Social mechanisms: An introductory essay. In P. Hedström & R. Swedberg (Eds.), *Social Mechanisms: An Analytical Approach to Social Theory* (pp. 1–31). https://doi.org/10.1017/CBO9780511663901

Heller-Sahlgren, G. (2015). *Real Finnish lessons: The true story of an education superpower*. Surrey, London: Centre for Policy Studies.

Heller-Sahlgren, G. (2018). Smart but unhappy: Independent-school competition and the wellbeing-efficiency trade-off in education. *Economics of Education Review*, *62*, 66–81. https://doi.org/10.1016/j.econedurev.2017.10.005

Heng, T. T., & Song, L. (2020). A proposed framework for understanding educational change and transfer: Insights from Singapore teachers' perceptions of differentiated instruction. *Journal of Educational Change*. https://doi.org/10.1007/s10833-020-09377-0

Hepburn, H. (2017, June 23). Scotland eyes Singapore in 'radical' overhaul of teaching career paths. *Tes*. Retrieved from https://www.tes.com/news/scotland-eyes-singapore-radical-overhaul-teaching-career-paths

Herbst, M., & Wojciuk, A. (2017). Common legacy, different paths: The transformation of educational systems in the Czech Republic, Slovakia, Hungary and Poland. *Compare*, *47*(1), 118–132. https://doi.org/10.1080/03057925.2016.1153410

Herzberg, F. (1966). *Work and the nature of man*. Cleveland: World Publishing.

Herzberg, F. (1968). One more time: How do you motivate employees? *Harvard Business Review*, *46*(1), 53–62.

Herzberg, F., Mausner, B., & Snyderman, B. B. (1959). *The motivation to work*. New York: John Wiley & Sons.

Hess, R. D., Chang, C., & McDevitt, T. M. (1987). Cultural variations in family beliefs about children's performance in mathematics: Comparisons among People's Republic of China, Chinese-American, and Caucasian-American families. *Journal of Educational Psychology*, *79*(2), 179–188. https://doi.org/10.1037/0022-0663.79.2.179

Hildebrandt, A., Trüdinger, E.-M., & Wyss, D. (2018). The Missing Link? Modernization, Tolerance, and Legislation on Homosexuality. *Political Research Quarterly*, 1065912918797464. https://doi.org/10.1177/1065912918797464

Hitt, C., Trivitt, J., & Cheng, A. (2016). When you say nothing at all: The predictive power of student effort on surveys. *Economics of Education Review*, *52*, 105–119. https://doi.org/10.1016/j.econedurev.2016.02.001

Ho, I. T., & Hau, K.-T. (2014). East meets West: Teacher motivation in the Chinese context. In P. W. Richardson, S. A. Karabenick, & H. M. G. Watt (Eds.), *Teacher motivation: Theory and practice* (pp. 133–149). New York: Routledge.

Ho, J.-M., & Koh, T.-S. (2018). Historical development of educational leadership in Singapore. In T.-S. Koh & D. Hung (Eds.), *Leadership for change: The Singapore schools' experience* (pp. 29–83). New Jersey: World Scientific.

Hofstede, G. (1980). *Culture's Consequences: International Differences in Work-Related Values*. London; Beverly Hills: SAGE Publications.

Hofstede, G. (2001). *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*. Thousand Oaks, California: SAGE.

Hofstede, G. (2015, December 8). 6 dimension data matrix (version 2015 12 08) [Dataset]. Retrieved from Geert Hofstede's official website: https://geerthofstede.com/wp-content/uploads/2016/08/6-dimensions-for-website-2015-08-16.csv

Hogan, D., Chan, M., Rahim, R., Kwek, D., Aye, K. M., Loo, S. C., … Luo, W. (2013). Assessment and the logic of instructional practice in Secondary 3 English and mathematics classrooms in Singapore. *Review of Education*, *1*(1), 57–106. https://doi.org/10.1002/rev3.3002

Hollins, M., & Reiss, M. J. (2016). A review of the school science curricula in eleven high achieving jurisdictions. *The Curriculum Journal*, *27*(1), 80–94. https://doi.org/10.1080/09585176.2016.1147968

Holloway, J., & Brass, J. (2018). Making accountable teachers: The terrors and pleasures of performativity. *Journal of Education Policy*, *33*(3), 361–382. https://doi.org/10.1080/02680939.2017.1372636

Holloway, S. D., Kashiwagi, K., Hess, R. D., & Azuma, H. (1986). Causal Attributions by Japanese and American Mothers and Children About Performance in Mathematics. *International Journal of Psychology*, *21*(1–4), 269–286. https://doi.org/10.1080/00207598608247590

Hölmstrom, B., & Milgrom, P. (1991). Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, & Organization*, *7*, 24–52.

Holzberger, D., Philipp, A., & Kunter, M. (2014). Predicting teachers' instructional behaviors: The interplay between self-efficacy and intrinsic needs. *Contemporary Educational Psychology*, *39*(2), 100–111. https://doi.org/10.1016/j.cedpsych.2014.02.001

Honig, D., & Pritchett, L. (2019). The Limits of Accounting-Based Accountability in Education (and Far Beyond): Why More Accounting Will Rarely Solve Accountability Problems. RISE Working Paper series. 19/030. https://www.riseprogramme.org/publications/rise-working-paper-19030-limits-accounting-based-accountability-education-and-far

Hoo, S. W. (2003). *Chaotic Thoughts from the Old Millennium.* Singapore: Cruxible.

Hood, C. (1983). *The Tools of Government.* London: Macmillan.

Hood, C. (2011). *The Blame Game: Spin, Bureaucracy, and Self-Preservation in Government.* Princeton: Princeton University Press.

Hood, C., & Margetts, H. Z. (2007). *The Tools of Government in the Digital Age.* New York: Palgrave Macmillan.

Hopfenbeck, T. N., Lenkeit, J., Masri, Y. E., Cantrell, K., Ryan, J., & Baird, J.-A. (2018). Lessons Learned from PISA: A Systematic Review of Peer-Reviewed Articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research*, *62*(3), 333–353. https://doi.org/10.1080/00313831.2016.1258726

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. https://doi.org/10.1080/00220270801989818

Hoy, A. W. (2008). What motivates teachers? Important work on a complex question. *Learning and Instruction*, *18*(5), 492–498. https://doi.org/10.1016/j.learninstruc.2008.06.007

Hudson, C. (2007). Governing the Governance of Education: The State Strikes Back? *European Educational Research Journal*, *6*(3), 266–282.

Hurley, A. (2013). An Education Fundamentalism? Let Them Eat Data! *Philosophical Studies in Education*, *44*, 60–74.

IBM. (2019). IBM Archives: Singapore chronology. Retrieved 24 June 2019, from //www.ibm.com/ibm/history/exhibits/asia/Singapore_ch1.html

Ingersoll, R. M. (2003). *Who Controls Teachers' Work? Power and Accountability in America's Schools.* Cambridge, Massachusetts; London: Harvard University Press.

Ingersoll, R. M., Merrill, L., & May, H. (2016). Do Accountability Policies Push Teachers Out? *Educational Leadership*, *73*(8), 44–49.

Inglehart, R. (1997). *Modernization and postmodernization: Cultural, economic, and political change in 43 societies.* Princeton: Princeton University Press.

Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., … Puranen, B. (Eds.). (2014a). *World Values Survey: Round Five—Country-Pooled Datafile Version: Www.worldvaluessurvey.org/WVSDocumentationWV5.jsp.* Madrid: JD Systems Institute.

Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., … Puranen, B. (Eds.). (2014b). *World Values Survey: Round Six—Country-Pooled Datafile Version: Http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp*. Madrid: JD Systems Institute.

Inglehart, R., & Welzel, C. (2005). *Modernization, cultural change, and democracy: The human development sequence*. Cambridge, UK: Cambridge University Press.

Ingram, D., Louis, K. S., & Schroeder, R. G. (2004). Accountability Policies and Teacher Decision Making: Barriers to the Use of Data to Improve Practice. *Teachers College Record*, *106*(6), 1258–1287. https://doi.org/10.1111/j.1467-9620.2004.00379.x

itim International. (2017). Countries—Geert Hofstede. Retrieved 17 August 2017, from Geert Hofstede website: http://geert-hofstede.com/countries.html

Iyengar, R. (2012). Social capital as the catalyst for school participation. *Compare*, *42*(6), 839–862. https://doi.org/10.1080/03057925.2012.657930

Iyengar, S. S., & Lepper, M. R. (1999). Rethinking the value of choice: A cultural perspective on intrinsic motivation. *Journal of Personality and Social Psychology*, *76*(3), 349–366. https://doi.org/10.1037/0022-3514.76.3.349

Jabbar, H. (2013). The case of 'payment-by-results': Re-examining the effects of an incentive programme in nineteenth-century English schools. *Journal of Educational Administration and History*, *45*(3), 220–243. https://doi.org/10.1080/00220620.2013.796912

Jaffer, K. (2010). School Inspection and Supervision in Pakistan: Approaches and Issues. *Prospects: Quarterly Review of Comparative Education*, *40*(3), 375–392.

Jang, H.-R. (2019). Teachers' intrinsic vs. Extrinsic instructional goals predict their classroom motivating styles. *Learning and Instruction*, *60*, 286–300. https://doi.org/10.1016/j.learninstruc.2017.11.001

Jaques, E. (1990). In praise of hierarchy. *Harvard Business Review*, *68*(1), 127–133.

Jerrim, J. (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. https://doi.org/10.1080/03054985.2015.1028525

Jerrim, J., Lopez-Agudo, L. A., Marcenaro-Gutierrez, O. D., & Shure, N. (2017). What happens when econometrics and psychometrics collide? An example using the PISA data. *Economics of Education Review*, *61*, 51–58. https://doi.org/10.1016/j.econedurev.2017.09.007

Jerrim, J., Micklewright, J., Heine, J.-H., Salzer, C., & McKeown, C. (2018). PISA 2015: How big is the 'mode effect' and what has been done about it? *Oxford Review of Education*, *44*(4), 476–493. https://doi.org/10.1080/03054985.2018.1430025

Johansson, S. (2018). Do students' high scores on international assessments translate to low levels of creativity? *Phi Delta Kappan*, *99*(7), 57–61. https://doi.org/10.1177/0031721718767863

Jones, S. A. (2019). Home school relations in Singaporean primary schools: Teachers', parents' and children's views. *Oxford Review of Education*, *45*(1), 32–49. https://doi.org/10.1080/03054985.2018.1481377

Judge, T. A., Thoresen, C. J., Bono, J. E., & Patton, G. K. (2001). The job satisfaction–job performance relationship: A qualitative and quantitative review. *Psychological Bulletin*, *127*(3), 376–407. https://doi.org/10.1037/0033-2909.127.3.376

Kalaoja, E., & Pietarinen, J. (2009). Small rural primary schools in Finland: A pedagogically valuable part of the school network. *International Journal of Educational Research*, *48*(2), 109–116. https://doi.org/10.1016/j.ijer.2009.02.003

Kan, Y. (2014, April). Your appraisal and career planning process streamlined. *Contact: The Teachers' Digest*, (14), 14–15.

Karachiwalla, N., & Park, A. (2017). Promotion incentives in the public sector: Evidence from Chinese schools. *Journal of Public Economics*, *146*, 109–128. https://doi.org/10.1016/j.jpubeco.2016.12.004

Kauko, J., Corvalán, J., Simola, H., & Carrasco, A. (2015). The historical dynamics in Chilean and Finnish basic education politics. In P. Sëppanen, A. Carrasco, M. Kalalahti, R. Rinne, & H. Simola (Eds.), *Contrasting Dynamics in Education Politics of Extremes: School Choice in Chile and Finland* (pp. 29–52). Rotterdam: Sense Publishers.

Kaur, B. (2010). Towards Excellence in Mathematics Education–Singapore's Experience. *Procedia - Social and Behavioral Sciences*, *8*, 28–34. https://doi.org/10.1016/j.sbspro.2010.12.004

Keller, M. M., Neumann, K., & Fischer, H. E. (2017). The impact of physics teachers' pedagogical content knowledge and motivation on students' achievement and interest. *Journal of Research in Science Teaching*, *54*(5), 586–614. https://doi.org/10.1002/tea.21378

Khattri, N., Ling, C., & Jha, S. (2012). The effects of school-based management in the Philippines: An initial assessment using administrative data. *Journal of Development Effectiveness*, *4*(2), 277–295. https://doi.org/10.1080/19439342.2012.692389

Khazan, O. (2013, July 11). The Secret to Finland's Success with Schools, Moms, Kids—and Everything. *The Atlantic*. Retrieved from https://www.theatlantic.com/international/archive/2013/07/the-secret-to-finlands-success-with-schools-moms-kids-and-everything/277699/

Kim, D. J. (1994, November 1). Is Culture Destiny? The Myth of Asia's Anti-Democratic Values. *Foreign Affairs*, (November/December 1994). Retrieved from https://www.foreignaffairs.com/articles/southeast-asia/1994-11-01/culture-destiny-myth-asias-anti-democratic-values

Kim, J., Sun, M., & Youngs, P. (2019). Developing the 'Will': The Relationship Between Teachers' Perceived Policy Legitimacy and Instructional Improvement. *Teachers College Record*, *121*(3), 1–44.

Kim, S. (2017). Culture matters in educational policy transfer: The case of curricular reforms in the two Koreas during the Soviet and US military occupation. *Journal of Education Policy*, *32*(3), 372–385. https://doi.org/10.1080/02680939.2016.1266034

Kirkman, B. L., Lowe, K. B., & Gibson, C. B. (2006). A quarter century of Culture's Consequences: A review of empirical research incorporating Hofstede's cultural values framework. *Journal of International Business Studies*, *37*(3), 285–320. https://doi.org/10.1057/palgrave.jibs.8400202

Kitayama, S. (2002). Culture and basic psychological processes--Toward a system view of culture: Comment on Oyserman et al (2002). *Psychological Bulletin*, *128*(1), 89–96. https://doi.org/10.1037/0033-2909.128.1.89

Klassen, R. M., & Chiu, M. M. (2010). Effects on teachers' self-efficacy and job satisfaction: Teacher gender, years of experience, and job stress. *Journal of Educational Psychology*, *102*(3), 741–756. https://doi.org/10.1037/a0019237

Klassen, R. M., Durksen, T. L., Al Hashmi, W., Kim, L. E., Longden, K., Metsäpelto, R.-L., … Györi, J. G. (2018). National context and teacher characteristics: Exploring the critical non-cognitive attributes of novice teachers in four countries. *Teaching and Teacher Education*, *72*, 64–74. https://doi.org/10.1016/j.tate.2018.03.001

Klassen, R. M., Tze, V. M. C., Betts, S. M., & Gordon, K. A. (2011). Teacher Efficacy Research 1998–2009: Signs of Progress or Unfulfilled Promise? *Educational Psychology Review*, *23*(1), 21–43. https://doi.org/10.1007/s10648-010-9141-8

Klassen, R. M., Usher, E. L., & Bong, M. (2010). Teachers' Collective Efficacy, Job Satisfaction, and Job Stress in Cross-Cultural Context. *The Journal of Experimental Education*, *78*(4), 464–486. https://doi.org/10.1080/00220970903292975

Klinge, M. (1984). Aspects of the Nordic Self. *Daedalus*, *113*(2), 257–277. Retrieved from JSTOR.

Klinge, M. (1993). Finland: From Napoleonic legacy to Nordic co-operation. In M. Teich & R. Porter (Eds.), *The National Question in Europe in Historical Context* (pp. 317–331). https://doi.org/10.1017/CBO9780511622298.014

Knack, S., & Keefer, P. (1997). Does Social Capital Have an Economic Payoff? A Cross-Country Investigation. *The Quarterly Journal of Economics*, *112*(4), 1251–1288. https://doi.org/10.1162/003355300555475

Komatsu, H., & Rappleye, J. (2017a). A new global policy regime founded on invalid statistics? Hanushek, Woessmann, PISA, and economic growth. *Comparative Education*, *53*(2), 166–191. https://doi.org/10.1080/03050068.2017.1300008

Komatsu, H., & Rappleye, J. (2017b). A PISA Paradox? An Alternative Theory of Learning as a Possible Solution for Variations in PISA Scores. *Comparative Education Review*, *61*(2), 269–297. https://doi.org/10.1086/690809

Koppell, J. G. (2005). Pathologies of Accountability: ICANN and the Challenge of "Multiple Accountabilities Disorder". *Public Administration Review*, *65*(1), 94–108. https://doi.org/10.1111/j.1540-6210.2005.00434.x

Kozlowski, K. P., & Lauen, D. L. (2019). Understanding Teacher Pay for Performance: Flawed Assumptions and Disappointing Results. *Teachers College Record*, *121*(2), 6.

Kraft, M. A., & Grace, S. (2016). *Teaching for tomorrow's economy? Teacher effects on complex cognitive skills and social-emotional competencies* [Working paper]. Retrieved from http://scholar.harvard.edu/files/mkraft/files/teaching_for_tomorrows_economy_-_final_public.pdf

Kumpulainen, K., & Lankinen, T. (2016). Striving for Educational Equity and Excellence: Evaluation and Assessment in Finnish Basic Education. In H. Niemi, A. Toom, & A. Kallioniemi (Eds.), *Miracle of Education: The Principles and Practices of Teaching and Learning in Finnish Schools* (2nd revised edition, pp. 71–82). Rotterdam: Sense Publishers.

Kunter, M., Tsai, Y.-M., Klusmann, U., Brunner, M., Krauss, S., & Baumert, J. (2008). Students' and mathematics teachers' perceptions of teacher enthusiasm and instruction. *Learning and Instruction*, *18*(5), 468–482. https://doi.org/10.1016/j.learninstruc.2008.06.008

Kurniasih, H., Utari, V. Y. D., & Akhmadi. (2018). Character Education Policy and Its Implications for Learning in Indonesia's Education System. RISE Insight Series. https://www.riseprogramme.org/publications/character-education-policy-and-its-implications-learning-indonesias-education-system

Kurzman, C. (2014, September 2). World values lost in translation. *Washington Post.* Retrieved from https://www.washingtonpost.com/news/monkey-cage/wp/2014/09/02/world-values-lost-in-translation/

Lahti, E. E. (2019). Embodied fortitude: An introduction to the Finnish construct of sisu. *International Journal of Wellbeing*, *9*(1). https://doi.org/10.5502/ijw.v9i1.672

Lam, W. L. K. (2014). *Singapore teachers' classroom assessment: Preparing students for the 'test of life,' or a 'life of tests'?* (Doctoral thesis, Lynch School of Education, Boston College). Retrieved from http://hdl.handle.net/2345/3804

Lassibille, G., Tan, J.-P., Jesse, C., & Van Nguyen, T. (2010). Managing for Results in Primary Education in Madagascar: Evaluating the Impact of Selected Workflow Interventions. *The World Bank Economic Review*, *24*(2), 303–329. https://doi.org/10.1093/wber/lhq009

Laukaityte, I., & Wiberg, M. (2017a). Importance of sampling weights in multilevel modeling of international large-scale assessment data. *Communications in Statistics - Theory and Methods, 47*(20), 4991-5012. https://doi.org/10.1080/03610926.2017.1383429

Laukaityte, I., & Wiberg, M. (2017b). Using plausible values in secondary analysis in large-scale assessments. *Communications in Statistics - Theory and Methods*, *46*(22), 11341–11357. https://doi.org/10.1080/03610926.2016.1267764

Lavy, V. (2009). Performance Pay and Teachers' Effort, Productivity, and Grading Ethics. *American Economic Review*, *99*(5), 1979–2011. https://doi.org/10.1257/aer.99.5.1979

Lawson, M., & Martin, M. (2018). *The Commitment to Reducing Inequality Index 2018: A global ranking of governments based on what they are doing to tackle the gap between rich and poor.* https://doi.org/10.21201/2018.3415

Lazear, E. P. (2000). Performance Pay and Productivity. *The American Economic Review*, *90*(5), 1346–1361.

Lazear, E. P. (2003). Teacher incentives. *Swedish Economic Policy Review*, *10*(2), 179–214.

Leaver, C., Ozier, O., Serneels, P., & Zeitlin, A. (2019). *Recruitment, effort, and retention effects of performance contracts for civil servants: Experimental evidence from Rwandan primary schools* [Working paper (preliminary)]. Retrieved from Innovations for Poverty Action website: https://www.poverty-action.org/publication/recruitment-effort-and-retention-effects-performance-contracts-civil-servants

Lee, H. L. (2004, August). *Prime Minister Lee Hsien Loong's National Day Rally 2004 speech, Sunday 22 August 2004, at the University Cultural Centre, NUS – Our future of opportunity and promise*. Speech presented at the Singapore. Retrieved from http://www.nas.gov.sg/archivesonline/speeches/view-html?filename=2004083101.htm

Lee, K. Y. (2000). *From third world to first: The Singapore story, 1965-2000*. New York: HarperCollins Publishers.

Lee, S. K., Goh, C. B., Fredriksen, B., & Tan, J. P. (Eds.). (2008). *Toward a better future: Education and training for economic development in Singapore since 1965*. Washington, D.C.; Singapore: World Bank; National Institute of Education.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255–275. https://doi.org/10.1037/0033-2909.125.2.255

Levi-Faur, D. (2012). From "Big Government" to "Big Governance"? In D. Levi-Faur (Ed.), *The Oxford Handbook of Governance* (pp. 3–18). Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199560530.013.0001

Levy, B. (2018). Improving basic education—The governance challenge. In B. Levy, R. Cameron, U. Hoadley, & V. Naidoo (Eds.), *The Politics and Governance of Basic Education: A Tale of Two South African Provinces* (pp. 3–26). Oxford: New York: Oxford University Press.

Lewin, K. (1935). *A Dynamic Theory of Personality* (D. K. Adams & K. E. Zener, Trans.). New York; London: McGraw Hill Book Company Inc.

Li, L., Hu, H., Zhou, H., He, C., Fan, L., Liu, X., Zhang, Z., Li, H., & Sun, T. (2014). Work stress, work motivation and their effects on job satisfaction in community health workers: A cross-sectional survey in China. *BMJ Open*, *4*(6), e004897. https://doi.org/10.1136/bmjopen-2014-004897

Li, J. (2012). *Cultural foundations of learning: East and West*. New York: Cambridge University Press.

Li, Y., & Dervin, F. (2018). *Continuing Professional Development of Teachers in Finland*. https://doi.org/10.1007/978-3-319-95795-1

Liaw, Y.-L., Wu, Y., Rutkowski, D., & Rutkowski, L. (2018). Evaluating PISA scales across Chinese economies. *Asia Pacific Journal of Education*, *38*(3), 432–451. https://doi.org/10.1080/02188791.2018.1491388

Liew, W. M. (2012). Perform or else: The performative enhancement of teacher professionalism. *Asia Pacific Journal of Education*, *32*(3), 285–303. https://doi.org/10.1080/02188791.2012.711297

Lipsky, M. (2010). *Street-Level Bureaucracy: Dilemmas of the Individual in Public Service* (30th Anniversary Edition). New York: Russell Sage Foundation.

Liu, X. S., & Ramsey, J. (2008). Teachers' job satisfaction: Analyses of the Teacher Follow-up Survey in the United States for 2000–2001. *Teaching and Teacher Education*, *24*(5), 1173–1184. https://doi.org/10.1016/j.tate.2006.11.010

Locke, E. A. (1969). What is job satisfaction? *Organizational Behavior and Human Performance*, *4*(4), 309–336. https://doi.org/10.1016/0030-5073(69)90013-0

Loh, J. (2016). From Fantasy to Depression: A Beginning Teacher's encounter with Performativity. *AsTEN Journal of Teacher Education*, *1*(1), 1-12. Retrieved from http://po.pnuresearchportal.org/ejournal/index.php/asten/article/view/143

Loh, J., & Hu, G. (2014). Subdued by the system: Neoliberalism and the beginning teacher. *Teaching and Teacher Education*, *41*, 13–21. https://doi.org/10.1016/j.tate.2014.03.005

Low, E.-L. (2016). Letter 3. Planning the visit by U.S. education leaders: From conceptualisation to realisation. In F. M. Reimers & E. B. O'Donnell (Eds.), *Fifteen Letters On Education In Singapore: Reflections from a Visit to Singapore In 2015 By a Delegation of Educators from Massachusetts* (E-book.). Retrieved from http://www.lulu.com/shop/fernando-m-reimers-and-e-b-odonnell/fifteen-letters-on-education-in-singapore-reflections-from-a-visit-to-singapore-in-2015-by-a-delegation-of-educators-from-massachusetts/ebook/product-22728679.html

Low, E.-L., & Tan, O.-S. (2017). Teacher Education Policy: Recruitment, Preparation and Progression. In *Teacher Education in the 21st Century* (pp. 11–32). https://doi.org/10.1007/978-981-10-3386-5_2

Luetsch, K., Twigg, M., Rowett, D., & Wong, G. (2019). In search for gold—The relevance of realist reviews and evaluations to pharmacy research and policy development. *Research in Social and Administrative Pharmacy*. https://doi.org/10.1016/j.sapharm.2019.07.002

Lui, J. (2018, July 2). Johnny Lau is the opposite of his creation Mr Kiasu. *The Straits Times*. Retrieved from https://www.straitstimes.com/lifestyle/the-opposite-of-mr-kiasu

Luna, L. (2015). Cooperative learning and embodied accountability: An ethnographic analysis of classroom participation in an English school. *Education Policy Analysis Archives*, *23*. https://doi.org/10.14507/epaa.v23.2050

Maas, C. J. M., & Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Macartney, H., McMillan, R., & Petronijevic, U. (2018). *Teacher Performance and Accountability Incentives* (Working Paper No. 24747). https://doi.org/10.3386/w24747

Mackie, J. L. (1965). Causes and Conditions. *American Philosophical Quarterly*, *2*(4), 245–264.

Mahoney, J. (2001). Beyond Correlational Analysis: Recent Innovations in Theory and Method. *Sociological Forum*, *16*(3), 575–593.

Malinen, O.-P., Väisänen, P., & Savolainen, H. (2012). Teacher education in Finland: A review of a national effort for preparing teachers for the future. *Curriculum Journal*, *23*(4), 567–584. https://doi.org/10.1080/09585176.2012.731011

Maluka, S., Kamuzora, P., SanSebastián, M., Byskov, J., Ndawi, B., Olsen, Ø. E., & Hurtig, A.-K. (2011). Implementing accountability for reasonableness framework at district level in Tanzania: A realist evaluation. *Implementation Science*, *6*(1), 11. https://doi.org/10.1186/1748-5908-6-11

Mansbridge, J. (2009). A "Selection Model" of Political Representation. *Journal of Political Philosophy*, *17*(4), 369–398. https://doi.org/10.1111/j.1467-9760.2009.00337.x

Mansbridge, J. (2014). A Contingency Theory of Accountability. In M. Bovens, R. E. Goodin, & T. Schillemans (Eds.), *The Oxford Handbook of Public Accountability* (pp. 55–68). Oxford: Oxford University Press. https://global.oup.com/academic/product/the-oxford-handbook-of-public-accountability-9780199641253

Manzano-Santaella, A. (2011). A realistic evaluation of fines for hospital discharges: Incorporating the history of programme evaluations in the analysis. *Evaluation*, *17*(1), 21–36. https://doi.org/10.1177/1356389010389913

Marchal, B., Dedzo, M., & Kegels, G. (2010). A realist evaluation of the management of a well-performing regional hospital in Ghana. *BMC Health Services Research*, *10*(1), 24. https://doi.org/10.1186/1472-6963-10-24

Markus, H. R., & Conner, A. (2013). *Clash! How to Thrive in a Multicultural World*. New York: Penguin Publishing Group.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*(2), 224–253. https://doi.org/10.1037/0033-295X.98.2.224

Markus, H. R., & Kitayama, S. (2010). Cultures and Selves: A Cycle of Mutual Constitution. *Perspectives on Psychological Science*, *5*(4), 420–430.

Maroy, C. (2009). Convergences and Hybridization of Educational Policies around 'Post-Bureaucratic' Models of Regulation. *Compare: A Journal of Comparative and International Education*, *39*(1), 71–84.

Martin, M. O., Mullis, I. V. S., Foy, P., & Hooper, M. (2016a). *TIMSS 2015 International Results in Mathematics*. Chestnut Hill, Massachusetts: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., Foy, P., & Hooper, M. (2016b). *TIMSS 2015 International Results in Science*. Chestnut Hill, Massachusetts: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., & Hooper, I. (Eds.). (2016). *Methods and Procedures in TIMSS 2015*. Chestnut Hill, Massachusetts: TIMSS & PIRLS International Study Center, Boston College.

Maslow, A. H. (1954). *Motivation and personality*. New York: Harper.

Mattei, P. (2012). Market Accountability in Schools: Policy Reforms in England, Germany, France and Italy. *Oxford Review of Education*, *38*(3), 247–266.

Mausethagen, S. (2013). A research review of the impact of accountability policies on teachers' workplace relations. *Educational Research Review*, *9*, 16–33. https://doi.org/10.1016/j.edurev.2012.12.001

Maxwell, J. A. (2012). *A realist approach for qualitative research*. Los Angeles: SAGE Publications.

Maxwell, J. A. (2013). *Qualitative research design: An interactive approach* (3rd ed). Thousand Oaks, California: SAGE Publications.

Maxwell, J. A. (2017). The validity and reliability of research: A realist perspective. In D. Wyse, N. Selwyn, E. Smith, & L. Suter (Eds.), *The BERA/SAGE handbook of educational research* (pp. 194–226). Los Angeles: SAGE.

Maxwell, J. A., & Mittapalli, K. (2010). Realism as a Stance for Mixed Methods Research. In A. Tashakkori & C. Teddlie, *SAGE Handbook of Mixed Methods in Social & Behavioral Research* (pp. 145–168). https://doi.org/10.4135/9781506335193.n6

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., & Rajani, R. (2019). Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania. *The Quarterly Journal of Economics*, *134*(3), 1627–1673. https://doi.org/10.1093/qje/qjz010

McDermott, K. A. (2011). *High-Stakes Reform: The Politics of Educational Accountability*. Washington, D.C.: Georgetown University Press.

McDonnell, L. M., & Elmore, R. F. (1987). Getting the Job Done: Alternative Policy Instruments. *Educational Evaluation and Policy Analysis*, *9*(2), 133–152. https://doi.org/10.3102/01623737009002133

McLaughlin, M. W. (1987). Learning From Experience: Lessons From Policy Implementation. *Educational Evaluation and Policy Analysis*, *9*(2), 171–178. https://doi.org/10.3102/01623737009002171

McMillan, C. (2017). *Churchill Fellowship report: To investigate how public education systems of government in Singapore and Finland develop and prepare leaders*. Retrieved from Winston Churchill Memorial Trust website: https://www.churchilltrust.com.au/media/fellows/McMillan_C_2015_educational_leaders_and_increased_autonomy.pdf

Mehta, J. (2013). *The Allure of Order: High Hopes, Dashed Expectations, and the Troubled Quest to Remake American Schooling*. Oxford: Oxford University Press.

Mellon, J. (2011). *Examining survey translation validity using corpus linguistics*. Nuffield College Working Papers Series in Politics No. 2011-08. Retrieved from Nuffield College website: http://www.nuffield.ox.ac.uk/politics/papers/2011/Jon%20Mellon_working%20paper%202011_08.pdf

Merritt, A. (2000). Culture in the Cockpit: Do Hofstede's Dimensions Replicate? *Journal of Cross-Cultural Psychology*, *31*(3), 283–301. https://doi.org/10.1177/0022022100031003001

Merton, R. K. (1968). *Social theory and social structure* (1968 enl. ed.). New York: Free Press.

Meyer, H.-D., & Rowan, B. (Eds.). (2006). *The new institutionalism in education*. Albany: State University of New York Press.

Meyer, H.-D., & Schiller, K. (2013). Gauging the role of non-educational effects in large-scale assessments: Socio-economics, culture and PISA outcomes. In H.-D. Meyer & A. Benavot (Eds.), *PISA, power, and policy: The emergence of global educational governance* (pp. 207–224). Oxford: Symposium Books.

Meyer, J. W., & Rowan, B. (1977). Institutionalized Organizations: Formal Structure as Myth and Ceremony. *American Journal of Sociology*, *83*(2), 340–363. https://doi.org/10.1086/226550

Microsoft Trust Center. (n.d.). Microsoft and the GDPR FAQs. Retrieved 6 August 2019, from https://www.microsoft.com/en-us/trust-center/privacy/gdpr-faqs

Mizel, O. (2009). Accountability in Arab Bedouin schools in Israel: Accountable to whom? *Educational Management Administration and Leadership*, *37*(5), 624–644. https://doi.org/10.1177/1741143209339654

Møller, J., & Skedsmo, G. (2013). Modernising education: New Public Management reform in the Norwegian education system. *Journal of Educational Administration and History*, *45*(4), 336–353. https://doi.org/10.1080/00220620.2013.822353

Monaghan, C., & King, E. (2018). How Theories of Change Can Improve Education Programming and Evaluation in Conflict-Affected Contexts. *Comparative Education Review*, *62*(3), 365–384. https://doi.org/10.1086/698405

Mourshed, M., Chijioke, C., & Barber, M. (2010). *How the world's most improved school systems keep getting better*. Retrieved from McKinsey & Company website: http://www.mckinsey.com/~/media/mckinsey/dotcom/client_service/social%20sector/pdfs/how-the-worlds-most-improved-school-systems-keep-getting-better_download-version_final.ashx

Muhonen, S. (2017, October 16). In Finland, it's easier to become a doctor or lawyer than a teacher—Here's why [News blog]. Retrieved 22 February 2019, from The Hechinger Report website: https://hechingerreport.org/teacher-voice-in-finland-its-easier-to-become-a-doctor-or-lawyer-than-a-teacher-heres-why/

Mulder, M. (1977). *The daily power game*. Leiden: Nijhoff.

Mulgan, R. (2000). 'Accountability': An Ever-Expanding Concept? *Public Administration*, *78*(3), 555–573. https://doi.org/10.1111/1467-9299.00218

Müller, J., & Hernández, F. (2010). On the geography of accountability: Comparative analysis of teachers' experiences across seven European countries. *Journal of Educational Change*, *11*(4), 307–322. https://doi.org/10.1007/s10833-009-9126-x

Muller, J. Z. (2018). *The tyranny of metrics*. Princeton: Princeton University Press.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2013). *TIMSS 2015 Assessment Frameworks*. Chestnut Hill, Massachusetts: TIMSS & PIRLS International Study Center, Boston College.

Munyengabe, S., He, H., & Yiyi, Z. (2016). The Analysis of Factors and Levels Associated with Lecturers' Motivation and Job Satisfaction in University of Rwanda. *Journal of Education and Practice*, *7*(30), 188–200.

Muralidharan, K., & Sundararaman, V. (2010). The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India. *The Economic Journal*, *120*(546), F187–F203. https://doi.org/10.1111/j.1468-0297.2010.02373.x

Muralidharan, K., & Sundararaman, V. (2011). Teacher Performance Pay: Experimental Evidence from India. *Journal of Political Economy*, *119*(1), 39–77. https://doi.org/10.1086/659655

Murnane, R. J., & Cohen, D. K. (1986). Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and a Few Survive. *Harvard Educational Review*, *56*(1), 1–18. https://doi.org/10.17763/haer.56.1.l8q2334243271116

Murnane, R. J., & Phillips, B. R. (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. https://doi.org/10.1016/0272-7757(81)90015-7

Murray, H. A. (1938). *Explorations In Personality*. Retrieved from http://archive.org/details/explorationsinpe031973mbp

Murtedjo, & Suharningsih. (2016). Contribution to Cultural Organization, Working Motivation and Job Satisfaction on the Performance of Primary School Teacher. *International Journal of Higher Education*, *5*(4), 86–95.

Nardon, L., & Steers, R. M. (2009). The culture theory jungle: Divergence and convergence in models of national culture. In R. S. Bhagat & R. M. Steers (Eds.), *Cambridge Handbook of Culture, Organizations, and Work* (pp. 3–22). https://doi.org/10.1017/CBO9780511581151.002

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. https://doi.org/10.1177/1478210314568242

National Research Council. (2011). *Incentives and Test-Based Accountability in Education* (M. Hout & S. W. Elliot, Eds.). Washington, D.C.: National Academies Press.

NCEE. (2016). *Empowered Educators country brief: Finland: Constructing teacher quality*. Retrieved from National Center on Education and the Economy website: http://ncee.org/wp-content/uploads/2017/02/FinlandCountryBrief.pdf

Neo, B. S., & Chen, G. (2007). *Dynamic governance: Embedding culture, capabilities and change in Singapore*. New Jersey: World Scientific.

Newton, K. (2001). Trust, Social Capital, Civil Society, and Democracy. *International Political Science Review*, *22*(2), 201–214. https://doi.org/10.1177/0192512101222004

Ng, P. T. (2013). An examination of school accountability from the perspective of Singapore school leaders. *Educational Research for Policy and Practice*, *12*(2), 121–131. https://doi.org/10.1007/s10671-012-9127-z

Ng, P. T. (2017). *Learning from Singapore: The power of paradoxes*. New York: Routledge, Taylor & Francis Group.

Ng, P. T., & Chan, D. (2008). A comparative study of Singapore's school excellence model with Hong Kong's school-based management. *International Journal of Educational Management*, *22*(6), 488–505. https://doi.org/10.1108/09513540810895426

Nikki, M.-L. (2000). The research context in Finland. In S. Hämäläinen, E. Kimonen, R. Nevalainen, & M.-L. Nikki (Eds.), *Consensus or Compromise? Making the School-based Curricula in Lower-level Comprehensive Schools in Finland* (pp. 9–20). Jyväskylä: University of Jyväskylä, Department of Teacher Education.

Nitsche, S., Dickhäuser, O., Fasching, M. S., & Dresel, M. (2011). Rethinking teachers' goal orientations: Conceptual and methodological enhancements. *Learning and Instruction*, *21*(4), 574–586. https://doi.org/10.1016/j.learninstruc.2010.12.001

Norris, P., & Inglehart, R. (2004). *Sacred and Secular: Religion and Politics Worldwide*. Cambridge, UK: Cambridge University Press.

Nunes, L. C., Reis, A. B., & Seabra, C. (2015). The publication of school rankings: A step toward increased accountability? *Economics of Education Review*, *49*, 15–23. https://doi.org/10.1016/j.econedurev.2015.07.008

Nylund, J. (2018). *Sisu: The Finnish art of courage*. London: Gaia/Octopus Books.

OAJ. (2018). *Opetusalan työolobarometri 2017 [Barometer of teachers' working conditions]* (No. OAJ:n julkaisusarja 5:2018). Retrieved from OAJ (Trade Union of Education in Finland) website: https://www.oaj.fi/ajankohtaista/julkaisut/2018/opetusalan-tyoolobarometri/

Oates, T. (2015). *Finnish Fairy Stories*. Retrieved from Cambridge Assessment website: https://www.cambridgeassessment.org.uk/Images/207376-finnish-fairy-stories-tim-oates.pdf

OECD. (1982). *Reviews of national policies for education: Finland*. Paris: OECD Publishing.

OECD. (2009). *PISA Data Analysis Manual: SPSS Second Edition*. Paris: OECD Publishing.

OECD. (2013a). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. Paris: OECD Publishing.

OECD. (2013b). *PISA 2012 Results: What Makes Schools Successful (Volume IV): Resources, Policies and Practices*. Paris: OECD Publishing.

OECD. (2014a). *PISA 2012 Results: What Students Know and Can Do (Volume I, Revised edition)*. Paris: OECD Publishing.

OECD. (2014b). *PISA 2012 Technical Report*. Paris: OECD Publishing.

OECD. (2014c). *TALIS 2013 Results*. Paris: OECD Publishing.

OECD. (2014d). *TALIS 2013 Technical Report*. Paris: OECD Publishing.

OECD. (2016a). Annex B1.4 Results (tables): Governance, assessment and accountability. Retrieved 2 September 2019, from PISA 2015 Results (Volume II): Policies and Practices for Successful Schools website: http://statlinks.oecdcode.org/982016071p1t004.xlsx

OECD. (2016b). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy*. Paris: OECD Publishing.

OECD. (2016c). PISA 2015 codebooks for the main files. Retrieved 6 December 2016, from PISA 2015 Database website: https://www.oecd.org/pisa/data/2015database/Codebook_CMB.xlsx

OECD. (2016d). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris: OECD Publishing.

OECD. (2016e). *PISA 2015 Results (Volume II): Policies and Practices for Successful Schools*. Paris: OECD Publishing.

OECD. (2016f). PISA 2015 Annex B1.4 Results (tables): Governance, assessment and accountability [Dataset]. Retrieved 17 August 2017, from OECD Code website: http://statlinks.oecdcode.org/982016071p1t004.xlsx

OECD. (2017). *PISA 2015 Technical Report*. Paris: OECD Publishing.

O'Neill, O. (2002). *A Question of Trust: The BBC Reith Lectures 2002*. Cambridge; New York: Cambridge University Press.

Osborn, M. (2006). Changing the context of teachers' work and professional development: A European perspective. *International Journal of Educational Research*, *45*(4–5), 242–253. https://doi.org/10.1016/j.ijer.2007.02.008

Overall, J. U., & Marsh, H. W. (1979). Midterm feedback from students: Its relationship to instructional improvement and students' cognitive and affective outcomes. *Journal of Educational Psychology*, *71*(6), 856–865. https://doi.org/10.1037/0022-0663.71.6.856

Oxford English Dictionary. (2006). kiasu, n. And adj. In *OED Online* (3rd edition). Retrieved from https://www.oed.com/view/Entry/251775

Page, E. C. (2008). The Origins of Policy. In R. E. Goodin, M. Moran, & M. Rein (Eds.), *The Oxford Handbook of Public Policy*. Retrieved from https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199548453.001.0001/oxfordhb-9780199548453-e-010

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. (2016). *Learning Job Skills from Colleagues at Work: Evidence from a Field Experiment Using Teacher Performance Data* (Working Paper No. 21986). https://doi.org/10.3386/w21986

Paronen, P., & Lappi, O. (2018). *Finnish teachers and principals in figures* (No. 2018:4). Retrieved from Finnish National Agency for Education website: https://www.oph.fi/download/189802_finnish_teachers_and_principals_in_figures.pdf

Partanen, A. (2016). *The Nordic theory of everything: In search of a better life*. New York: Harper.

Pawson, R. (2000). Middle-range realism. *European Journal of Sociology / Archives Européennes de Sociologie*, *41*(2), 283–325. https://doi.org/10.1017/S0003975600007050

Pawson, R. (2013). *The Science of Evaluation: A Realist Manifesto.* https://doi.org/10.4135/9781473913820

Pawson, R. (2018). Realist memorabilia. In N. Emmel, J. Greenhalgh, A. Manzano, M. Monaghan, & S. Dalkin (Eds.), *Doing realist research* (pp. 203–220). Los Angeles: Sage.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. London: SAGE Publications.

Pawson, R., & Tilley, N. (2004). *Realist evaluation*. Retrieved from http://www.communitymatters.com.au/RE_chapter.pdf

Pearson, M., Chilton, R., Wyatt, K., Abraham, C., Ford, T., Woods, H., & Anderson, R. (2015). Implementing health promotion programmes in schools: A realist systematic review of research and experience in the United Kingdom. *Implementation Science*, *10*(1), 149. https://doi.org/10.1186/s13012-015-0338-6

Peters, B. G. (2012). Governance As Political Theory. In D. Levi-Faur (Ed.), *The Oxford Handbook of Governance* (pp. 19–32). OxfordL Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199560530.013.0001

Pierson, D. (2019, January 18). Singapore's 'kiasu' culture makes FOMO look like child's play. *Los Angeles Times*. Retrieved from https://www.latimes.com/world/asia/la-fg-singapore-kiasu-fomo-20190118-story.html

Pillinger, R. (2011, September 8). *Weighting in MLwiN*. Centre for Multilevel Modelling, University of Bristol.

Pinder, C. C. (1992). Valence-Instrumentality-Expectancy Theory. In V. H. Vroom & E. L. Deci (Eds.), *Management and motivation: Selected readings* (2nd ed., pp. 90–102). Harmondsworth: Penguin Books.

Plaut, V. C., & Markus, H. R. (2005). The 'Inside' Story: A Cultural-Historical Analysis of Being Smart and Motivated, American Style. In *Handbook of competence and motivation* (pp. 457–488). New York, NY, US: Guilford Publications.

Podgursky, M. J., & Springer, M. G. (2007). Teacher performance pay: A review. *Journal of Policy Analysis and Management*, *26*(4), 909–950. https://doi.org/10.1002/pam.20292

Poland, B. D. (2001). Transcription Quality. In J. Gubrium & J. Holstein (Eds.), *Handbook of Interview Research* (pp. 628–649). https://doi.org/10.4135/9781412973588.n36

Pollitt, C., & Bouckaert, G. (2017). *Public management reform: A comparative analysis - into the age of austerity* (4th edition). New York, NY: Oxford University Press.

Porpora, D. V. (2015). *Reconstructing Sociology: The Critical Realist Approach*. Cambridge, United Kingdom: Cambridge University Press.

Pritchett, L. (2015). *Creating Education Systems Coherent for Learning Outcomes: Making the Transition from Schooling to Learning* (No. RISE-WP-15/005). Retrieved from RISE Programme website: http://www.riseprogramme.org/sites/www.riseprogramme.org/files/RISE_WP-004_Moore-REV%20copy.pdf

Pritchett, L. (2017). *"The Evidence" About "What Works" in Education: Graphs to Illustrate External Validity and Construct Validity*. Retrieved from RISE Programme website: https://www.riseprogramme.org/publications/evidence-about-what-works-education-graphs-illustrate-external-validity-and-construct

Public School Insights. (2008, September 29). In Teachers We Trust: An Interview with Finnish Education Expert Reijo Laukkanen. Retrieved 10 June 2019, from Learning First Alliance website: https://learningfirst.org/blog/teachers-we-trust-interview-finnish-education-expert-reijo-laukkanen

Putnam, R. D. (1995). Bowling Alone: America's Declining Social Capital. *Journal of Democracy*, *6*(1), 65–78. https://doi.org/10.1353/jod.1995.0002

Quah, J. S. T. (2010). *Public Administration Singapore-style*. https://doi.org/10.1108/S0732-1317(2010)19

Quintelier, A., Vanhoof, J., & De Maeyer, S. (2018). Understanding the influence of teachers' cognitive and affective responses upon school inspection feedback acceptance. *Educational Assessment, Evaluation and Accountability*, *30*(4), 399–431. https://doi.org/10.1007/s11092-018-9286-4

Rapple, B. A. (1994). Payment by Results: An Example of Assessment in Elementary Education from Nineteenth Century Britain. *Education Policy Analysis Archives*, *2*, 1.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed). Thousand Oaks: Sage Publications.

Redden, G., & Low, R. (2012). My School, Education, and Cultures of Rating and Ranking. *Review of Education, Pedagogy & Cultural Studies*, *34*(1–2), 35–48.

Reimers, F. M., & O'Donnell, E. B. (Eds.). (2016). *Fifteen Letters On Education In Singapore: Reflections from a Visit to Singapore In 2015 By a Delegation of Educators from Massachusetts* (E-book). Retrieved from http://www.lulu.com/shop/fernando-m-reimers-and-e-b-odonnell/fifteen-letters-on-education-in-singapore-reflections-from-a-visit-to-singapore-in-2015-by-a-delegation-of-educators-from-massachusetts/ebook/product-22728679.html

Rinne, R., Kivirauma, J., & Simola, H. (2002). Shoots of revisionist education policy or just slow readjustment? The Finnish case of educational reconstruction. *Journal of Education Policy*, *17*(6), 643–658. https://doi.org/10.1080/0268093022000032292

Ripley, A. (2013). *The Smartest Kids in the World: And How They Got That Way*. New York: Simon & Schuster.

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. Policy Sciences, 4(2), 155–169. https://doi.org/10.1007/BF01405730

Robertson, S. (2015). What teachers need to know about the 'Global Education Reform Movement'. In G. Little (Ed.), *Global education 'reform': Building resistance and solidarity* (pp. 10–17). Croydon: Manifesto Press.

Robertson-Kraft, C. (2014). Teachers' Motivational Responses to New Teacher Performance Management Systems: An Evaluation of the Pilot of Aldine ISD's inVEST System. *Publicly Accessible Penn Dissertations*. Retrieved from https://repository.upenn.edu/edissertations/1420

Rockoff, J., & Turner, L. J. (2010). Short-Run Impacts of Accountability on School Quality. *American Economic Journal: Economic Policy*, *2*(4), 119–147. https://doi.org/10.1257/pol.2.4.119

Rodríguez-Planas, N., & Nollenberger, N. (2018). Let the girls learn! It is not only about math … it's about gender social norms. *Economics of Education Review*, *62*, 230–253. https://doi.org/10.1016/j.econedurev.2017.11.006

Romzek, B. S. (2014). Accountable Public Services. In M. Bovens, R. E. Goodin, & T. Schillemans (Eds.), *The Oxford Handbook of Public Accountability* (pp. 307–323). Retrieved from http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199641253.001.0001/oxfordhb-9780199641253-e-002

Romzek, B. S., & Dubnick, M. J. (1987). Accountability in the Public Sector: Lessons from the Challenger Tragedy. *Public Administration Review*, *47*(3), 227–238. https://doi.org/10.2307/975901

Roth, G., Assor, A., Kanat-Maymon, Y., & Kaplan, H. (2007). Autonomous motivation for teaching: How self-determined teaching may lead to self-determined learning. *Journal of Educational Psychology*, *99*(4), 761–774. https://doi.org/10.1037/0022-0663.99.4.761

Rowan, B. (1990). Chapter 7: Commitment and Control: Alternative Strategies for the Organizational Design of Schools. *Review of Research in Education*, *16*(1), 353–389. https://doi.org/10.3102/0091732X016001353

Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, *91*(434), 473–489. https://doi.org/10.2307/2291635

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International Large-Scale Assessment Data: Issues in Secondary Analysis and Reporting. *Educational Researcher*, *39*(2), 142–151. https://doi.org/10.3102/0013189X10363170

Ryan, R. M., & Deci, E. L. (2000a). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology*, *25*(1), 54–67. https://doi.org/10.1006/ceps.1999.1020

Ryan, R. M., & Deci, E. L. (2000b). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Ryan, R. M., & Deci, E. L. (2000c). The Darker and Brighter Sides of Human Existence: Basic Psychological Needs as a Unifying Concept. *Psychological Inquiry*, *11*(4), 319–338. https://doi.org/10.1207/S15327965PLI1104_03

Sahlberg, P. (2010). Rethinking accountability in a knowledge society. *Journal of Educational Change*, *11*(1), 45–61. https://doi.org/10.1007/s10833-008-9098-2

Sahlberg, P. (2012). *Finnish Lessons: What Can the World Learn from Educational Change in Finland?* New York: Teachers' College Press.

Sahlberg, P. (2015a). *Finnish Lessons 2.0: What Can the World Learn from Educational Change in Finland?* (2nd edition). New York: Teachers College Press.

Sahlberg, P. (2015b, March 31). Q: What makes Finnish teachers so special? A: It's not brains. *The Guardian*. Retrieved from https://www.theguardian.com/education/2015/mar/31/finnish-teachers-special-train-teach

Sahlberg, P. (2016). The Global Educational Reform Movement and Its Impact on Schooling. In *The Handbook of Global Education Policy* (pp. 128–144). https://doi.org/10.1002/9781118468005.ch7

SAK. (2016, March 7). The Competitiveness Pact in brief [Organisation website]. Retrieved 21 February 2019, from SAK: The Central Organisation of Finnish Trade Unions website: https://www.sak.fi/en/whats-new/news/competitiveness-pact-brief

Schleicher, A. (2018). *World class: How to build a 21st-century school system*. Paris: OECD.

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology*, *57*(1), 1.

Schunk, D. H., Pintrich, P. R., & Meece, J. L. (2010). *Motivation in education: Theory, research, and applications* (3rd edition). London: Pearson, Merrill, Prentice-Hall.

Schwab, D. P., & Cummings, L. L. (1970). Theories of Performance and Satisfaction: A Review. *Industrial Relations: A Journal of Economy and Society*, *9*(4), 408–430. https://doi.org/10.1111/j.1468-232X.1970.tb00524.x

Schwartz, R. (2000). School accountability - An elusive policy solution: The Israeli experience in comparative perspective. *Journal of Public Policy*, *20*(2), 195–218. https://doi.org/10.1017/S0143814X00000817

Sclafani, S., & Lim, E. (2008). *Rethinking Human Capital in Education: Singapore as a Model for Teacher Development*. Retrieved from Aspen Institute website: http://eric.ed.gov/?id=ED512422

Scott, J. (1998). *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed* (New edition). New Haven, Connecticut: Yale University Press.

Seah, C. N. (2006, July 9). Insight Down South: Falling back on autopilot. *The Star (Malaysia)*. Retrieved from https://web.archive.org/web/20110521150911/http://thestar.com.my/news/story.asp?file=%2F2006%2F7%2F9%2Ffocus%2F14774318&sec=focus

Sen, A. K. (1999). Democracy as a Universal Value. *Journal of Democracy*, *10*(3), 3–17. https://doi.org/10.1353/jod.1999.0055

Sharp, C., Walker, M., Lynch, S., Worth, J., Greaves, E., Bernardinelli, D., … Murphy, R. (2017). *Evaluation of Teachers' Pay Reform: Final Report*. Retrieved from Department for Education website: https://www.gov.uk/government/publications/teachers-pay-reform-evaluation

Sim, J. B.-Y., & Print, M. (2005). Citizenship education and social studies in Singapore: A national agenda. *International Journal of Citizenship and Teacher Education*, *1*(1), 58–73.

Simola, H. (2014). *The Finnish Education Mystery: Historical and sociological essays on schooling in Finland* (1st edition). London; New York: Routledge.

Simola, H., Kauko, J., Varjo, J., Kalalahti, M., & Sahlstrom, F. (2017). *Dynamics in Education Politics: Understanding and explaining the Finnish case*. Abingdon, Oxfordshire; New York: Routledge.

Simola, H., Rinne, R., Varjo, J., Pitkänen, H., & Kauko, J. (2009). Quality assurance and evaluation (QAE) in Finnish compulsory schooling: A national model or just unintended effects of radical decentralisation? *Journal of Education Policy*, *24*(2), 163–178. https://doi.org/10.1080/02680930902733139

Singapore Department of Statistics. (2019). *Report on the Household Expenditure Survey 2017/18* (No. ISSN 2661-4103). Retrieved from Department of Statistics, Ministry of Trade and Industry website: https://www.singstat.gov.sg/-/media/files/publications/households/hes201718.pdf

Singapore MOE. (2013). *Engaging our learners: Teach less, learn more*. Singapore: Ministry of Education.

Singapore MOE. (2018a). *Our Home, Our Say: National Education Review 2016-2017*. Retrieved from Ministry of Education website: https://www.moe.gov.sg/docs/default-source/document/education/programmes/national-education/ne-review-2016-2017-booklet.pdf

Singapore MOE. (2018b, November 20). Parliamentary replies: Teachers performance and appraisal. Retrieved 18 July 2018, from https://www.moe.gov.sg/news/parliamentary-replies/teachers-performance-and-appraisal

Singapore Teachers' Union. (2009, April–June). Performing up to expectations: The D grade. *The Mentor*, 4–5.

Singapore Teachers' Union. (2011, April–June). Impartiality of the School Climate Survey. *The Mentor*, 6.

Singapore Teachers' Union. (2014, Term 1). Performance & Expectations – The 'D' Grade. *The Mentor*, 4–5.

Singapore Teachers' Union. (2015, Term 4). Revisions to the Performance Review Process. *The Mentor*, 4.

Skaalvik, E. M., & Skaalvik, S. (2009). Does school context matter? Relations with teacher burnout and job satisfaction. *Teaching and Teacher Education*, *25*(3), 518–524. https://doi.org/10.1016/j.tate.2008.12.006

Skrla, L., Mckenzie, K. B., Scheurich, J. J., & Dickerson, K. L. (2011). Home-town values and high accountability: A Texas recipe for districtwide success in an urban school district. *Journal of Education for Students Placed at Risk*, *16*(2), 137–165. https://doi.org/10.1080/10824669.2011.559902

Snijders, T., & Bosker, R. (2011). *Multilevel Analysis: An Introduction To Basic And Advanced Multilevel Modeling* (2nd edition). London: Sage Publications Ltd.

Spaull, N. (2019). Who makes it into PISA? Understanding the impact of PISA sample eligibility using Turkey as a case study (PISA 2003–PISA 2012). *Assessment in Education: Principles, Policy & Practice*, *26*(4), 397–421. https://doi.org/10.1080/0969594X.2018.1504742

Speckesser, S., Runge, J., Foliano, F., Bursnall, M., Hudson-Sharp, N., Rolfe, H., & Anders, J. (2018). *Embedding Formative Assessment: Evaluation report and executive summary*. Retrieved from Education Endowment Foundation website: https://educationendowmentfoundation.org.uk/public/files/EFA_evaluation_report.pdf

Spillane, J. P. (2009). *Standards Deviation: How Schools Misunderstand Education Policy*. Cambridge, Massachusetts; London: Harvard University Press.

Springer, M. G., Ballou, D., Hamilton, L., Le, V.-N., Lockwood, J. R., McCaffrey, D. F., … Stecher, B. M. (2010). *Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching (POINT).* Nashville, Tennessee: National Center on Performane Incentives at Vanderbilt University.

Springer, M. G., Pane, J. F., Le, V.-N., McCaffrey, D. F., Burns, S. F., Hamilton, L. S., & Stecher, B. (2012). Team Pay for Performance: Experimental Evidence From the Round Rock Pilot Project on Team Incentives. *Educational Evaluation and Policy Analysis*, *34*(4), 367–390. https://doi.org/10.3102/0162373712439094

Stanfill, B. A., Villarreal, A. D., Medina, M. R., Esquivel, E. P., de la Rosa, E., & Duncan, P. A. (2016). Beyond the Culture of Corruption: Staying Ethical While Doing Business in Latin America. *Journal of Organizational Culture, Communications and Conflict*, *20*, 56.

Statistics Finland. (2018). *Finland in Figures 2018*. Helsinki: Statistics Finland.

Steinberg, M. P., & Sartain, L. (2015). Does Teacher Evaluation Improve School Performance? Experimental Evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy*, *10*(4), 535–572. https://doi.org/10.1162/EDFP_a_00173

Steiner-Khamsi, G. (2014). Cross-national policy borrowing: Understanding reception and translation. *Asia Pacific Journal of Education*, *34*(2), 153–167. https://doi.org/10.1080/02188791.2013.875649

Steinmetz, G. (2004). Odious Comparisons: Incommensurability, the Case Study, and "Small N's" in Sociology. *Sociological Theory*, *22*(3), 371–400. https://doi.org/10.1111/j.0735-2751.2004.00225.x

Stevenson, H. W., & Stigler, J. W. (1992). *The learning gap: Why our schools are failing and what we can learn from Japanese and Chinese education*. New York: Summit Books.

Stigler, J. W., & Hiebert, J. (1999). *The Teaching Gap: Best Ideas from the World's Teachers for Improving Education in the Classroom*. New York: The Free Press.

Stout, L. (2010). *Cultivating Conscience: How Good Laws Make Good People*. Princeton: Princeton University Press.

Strauss, V., & Sahlberg, P. (2015, February 12). Teach For Finland? Why it won't happen. *Washington Post*. Retrieved from https://www.washingtonpost.com/news/answer-sheet/wp/2015/02/12/teach-for-finland-why-it-wont-happen/

Strode, H. (1940). Sisu: A Word That Explains Finland. *New York Times*, p. SM4.

Survey Research Center. (2016). *Guidelines for Best Practice in Cross-Cultural Surveys* (4th edition). Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan.

Takayama, K., Waldow, F., & Sung, Y.-K. (2013). Finland Has it All? Examining the Media Accentuation of 'Finnish Education' in Australia, Germany and South Korea. *Research in Comparative and International Education*, *8*(3), 307–325. https://doi.org/10.2304/rcie.2013.8.3.307

Tan, C. (2017). Teaching critical thinking: Cultural challenges and strategies in Singapore. *British Educational Research Journal*, *43*(5), 988–1002. https://doi.org/10.1002/berj.3295

Tan, J. (2010). Education in Singapore: Sorting Them Out? In T. Chong (Ed.), *Management of success: Singapore revisited* (pp. 288–308). Singapore: Institute of Southeast Asian Studies.

Tan, J., & Gopinathan, S. (2000). Education reform in Singapore: Towards greater creativity and innovation? *NIRA Review*, *7*(3), 5–10.

Tan, K. P. (2010). The Transformation of Meritocracy. In T. Chong (Ed.), *Management of success: Singapore revisited* (pp. 272–287). Singapore: Institute of Southeast Asian Studies.

Tan, K. P. (2018). *Singapore: Identity, Brand, Power* (1st ed.). https://doi.org/10.1017/9781108561273

Tao, S. (2013). Why are teachers absent? Utilising the Capability Approach and Critical Realism to explain teacher performance in Tanzania. *International Journal of Educational Development*, *33*(1), 2–14. https://doi.org/10.1016/j.ijedudev.2012.01.003

Taras, V., Kirkman, B. L., & Steel, P. (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. https://doi.org/10.1037/a0018938

Teh, L., Hogan, D., & Dimmock, C. (2013). Knowledge Mobilisation and Utilisation in the Singapore Education System: The Nexus Between Researchers, Policy Makers and Practitioners. In B. Levin, J. Qi, H. Edelstein, & J. Sohn (Eds.), *The impact of research in education* (pp. 41–64). https://doi.org/10.1332/policypress/9781447306207.003.0003

Teivainen, A. (2019, July 26). Etla: Declining education level threatening economic growth in Finland. *Helsinki Times*. Retrieved from https://www.helsinkitimes.fi/finland/finland-news/domestic/16590-etla-declining-education-level-threatening-economic-growth-in-finland.html

Teng, A. (2019, September 6). Singapore families spent $1.4b on private tuition for kids last year. *The Straits Times*. Retrieved from https://www.straitstimes.com/singapore/education/families-spent-14b-on-private-tuition-for-kids-last-year-as-parents-fork-out

Teo, C. H. (2001, April). *Speech by RADM (NS) Teo Chee Hean, Minister for Education and Second Minister for Defence, at the Senior Education Officer Promotion Ceremony 2001*. Speech presented at the Westin Ballroom, Singapore. Retrieved from http://www.nas.gov.sg/archivesonline/speeches/view-html?filename=2001041401.htm

Teo, C. H. (2018, January 8). Written Reply to Parliamentary Question on the Civil Service Appraisal System. Retrieved 18 July 2018, from http://www.psd.gov.sg/press-room/parliamentary-replies/written-reply-to-parliamentary-question-on-the-civil-service-appraisal-system

Teo, Y. Y. (2018). *This is what inequality looks like*. Singapore: Ethos Books.

Tetlock, P. E. (1991). An Alternative Metaphor in the Study of Judgment and Choice: People as Politicians. *Theory & Psychology*, *1*(4), 451–475. https://doi.org/10.1177/0959354391014004

Thiel, C., Schweizer, S., & Bellmann, J. (2017). Rethinking side effects of accountability in education: Insights from a multiple methods study in four German school systems. *Education Policy Analysis Archives*, *25*, 93. https://doi.org/10.14507/epaa.25.2662

Thomas, P. L. (2013). Testing Capitalism: Perpetuating Privilege behind the Masks of Merit and Objectivity. *International Education Journal: Comparative Perspectives*, *12*(2), 85–103.

Thompson, M., Ellis, R., & Wildavsky, A. (1990). *Cultural theory*. Boulder, Colorado: Westview Press.

Tirri, K. (2014). The last 40 years in Finnish teacher education. *Journal of Education for Teaching*, *40*(5), 600–609. https://doi.org/10.1080/02607476.2014.956545

Tobin, J. J., Wu, D. Y. H., & Davidson, D. H. (1989). Preschool in Three Cultures: Japan, China, and the United States. New Haven, Connecticut: Yale University Press.

Tomlinson, P. (1989). Having it Both Ways: Hierarchical focusing as research interview method. *British Educational Research Journal*, *15*(2), 155–176. https://doi.org/10.1080/0141192890150205

Törnroos, J. (2005). Mathematics textbooks, opportunity to learn and student achievement. *Studies in Educational Evaluation*, *31*(4), 315–327. https://doi.org/10.1016/j.stueduc.2005.11.005

Tremewan, C. (1994). *The Political Economy of Social Control in Singapore*. London: Palgrave Macmillan.

Triandis, H. C. (2004). The Many Dimensions of Culture. *The Academy of Management Executive (1993-2005)*, *18*(1), 88–93.

Tucker, M. S. (Ed.). (2011). *Surpassing Shanghai: An agenda for American education built on the world's leading systems*. Cambridge, Massachusetts: Harvard Education Press.

Tulowitzki, P. (2016). Educational Accountability Around The Globe: Challenges and Possibilities for School Leadership. In J. I. Easley & P. Tulowitzki (Eds.), *Educational Accountability: International perspectives on challenges and possibilities for school leadership*. London; New York: Routledge.

Tyler, T. R. (2006). Psychological Perspectives on Legitimacy and Legitimation. *Annual Review of Psychology*, *57*(1), 375–400. https://doi.org/10.1146/annurev.psych.57.102904.190038

UNESCO. (2017). *Global Education Monitoring Report 2017/2018: Accountability in Education – Meeting Our Commitments*. Paris: UNESCO.

United Nations. (2017). *World Population Prospects: The 2017 Revision, Key Findings and Advance Tables* (No. ESA/P/WP/248). Retrieved from United Nations, Department of Economic and Social Affairs, Population Division website: https://population.un.org/wpp/Publications/Files/WPP2017_KeyFindings.pdf

Urdan, T. (2014). Understanding teacher motivation: What is known and what more there is to learn. In P. W. Richardson, S. A. Karabenick, & H. M. G. Watt (Eds.), *Teacher motivation: Theory and practice* (pp. 133–149). New York: Routledge.

Uslaner, E. M. (2008). Where You Stand Depends Upon Where Your Grandparents Sat: The Inheritability of Generalized Trust. *Public Opinion Quarterly*, *72*(4), 725–740. https://doi.org/10.1093/poq/nfn058

van Kersbergen, K., & van Waarden, F. (2004). 'Governance' as a bridge between disciplines: Cross-disciplinary inspiration regarding shifts in governance and problems of governability, accountability and legitimacy. *European Journal of Political Research*, *43*(2), 143–171. https://doi.org/10.1111/j.1475-6765.2004.00149.x

Vainikainen, M.-P., Thuneberg, H., Marjanen, J., Hautamäki, J., Kupiainen, S., & Hotulainen, R. (2017). How Do Finns Know? Educational Monitoring without Inspection and Standard Setting. In S. Blömeke & J.-E. Gustafsson (Eds.), *Standard Setting in Education: The Nordic Countries in an International Perspective* (pp. 243–259). https://doi.org/10.1007/978-3-319-50856-6_14

Velayutham, S., & Perera, M. H. B. (2004). The influence of emotions and culture on accountability and governance. *Corporate Governance: The International Journal of Business in Society*, *4*(1), 52–64. https://doi.org/10.1108/14720700410521961

Verger, A., & Parcerisa, L. (2017a). A Difficult Relationship. Accountability Policies and Teachers: International Evidence and Key Premises for Future Research. In *International Handbook of Teacher Quality and Policy* (pp. 241–254). https://doi.org/10.5281/zenodo.1256602

Verger, A., & Parcerisa, L. (2017b). *Accountability and education in the post-2015 scenario: International trends, enactment dynamics and socio-educational effects. Background paper prepared for the 2017/8 Global Education Monitoring Report – Accountability in education: Meeting our commitments* (No. ED/GEMR/MRT/2017/P1/1/REV). Retrieved from UNESCO website: https://unesdoc.unesco.org/ark:/48223/pf0000259559

Vieira, E., & Gomes, J. (2009). A comparison of Scopus and Web of Science for a typical university. *Scientometrics*, *81*(2), 587–600. https://doi.org/10.1007/s11192-009-2178-0

Viholainen, A., Partanen, M., Piiroinen, J., Asikainen, M., & Hirvonen, P. (2015). The role of textbooks in Finnish upper secondary school mathematics: Theory, examples and exercises. *Nordic Studies in Mathematics Education*, *20*, 157–178.

Visschedijk, M., Hendriks, R., & Nuyts, K. (2005). How to set up and manage quality control and quality assurance. *The Quality Assurance Journal*, *9*(2), 95–107. https://doi.org/10.1002/qaj.325

Vogell, H. (2011, July 26). Investigation into APS cheating finds unethical behavior across every level. *The Atlanta Journal-Constitution*. Retrieved from https://www.ajc.com/news/local/investigation-into-aps-cheating-finds-unethical-behavior-across-every-level/bX4bEZDWbeOH33cDkod1FL/

von der Embse, N. P., Pendergast, L. L., Segool, N., Saeki, E., & Ryan, S. (2016). The influence of test-based accountability policies on school climate and teacher stress across four states. *Teaching and Teacher Education*, *59*, 492–502. https://doi.org/10.1016/j.tate.2016.07.013

Vroom, V. H. (1964). *Work and motivation*. Malabar, Florida: Robert E Krieger.

Vroom, V. H., & Deci, E. L. (1992). *Management and motivation: Selected readings* (2nd ed.). Harmondsworth: Penguin Books.

Wagemaker, H. (2010). IEA: Globalization and assessment. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (pp. 663–668). https://doi.org/10.1016/B978-0-08-044894-7.01477-9

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

Walker, A., Qian, H., & Zhang, S. (2011). Secondary school principals in curriculum reform: Victims or accomplices? *Frontiers of Education in China*, *6*(3), 388–403. https://doi.org/10.1007/s11516-011-0138-y

Walker, A., & Dimmock, C. A. J. (Eds.). (2002). *School leadership and administration: Adopting a cultural perspective*. New York: RoutledgeFalmer.

Walker, T. D. (2017). *Teach like Finland: 33 simple strategies for joyful classrooms*. New York: W.W. Norton & Company.

Wallace, A. F. C. (1970). *Culture and personality* (2nd edition). New York: Random House.

Waller, W. (1932). *Sociology of Teaching*. New York: John Wiley & Sons.

Wallner, J. (2008). Legitimacy and Public Policy: Seeing Beyond Effectiveness, Efficiency, and Performance. *Policy Studies Journal*, *36*(3), 421–443. https://doi.org/10.1111/j.1541-0072.2008.00275.x

Walsh, P. (2006). Narrowed horizons and the impoverishment of educational discourse: Teaching, learning and performing under the new educational bureaucracies. *Journal of Education Policy*, *21*(1), 95–117. https://doi.org/10.1080/02680930500393492

Watt, H. M. G., Richardson, P. W., Klusmann, U., Kunter, M., Beyer, B., Trautwein, U., & Baumert, J. (2012). Motivations for choosing teaching as a career: An international comparison using the FIT-Choice scale. *Teaching and Teacher Education*, *28*(6), 791–805. https://doi.org/10.1016/j.tate.2012.03.003

Weaver, R. K. (2010). *But will it work? Implementation analysis to improve government performance* (No. 32). Retrieved from Brookings website: http://www.brookings.edu/~/media/research/files/papers/2010/2/implementation-analysis-weaver/02_implementation_analysis_weaver.pdf

Webb, R., Vulliamy, G., Häkkinen, K., & Hämäläinen, S. (1998). External inspection or school self-evaluation? A comparative analysis of policy and practice in primary schools in England and Finland. *British Educational Research Journal*, *24*(5), 539–556. Retrieved from Scopus.

Webb, R., Vulliamy, G., Hämäläinen, S., Sarja, A., Kimonen, E., & Nevalainen, R. (2004). A comparative analysis of primary teacher professionalism in England and Finland. *Comparative Education*, *40*(1), 83–108. https://doi.org/10.1080/0305006042000184890

Webber, D. J. (2010). School district democracy: School board voting and school performance. *Politics and Policy*, *38*(1), 81–95. https://doi.org/10.1111/j.1747-1346.2009.00229.x

Weiss, T. G. (2000). Governance, good governance and global governance: Conceptual and actual challenges. *Third World Quarterly*, *21*(5), 795–814. https://doi.org/10.1080/713701075

Weninger, C. (2016). A contextual critique of critical literacy: Freirean 'generative themes' and their impact on pedagogic practice. *Discourse: Studies in the Cultural Politics of Education*, 1–14. https://doi.org/10.1080/01596306.2016.1234432

West, M. R., & Woessmann, L. (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–F255. https://doi.org/10.1111/j.1468-0297.2010.02375.x

Westhorp, G. (2018). Understanding mechanisms in realist evaluation and research. In N. Emmel, J. Greenhalgh, A. Manzano, M. Monaghan, & S. Dalkin (Eds.), *Doing realist research* (pp. 41–57). Los Angeles: Sage.

Westhorp, G., Walker, B., Rogers, P., Overbeeke, N., Ball, D., & Brice, G. (2014). *Enhancing community accountability, empowerment and education outcomes in low and middle-income countries: A realist review*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–Value Theory of Achievement Motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Wiksten, S. (2018). *Teacher Training in Finland: A Case Study* (PhD thesis, UCLA). Retrieved from https://escholarship.org/uc/item/1kk7c3bz

Wilkinson, L., & Friendly, M. (2009). The History of the Cluster Heat Map. *The American Statistician*, *63*(2), 179–184. https://doi.org/10.1198/tas.2009.0033

Williams, M. J. (2017). *External validity and policy adaptation* (Working Paper No. BSG-WP-2017/019). Retrieved from Blavatnik School of Government website: https://www.bsg.ox.ac.uk/research/working-paper-series/external-validity-and-policy-adaptation

Williams, J. (2016). Quality assurance and quality enhancement: Is there a relationship? *Quality in Higher Education*, *22*(2), 97–102. https://doi.org/10.1080/13538322.2016.1227207

Wiseman, A. W. (2010). The Uses of Evidence for Educational Policymaking: Global Contexts and International Trends. *Review of Research in Education*, *34*(1), 1–24. https://doi.org/10.3102/0091732X09350472

Woessmann, L. (2005). The effect heterogeneity of central examinations: Evidence from TIMSS, TIMSS-Repeat and PISA. *Education Economics*, *13*(2), 143–169. https://doi.org/10.1080/09645290500031165

Woessmann, L. (2007). International evidence on school competition, autonomy, and accountability: A review. *Peabody Journal of Education*, *82*(2–3), 473–497. Retrieved from Scopus.

Woessmann, L. (2016). The Importance of School Systems: Evidence from International Differences in Student Achievement. *Journal of Economic Perspectives*, *30*(3), 3–32. https://doi.org/10.1257/jep.30.3.3

Woessmann, L., Luedemann, E., Schuetz, G., & West, M. R. (2009). *School Accountability, Autonomy and Choice Around the World*. Cheltenham; Northhampton, Massachusetts: Edward Elgar Publishing Ltd.

Wolters, C. A., & Daugherty, S. G. (2007). Goal structures and teachers' sense of efficacy: Their relation and association to teaching experience and academic level. *Journal of Educational Psychology*, *99*(1), 181–193. https://doi.org/10.1037/0022-0663.99.1.181

Wong, G., Greenhalgh, T., Westhorp, G., & Pawson, R. (2012). Realist methods in medical education research: What are they and what can they contribute? *Medical Education*, *46*(1), 89–96. https://doi.org/10.1111/j.1365-2923.2011.04045.x

Woolcock, M. (2018). Culture, Politics, and Development. In C. Lancaster & N. van de Walle (Eds.), *The Oxford Handbook of the Politics of Development* (Vol. 1, pp. 107–122). https://doi.org/10.1093/oxfordhb/9780199845156.013.11

World Bank. (2003). *World Development Report 2004: Making services work for poor people* (No. 26886; pp. 1–36). Retrieved from The World Bank website: http://documents.worldbank.org/curated/en/527371468166770790/World-Development-Report-2004-Making-services-work-for-poor-people-Overview

World Bank. (2019). World Development Indicators: GINI Index. Retrieved 26 February 2019, from DataBank website: https://databank.worldbank.org/data/reports.aspx?source=2&series=SI.POV.GINI&country=

WVS Association. (2012, June). *WVS 6 Official Questionnaire Version 4*. Retrieved from http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp

WVS Association. (n.d.). WVS Database: Fieldwork and sampling [Official web site]. Retrieved 11 August 2017, from World Values Survey website: http://www.worldvaluessurvey.org/WVSContents.jsp

Yle Uutiset. (2019, September 8). Teachers union calls for more consistent grading criteria in comprehensive schools. *Yle Uutiset*. Retrieved from https://yle.fi/uutiset/osasto/news/teachers_union_calls_for_more_consistent_grading_criteria_in_comprehensive_schools/10961151

You, Y. (2017). Comparing school accountability in England and its East Asian sources of 'borrowing'. *Comparative Education*, *53*(2), 224–244. https://doi.org/10.1080/03050068.2017.1294652

Yuan, K., Le, V.-N., McCaffrey, D. F., Marsh, J. A., Hamilton, L. S., Stecher, B. M., & Springer, M. G. (2013). Incentive Pay Programs Do Not Affect Teacher Motivation or Reported Practices: Results From Three Randomized Studies. *Educational Evaluation and Policy Analysis*, *35*(1), 3–22. https://doi.org/10.3102/0162373712462625

Zamarro, G., Hitt, C., & Mendez, I. (2016). *When Students Don't Care: Reexamining International Differences in Achievement and Non-cognitive Skills* (No. 2016–18). Retrieved from University of Arkansas, Department of Education Reform (EDRE) website: http://www.uaedreform.org/downloads/2016/10/when-students-dont-care-reexamining-international-differences-in-achievement-and-non-cognitive-skills.pdf

Zhao, Y. (2017). What works may hurt: Side effects in education. *Journal of Educational Change*, *18*(1), 1–19. https://doi.org/10.1007/s10833-016-9294-4

Zhao, Y. (2018). *What works may hurt: Side effects in education*. New York: Teachers College Press.

Zumbansen, P. (2012). Governance: An Interdisciplinary Perspective. In D. Levi-Faur (Ed.), *The Oxford Handbook of Governance* (pp. 83–96). Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199560530.013.0001