Management in Education Systems

Yue-Yi Hwa and Clare Leaver

Abstract

There is increasing interest in measuring management in schools. This paper discusses a popular measurement tool: the World Management Survey (WMS) for schools. Drawing on WMS data, secondary sources, and the recent literature on school management, we take stock of the WMS and make recommendations for its use in future research and policy. We conclude that the WMS remains a highly useful tool for its stated purpose—the standardized measurement of (a subset of) management practices within schools—and make two sets of recommendations. First, we encourage those seeking to benchmark management practices in schools to take a systems perspective by extending the WMS approach upwards into the education bureaucracy. Second, when measuring practices within schools, we recommend that researchers consider: how best to assess alignment across practices in the operations domain; the challenge of measuring student learning for monitoring and target-setting; and the context specificity of people management.

Management in Education Systems

Yue-Yi Hwa
Blavatnik School of Government, University of Oxford

Clare Leaver
Blavatnik School of Government, University of Oxford; CEPR

Management in Education Systems

Yue-Yi Hwa and Clare Leaver

10 January, 2021

1. Introduction

There is increasing interest in measuring management in schools, as researchers and policymakers try to understand which management practices can improve classroom teaching and student learning outcomes. Among economists, the World Management Survey for schools is increasingly seen as the gold standard for standardised comparisons of school management, with the instrument (and associated data) being used dozens of times since its inception in 2009.

In a spirit similar to the article by Scur et al. (2020) in this issue, our goal is to take stock of the WMS for schools and to offer recommendations for its future use in research and policy. We begin in Section 2 by setting out the objective of the WMS for schools and the management practices that were selected for inclusion using its monotonic 'more is better' scale. Drawing on data from the original survey waves, we also describe which types of schools score highly on which management practices. Next, in Section 3, we offer our reflections on the WMS for schools, based on this analysis of the original data, secondary sources, and a review of the recent literature on school management.[1] Finally, in Section 4 we present our main conclusions. We argue that the WMS for schools remains a highly useful tool for the purpose for which it was designed: benchmarking of specific school-level management practices. However, we encourage researchers to view management in schools from a systems perspective (and hence to expand the WMS approach 'upwards' into the education bureaucracy), and to think carefully about the importance of alignment, the challenge of measuring student learning, and the context-specific nature of people management when applying the WMS instrument within schools.

2. The World Management Survey for Schools

2.1. What is the objective of the WMS for schools, and which practices are included?

As Scur et al. (2020) note in their article in this issue, the goal of the WMS project is to systematically collect data on the usage of management practices at scale, while ensuring comparability over settings and time. For this reason, the original WMS was based on a *selected set* of management practices—those for which the WMS Team felt that there was a high degree of consensus *ex ante* (among consultants and industry experts) that the practice was 'good', in the sense of being a causal determinant of better firm performance. For these selected practices, it

---

[1] We focus on the substance of the WMS for schools—i.e. the selection of management practices —rather than the survey methodology (in which trained enumerators pose open-ended interview questions to school principals, and then numerically score their responses based on a detailed rubric; see Bloom et al., 2015).

was felt that adoption (of the practice) could be scored on a monotonic 'more is better' scale, with limited adoption being given a score of 1 and thorough adoption a score of 5.

The WMS for schools was developed in 2009, seven years after the original survey.[2] Again, the aim was not to be exhaustive but rather to select school management practices that were seen (by education practitioners and academics whom the team consulted) at that time to be important, and where adoption could plausibly be scored on a monotonic 'more is better' scale.

The management practices that were selected for inclusion in the WMS for schools fall under the same four domains used in the original survey: operations, monitoring, target setting, and people management. The 20 individual practices are listed in Appendix Table A1, together with a descriptor for thorough adoption of each practice—the top of the monotonic 'more is better' scale. It is clear from the table that the people management practices relate to individual teachers, whereas the operations, monitoring, and target-setting practices relate to the school as an organisation. For this reason, we follow the WMS convention and refer to the grouping of operations, monitoring, and target-setting practices as 'non-people' management.

2.2. Which types of schools score highly on which management practices?

Results from the initial waves of the WMS for schools are reported in Bloom et al. (2015) and are based on 1,849 schools, across 8 countries. The authors classify these schools into three groups depending on their source of funding and their degree of operational autonomy. As Table 1 shows there are: 1,237 'regular government schools' that are publicly funded and operate within the public regulatory framework; 483 'private schools' that are privately funded and operate using a school-specific charter; and 129 'autonomous government schools' that receive some public funding and have operational autonomy in at least one of curriculum content, teacher selection, and student admissions. These totals do, of course, mask differences across countries. Notably, the WMS for schools sample for Sweden does not include any private schools, and the sample for Italy does not include any autonomous government schools.

Bloom et al. (2015) explore whether there are differences in the overall management score—obtained by aggregating the individual scores (measured on the 1–5 scale) for each of the 20 practices in Table A1—by this tripartite classification of school type. They report that, among the OECD WMS countries (Canada, Germany, Italy, Sweden, the UK, and the US) and also for Brazil, autonomous government schools have higher overall management scores than regular government schools (see their Figure 3). This difference in mean scores persists even after including a variety of school and survey controls (see their Table 5).

We extend the analysis in Bloom et al. (2015) in two ways, first by *disaggregating* the overall management score into a non-people management score vs. a people management score, and second by exploring whether there are differences in the *variation* of management scores across school types and not simply the mean.[3] Since there are so few autonomous government schools, we pool this group with private schools, referring to this category simply as 'other schools'.

---

[2] The full survey instrument for the WMS for schools can be found online at www.worldmanagementsurvey.org. The Development WMS for schools (Lemos and Scur, 2016) is an expanded instrument for use in developing countries and is available at www.developingmanagement.org.

[3] Bloom et al. (2015) compare the overall management score across school types (see their Figure 5) but do not disaggregate into the different management domains. They also look at the variance of non-people and people management scores (see their Appendix Figure B1) but do not compare these distributions across school types.

Table 2 reports results for the non-people management score (the average of operations, monitoring, and target setting) and the people management score.[4] This disaggregation turns out to be revealing.

Non-people management. Pooling across countries, the average non-people management score is *higher* among regular government schools (2.43) than among other schools (2.31). The difference is small at 0.12 points (on a scale of 1–5) but is statistically significant at 1 percent. This pooled result runs counter to the finding in Bloom et al. (2015) for the overall management score and is clearly driven by a subset of countries, namely Canada, Italy, and the US. In the two non-OECD countries, Brazil and India, as well as Sweden, the average non-people management score is lower among regular government schools than in other schools. The difference in means is small in India, but larger in Sweden (0.34 points) and Brazil (0.22 points), and is statistically significant at 1 percent in all three countries.

Brazil and Sweden are also notable in terms of the variance of the non-people management score. In these two countries, non-people management scores are *less dispersed* among regular government schools than among other schools, as shown by the smaller standard deviations in Table 2. We can reject the null of equal variances at 5 percent (or less) for both countries. For the other 6 countries, the difference in standard deviation across school types is small and not statistically significant at conventional levels. In Figure 1, we present these results graphically by plotting the discrete density of the non-people management score in regular government schools (yellow bars) in each of the 8 WMS countries, overlaying the smoothed kernel density of the non-people management score in other schools (red plot) for comparison. For Brazil and Sweden, the discrete density is clearly less dispersed than the smoothed kernel density, but this is not the case for the other 6 countries.

People management. Pooling across countries, the average people management score is *lower* among regular government schools (1.92) than among other schools (2.18). The difference is large at 0.26 points and is again statistically significant at 1 percent. In contrast to the non-people management score, this finding is not driven by a subset of countries; the sign of the difference in means is the same—lower in regular government schools—across all 8 countries. The difference in means is largest in Brazil (0.72 points), Germany (0.47 points), Sweden (0.39 points) and Italy (0.28 points), and is statistically significant at 5 percent or less in every country except Canada and the UK.

In Brazil, Germany, Sweden, and Italy people management scores are also *less dispersed* among regular government schools than among other schools, as shown by the smaller standard deviations in Table 2. We can reject the null of equal variances at 5 percent or less for all 4 countries. In Canada, India, the UK, and the US, the standard deviation of the people management score is lower in regular government schools than in other schools, but the differences are small and (with the exception of the UK) not statistically significant at conventional levels. Figure 2 reproduces Figure 1 for the people management score. The plots for Brazil, Sweden, and Germany are particularly striking: people management scores are substantially less dispersed in regular government schools than in other schools.

---

[4] We have also disaggregated further into the four domains. Results for each of operations, monitoring, and target setting are qualitatively identical to those for the non-people management score reported in Table 2.

To sum up, there are two takeaways from this re-examination of the Bloom et al. (2015) data. First, when studying adoption patterns across types of school it is important to distinguish between non-people and people management practices. Although the mean overall management score is lower in regular government schools than other schools, in several OECD countries the mean non-people management score is actually higher in regular government schools than other schools. It is not true that management practices are universally worse (on the WMS coding) in schools that are publicly funded and that operate within the public regulatory framework, as the regular government schools in the WMS for schools dataset for Canada, Italy, and the US illustrate.

Second, when studying adoption patterns across types of school it is instructive to look at the variation in management practices, and not simply the mean. In Brazil and Sweden, both non-people and people management scores are lower and less dispersed in regular government schools than in other schools. In Germany and Italy, this is true for people management scores, although not for non-people management scores. The fact that management scores vary less in schools that have been granted less operational autonomy is, of course, intuitive and points to the importance of thinking about school management in the context of the wider education system. We return to this issue in Section 3.1 below.

3. Reflections on the WMS for Schools at 11

In their article in this issue, Scur et al. (2020) reflect on the WMS at 18. The WMS for schools is much younger but, over the last decade since its creation, much research has been undertaken on the topic of management in education and so reflection is also timely. In this section, we focus on two issues. The first is scope—whether the WMS approach should be broadened to bring in the wider education system; and the second is measurement—whether the management practices that were selected for inclusion in the WMS for schools in 2009 (on the monotonic 'more is better' scale) remain appropriate today, and whether there are new areas worthy of focus.

3.1. Bringing in the Education System

Bloom et al. (2015) report that the overall management score is, on average, higher in autonomous government schools than in regular public schools. Responding to this finding, they attempt to answer the question 'what explains the advantage of autonomous government schools?' by regressing the overall management score on a set of school type dummies (with regular government school as the omitted category) and a wide range of school and principal controls. The positive coefficient on autonomous government school remains more or less stable until controls are added for 'principal accountability' (the degree to which the principal is accountable to institutional stakeholders such as school external boards) and 'principal strategy' (the degree to which the principal communicates a well-articulated strategy for the school over the next five years). In these specifications, the coefficient halves in size (0.129 compared to 0.233 in the specification with no additional controls) but remains both economically and statistically significant. Differences in school and principal characteristics therefore account for some of the overall management score advantage of autonomous government schools, but certainly not all of it. Another way to think about the difference in management scores across school types is from a *systems perspective*.

3.1.1. What is a systems perspective?

Figure 3 provides a visual conceptualisation. The solid line depicts the management relationship within schools between principals and their teachers and students. This is the focus of the WMS for schools—the questions prompt principals to describe the (operations, monitoring, target-setting, and people management) practices that they use to manage teachers and students in their schools. The dotted line depicts the management relationship in the wider education system between education authorities (the Ministry of Education and its bureaucracy)[5] and schools (principals, teachers, and students). The WMS for schools does not focus on this management relationship, but there are frameworks that do. The RISE Accountability Framework, for instance, proposes five distinct 'design elements' for this relationship:[6]

- *Delegation*. Education authorities specify what they want done—the objectives or goals that schools and teachers should achieve.
- *Finance*. Education authorities set a budget for schools. This may include line items for teacher pay and bonuses or teacher pay may be determined centrally.
- *Support*. Education authorities design and deliver curricula, learning materials, and training, to schools and teachers.
- *Information*. Education authorities determine the information that will be collected from schools (e.g. national assessment systems, school inspections, EMIS reporting, etc.).
- *Motivation*. Education authorities specify what will happen to schools and teachers if the outcomes are good (relative to the delegation specified and based on the information available) versus if the outcomes are bad. These can be positive or negative and intrinsic or extrinsic (pecuniary) motivators.[7]

Evidently, the choices made under each design element will impact the ability of principals in government-funded and regulated schools to adopt management practices that score highly on the WMS coding. If, under the finance element, education authorities decide to retain control of teacher pay and allocate school budgets for non-people line items only, principals will have little scope to adopt practices that score highly on the question about 'attracting talent'. Similarly, if, under the motivation element, education authorities decide not to allow performance-based rewards (bonuses and/or promotion) and sanctions (dismissals or job reallocation/reposting), principals will have little scope to adopt practices that score highly on questions about 'rewarding high performers' and 'removing poor performers'. In short, principals in regular government schools do not operate in a vacuum but rather are nested within a hierarchy of management relationships within the wider public education system, and this may impose constraints that are

---

[5] Education authorities may also spread beyond the education ministry, e.g. when teacher salaries are administered by the finance ministry and hiring rules are controlled by the civil service authority (e.g. Huang et al., 2020). Also, not all education authorities are government entities. Certain non-state actors may have significant discretion over school management, such as the head offices of large private school chains such as Bridge International Academies, or commercial examination boards like Cambridge Assessment.

[6] The original RISE Accountability Framework built on the WDR (World Bank, 2004) and proposed four design elements (see Pritchett, 2015); 'support' was added to the list subsequently.

[7] Additionally, teachers and school principals may face affective, social, or reputational motivators that are beyond the formal control of education authorities, but are nevertheless contingent on their fulfilment of the delegation specified by education authorities (e.g. being shunned by colleagues for excessive absenteeism—or 'excessive' conscientiousness).

not present (or present to a lesser degree) in autonomous government schools and private schools.[8]

If principals in regular government schools do face such systemic constraints, then we would expect to see management scores that are uniformly lower than in other schools (because the explanation lies at the system level rather than school level). This is indeed what we find in 4 of the 8 WMS countries. Recall that in Brazil and Sweden, both non-people and people management scores are significantly lower and *significantly less dispersed* among regular government schools than among the group of other schools (autonomous government schools and/or private schools depending on the country). And in Germany and Italy, this is true for people management scores although not scores in the non-people domains.

3.1.2. How do system features vary across WMS countries?

To illustrate these arguments about the influence of the wider education system on school management, we compare the distribution of WMS scores from Section 2.2 with data on education systems from other cross-country sources. Specifically, we outline possible correspondences between people management scores and *one* element of the management relationship between education authorities and schools—delegation (or decentralisation of decision-making discretion)—as well as *one* aspect of the political economy of education systems—teacher unions. We do not discuss non-people management scores. Neither do we attempt to elucidate the full range of contextual features that can facilitate or constrain effective school management. As such, our analysis should be seen as exploratory and non-exhaustive.

As noted in Section 2.2, there is substantial between-country variation in average people management scores in regular government schools. Among the six OECD countries surveyed, these scores range from 1.83 in Italy to 2.69 in the UK. This variation may reflect, in part, differences in the degree to which decision-making authority is decentralised across levels of government. The more decentralised an area of management may be, the more flexibility school principals would have to improve management practices within their school, and the higher we would expect management scores to be. This is borne out when comparing the WMS for schools data with country-level data from the OECD's (2012) *Education at a Glance* on aspects of people management that span four of the five 'design elements' from the RISE framework: determining teachers' duties (delegation), fixing salary levels (finance), allocating resources for teacher professional development (support), and teacher dismissal (motivation). As shown in Table 3, decision-making discretion across these four areas is primarily centralised in low-scoring Italy, and fully decentralised to the school level in high-scoring England. Among the four countries with people management scores in the middle of the range, decision-making discretion over these four areas is held at the state and regional levels in Germany, and is distributed between the state, local, and school levels in Canada, Sweden, and the US.

The distribution of decision-making discretion may also explain some of the within-country differences in the variance of people management scores across school types. If some areas of management are decided by higher administrative levels for regular government schools, but are

---

[8] Bloom et al. (2015) plot the distribution of people management within countries and, on that basis, argue that 'national regulations are not homogenous completely binding on schools' (p. 663). However, their plots combine all school types. As we have shown, it is in the regular government schools where people management scores are significantly less dispersed.

decentralised for private and autonomous government schools, then we would expect to see more dispersion in the management scores of the latter. Again, there is suggestive evidence that this may be the case in some of the WMS countries. In Sweden, which has the largest difference in the variance of people management scores between regular and autonomous government schools (difference = 0.490 SD), procedures for teacher appraisal and teacher reward schemes are determined by subregional, municipal, or local authorities in regular government schools, but by school-level committees in autonomous government schools (OECD, 2015). In Brazil and Italy, which have the next largest differences in the variances of people management scores between school types, teachers in regular government schools are civil servants (OECD, 2015, on Brazil; Eurydice, 2020, on Italy). Accordingly, people management in regular government schools in these countries would be subject to civil service regulations that would not apply to other schools. At the other end of the spectrum, there is neither a significant nor substantive difference in the dispersion of people management scores across school types in the US (difference = 0.013 SD). This may be due in part to the fact that regulations for teacher certification and appraisal for both public and private schools vary across the 50 states (OECD, 2015), such that any differences in variance between school types are cross-cut by differences between states.

Besides the decentralisation of discretion between education authorities and schools, other relationships, actors, and dynamics within the wider education system may also constrain school principals' management practices (Pritchett, 2015). One such dynamic is the political economy of teacher unions. For example, teacher unions with wide membership bases may mobilise in support of labour protections that restrict governments' and school principals' leeway to implement the performance-based incentives at the centre of the WMS people management indicators. While there are no publicly accessible cross-country datasets on teacher union membership, International Labour Organization data on the proportion of employees across all industries who are covered by collective bargaining agreements are available for all of the WMS for schools countries besides India. Among these seven countries, collective bargaining agreements covered more than half of all employees in Brazil, Italy, Germany, and Sweden, where the people management scores were significantly lower and less dispersed in regular government schools than in other schools. In Canada, the UK, and the US, where people management scores differed less markedly across school types, collective bargaining agreements covered less than half of all employees. This is true for all years for which data are available, i.e. 2000–2016 (International Labour Organization, 2020).

Country-specific data on teacher unionisation paints a similar picture. In Sweden, where people management scores were far less dispersed in regular government schools than in autonomous schools, collective agreements cover 100 percent of teachers in regular government schools and 85 percent of teachers in autonomous schools. Notwithstanding these comparably high coverage levels, salary negotiations for regular government school teachers take place jointly at the national level, whereas there is fragmentation in the autonomous sector, with an estimated 3,000 collective agreements across autonomous schools (Education International, 2013b). Another variable in the political economy of teacher unions is how much power a union has to disrupt business-as-usual. This can vary considerably. In Brazil, where we see the largest difference in magnitude of people management scores between regular government and other schools, the public-sector education union has substantial veto power. In 2011, there were six different states or municipalities which saw more than 50 days of strike action, and a further four such states or municipalities in 2012—including the state of Bahia, which saw 115 days of strike action that year (Education International, 2013a). In contrast, teacher unions have far less power in the US.

A recent difference-in-difference analysis across 33 US states that had introduced mandatory collective bargaining regulations found that these laws did not lead to higher teacher salaries or larger education budgets, because these same laws typically restricted union power by instituting costly individual and collective penalties for strikes (Paglayan, 2019). 3.1.3. Moving forwards

The discussion above, while exploratory and non-exhaustive, does support the argument that the wider education system can influence the management practices in place in schools. We therefore see the extension of the WMS approach 'upwards' into the education bureaucracy as an important agenda for research. There has been some recent progress in this area. For example, RISE researchers recently adapted the Development WMS for schools to study the adoption of management practices within district education offices in Tanzania (Cilliers, Dunford, and Habriyamana, 2020), building on earlier work on district education offices in Zambia (Walter, 2018) and on the civil service in Ghana (Rasul, Rogger, and Williams, 2020). We encourage researchers seeking to understand school management and its relationship to student learning to take a systems perspective. Our specific suggestion is, alongside the WMS for schools, to administer surveys within the education ministry and its bureaucracy that aim to capture the nature of the management relationship between education authorities and their schools (e.g. via constructs built around the five design elements mentioned above), as well as political economy factors such as the influence of trade unions. A further practical suggestion is to build on existing data sources from other organisations and disciplines, for to reanalyse prior surveys using a WMS-style framework (e.g. Leaver, Lemos, and Scur, 2019), or to complement WMS surveys with data on other system features (e.g. the illustrative discussion in this section).

3.2. Measuring Management Practices in Schools

As noted in Section 2, the aim of the WMS for schools was to select school management practices for which adoption could plausibly (in 2009) be scored on a monotonic 'more is better' scale. In this section, we draw on the literature that has emerged over the last decade to assess whether this selection remains appropriate today. We focus on the domains of operations, monitoring and target-setting (for brevity, grouped together), and people management in turn and, in each case, highlight what we see as the key theme to have emerged from recent research.[9]

3.2.1. Operations: The importance of alignment

The unifying principle behind all four practices in the operations domain of the WMS for schools is alignment. Specifically, *standardisation of instructional processes* entails alignment across classrooms and between curriculum, instructional materials, and classroom practice. *Personalisation of instruction* and *data-driven transitions* entail alignment of, respectively, day-to-day classroom instruction and critical transition points with students' learning needs. Finally, *adopting best practices* entails alignment of instruction with effective pedagogical strategies.

If anything, there is even greater consensus today that alignment is an important determinant of school performance than there was back in 2009 when the WMS for schools was created (see, e.g., World Bank, 2018). This is particularly true for alignment of instruction with students' learning needs, a long-established principle in educational research (e.g. National Research

---

[9] A systematic literature review of school management is beyond the scope of this article. This section should be read as an (informal) review that draws on the economics and education literatures, guided by the authors' experiences working on the RISE Programme.

Council, 2000; Tomlinson et al., 2003; Vygotsky, 1978). Education interventions that are premised on such alignment to learning needs have shown significant learning gains in recent experimental evaluations, whether in the Teaching at the Right Level approach for foundational literacy and numeracy that was pioneered in India and is currently implemented in several African countries (Banerjee et al., 2016, 2017) or in the more recent Mindspark programme for computer-adaptive instruction (Muralidharan, Singh, and Ganimian, 2019). At the classroom level, alignment between curriculum, instructional materials, and assessments can greatly aid teachers in navigating and fulfilling the many priorities of classroom learning. Some education systems have badly misaligned curricula, exams, and instructional practices (Atuhurra and Kaffenberger, 2020), but interventions that introduce well-aligned instructional components can yield learning gains, such as Kenya's nationwide Tusome literacy programme (Piper et al., 2018).

It is also increasingly recognised that it is not just alignment within each management practice that can influence instructional effectiveness, but also alignment across these practices and with the wider education system as a whole (Pritchett, 2015). To illustrate, adoption of new 'best practice' teaching techniques into classrooms may do little to cultivate student learning unless they are compatible with individual student learning needs (e.g. Glewwe, Kremer, and Moulin, 2009).[10] Achieving alignment across management practices may be challenging, however, especially in low-capacity settings. For example, there may be a tension between *personalisation* and *standardisation* of instruction. Concurrently achieving high scores for both practices would require extensive educational resources, such as highly trained support staff who can offer tailored out-of-lesson remedial instruction—this may be a reality in some high-income settings, but not in countries where large shares of primary school teachers are yet to master the content that they are supposed to teach (Bold et al., 2017). In such low-capacity settings, school principals and education authorities face a twin challenge: aligning across management practices and prioritising between these practices.

3.2.2. Monitoring and target setting: The challenge of measuring student learning

Most of the management practices in the monitoring and targeting domains focus on process, i.e. 'how', rather than 'what', to monitor and target. The descriptor for *target balance* does, however, indicate a judgement that it is good management practice to include targets based on (absolute and value-added) measures of student learning.[11]

Such emphasis on student learning, while fundamental to the purpose of schooling, cannot be taken for granted in educational management. For example, the State Report Cards 2016–2017 in India's District Information System for Education reported on 977 distinct figures, none of which was a direct measure of student learning (Pritchett, 2018). Nonetheless, frequent monitoring of student, school, and classroom progress has consistently been identified as an important management process in educational effectiveness research (see Reynolds et al., 2015, for a review). More recently, an analysis of PISA data for 59 countries from 2000 to 2015 found

---

[10] It is also worth noting that there has been pushback against the idea of universal best practices that can improve education across all contexts (e.g. Coffield, 2012, critiquing McKinsey's oft-cited education reports; Sellar and Lingard, 2013, on the OECD's influence on educational governance). These critiques are supported by arguments in public policy and development studies about the importance of identifying context-specific mechanisms for achieving desired changes, especially when these changes require shifts in people's behaviour or decision-making (Bates and Glennerster, 2017; Cartwright and Hardie, 2012; Monaghan and King, 2018; Pawson and Tilley, 1997; Williams, 2020).

[11] In the newer Development WMS instrument, student learning outcomes are also explicitly included in *performance tracking*, under monitoring (Lemos and Scur, 2016).

that expanding the use of standardised tests to compare learning outcomes across students or across schools was positively and significantly associated with student achievement (Bergbauer, Hanushek, and Woessmann, 2018).

In addition to providing valuable information to school principals and teachers, some approaches for monitoring student outcomes can themselves reinforce students' mastery of new learning. Cognitive science research has found that retrieval practice—trying to retrieve information from one's memory, ideally with feedback about the accuracy of retrieval—can reinforce long-term retention of information (Roediger and Butler, 2011; see also Dunlosky et al., 2013, on practice testing; and Ericsson, Krampe, and Tesch-Römer, 1993, on deliberate practice for the development of expertise). Similarly, meta-analyses of interventions that promote formative assessment in K–12 classrooms found a weighted mean effect size of 0.20 on student achievement (Kingston and Nash, 2011). That said, a recent randomised controlled trial of a formative assessment programme in primary schools in Haryana, India, found that the programme did not improve test scores (Berry et al., 2020), partly because the assessment was treated as an administrative task and was not used to provide feedback to students, nor to inform teaching practices.

In short, there is growing evidence that monitoring/targeting student learning is associated with student achievement, alongside increasing recognition that this area of management practice brings complex challenges.

One such challenge is administration. Bloom and Van Reenen (2007) note in their article introducing the WMS that firms may choose not to adopt certain productivity-boosting management practices if the productivity gains do not offset the costs of adoption. This trade-off certainly applies to the costs that would be incurred in principals' and teachers' time in order to monitor school performance with the frequency and formality recommended in the WMS for schools. In the 2018 Teaching and Learning International Survey (TALIS), nationally representative samples of lower secondary school teachers across 48 countries reported spending an average of 8.2 percent of classroom lesson time on 'general administrative tasks', such as recording attendance or distributing forms (OECD, 2019). One study of teacher accountability in the US found that, under the No Child Left Behind Act, teachers in some schools had to submit 'up to sixty pages of documentation each week' (p. 366) for administrative monitoring of compliance with curricular standards (Holloway and Brass, 2018).

A further challenge is measurement. Conducting student assessments that are valid, reliable, and appropriately calibrated is a technically demanding and potentially resource-intensive task (Koretz, 2008). Yet high-quality student assessment is both a precondition of meaningful monitoring and a casualty when the incentives embedded in monitoring systems go awry (whether those incentives are pecuniary, reputational, or otherwise). In the WMS for schools, the quality of student assessment is mentioned in the operations domain under *data-driven transitions,* which notes that student transitions should be 'supported by formative assessment tightly linked to learning expectations'. However, the monitoring and target-setting domains appear to take assessment quality for granted.

For a learning assessment to be valid, it has to measure the knowledge and skills that it purports to test. To give an egregious but actual example, if a test item is meant to measure students' knowledge of place values, then it would be inadvisable to ask them to identify the number in the

units position of the cubed root of 531,441—since many students who have mastered place values would not have mastered the far more difficult area of cubed roots (Burdett, 2016). In light of validity concerns, some education experts recommend replacing standardised tests with performance assessments that demonstrate real-world skills, such as portfolios or capstone projects that are assessed using rubrics rather than percentage-correct or item-response theory scoring (Guha et al., 2018). Using test scores as a measure of schools' or teachers' performance introduces another dimension of validity, i.e. whether the measures accurately reflect teachers' and schools' contribution to learning, rather than reflecting factors beyond their control. In the WMS for schools, attaining a top score on *target balance* entails setting targets for 'both absolute and value-added measures of student outcomes'. Value-added measures that take into account students' prior performance (and sometimes home backgrounds) can be more appropriate indicators of teacher and school performance than absolute scores, although these measures have their own share of issues (Amrein-Beardsley, 2014; Bitler et al., 2019; see also recommendations from the AERA, 2015, and the ASA, 2014).

Besides validity, high-quality student assessments also need to be reliable. In addition to reliability at the level of item and test design, a potential issue when assessment data is used in management processes is corruption of test score data when agents respond to self-serving incentives, as observed in Campbell's (1979) and Goodhart's (1984) laws. Test score manipulation can range from focusing on children whose test scores hover near accountability thresholds (e.g. Booher-Jennings, 2005, on Texas), to inflationary leniency in grading (e.g. Hinnerich and Vlachos, 2017, on Sweden), to outright cheating (e.g. Buckner and Hodges, 2016, on Jordan and Morocco; Patrick et al., 2018, on Atlanta). Statistical analyses of assessment data have found answering patterns indicating test score manipulation in Sweden (Diamond and Persson, 2016), the US (Dee et al., 2019; Jacob and Levitt, 2003), southern Italy (Angrist, Battistin, and Vuri, 2017), Mexico (Martinelli et al., 2018), India (Johnson and Parrado, 2020; Singh, 2020), Indonesia (Berkhout et al., 2020), and on a regional assessment of southern and eastern African countries (Gustafsson and Nuga Deliwe, 2017). Test score manipulation can compromise the achievement of learning targets not only because it redirects student, teacher, and administrator effort toward manipulation rather than learning, but also because managers may make counterproductive decisions when they treat manipulated data as if it were accurate.

Finally, high-quality student assessments must also be appropriately calibrated. The learning crisis in many low- and middle-income countries is such that actual student learning falls far below the levels that many policymakers and school leaders can comfortably acknowledge. A recent evaluation of a computer-adaptive instruction programme in India found that grade 6 students in the treatment group were, at baseline, an average of 2.5 years behind the mathematics curriculum (Muralidharan, Singh, and Ganimian, 2019). This is not an isolated result. For example, across 51 developing countries in the Demographic and Health Surveys, only half of all women aged 25–34 who had completed grade 6 (but had not attended secondary school) could read a simple sentence such as 'Parents love their children' in a language of their choosing (Pritchett and Sandefur, 2017). When the official curriculum is far above actual learning levels, emphasising curricular completion is likely to make students fall further behind learning targets over time, rather than boosting achievement (Pritchett and Beatty, 2012).

3.2.3. People management: Context matters

People management is the one domain where there is perhaps less consensus on the management practices that should be included in the WMS for schools with its monotonic 'more is better' scale. One factor, discussed in Section 3.1 above, is that principals in regular public schools may have limited discretion to adopt the high scoring practices (flexible compensation, promotion, hiring, firing, and retention) due to system-level constraints. To the extent that this is true, the people management domain may not be capturing management *by* a school principal but rather management *of* a school by education authorities. There is also accumulating evidence that for some practices, notably performance pay under *rewarding high performers*, more may enhance student learning in some contexts but not in others.

To focus in on teacher compensation, two recent reviews (Breeding, Béteille, and Evans, 2020; Glewwe and Muralidharan, 2015) and one recent meta-analysis (Pham, Nguyen, and Springer, 2020) have found that, averaging over studies, performance pay schemes have positive treatment effects on student learning outcomes, although the individual impacts vary considerably in size and significance.

One possible explanation for this variation is that the interplay between extrinsic incentives and intrinsic motivation depends on context. Frey (1997) argues that extrinsic incentives are only likely to crowd out intrinsic motivation when the agent has high levels to begin with. Experimental studies support this hypothesis. Teacher performance pay schemes have raised student achievement *and* been viewed positively by teachers in a number of less well-functioning education systems (where a sizeable fraction of teachers fail to perform even 'the basics' such as turning up for work and being present in class) including: India (Muralidharan and Sundararaman, 2011a, 2011b), Tanzania (Mbiti et al., 2019; Mbiti and Schipper, 2020), and Rwanda (Leaver et al., 2020). Relatedly, Deci and Ryan (1980) suggest that extrinsic incentives may inhibit intrinsic motivation to the extent that they are felt to constrain autonomy, but may reinforce intrinsic motivation when they are felt to affirm competence. The Tanzania performance pay scheme also supports this hypothesis. McAlpine et al. (2018) suggest that one reason why teachers viewed the Tanzanian scheme favourably was that they saw it as a status-raising social affirmation, because the scheme involved visits from the external implementers.

In other settings, however, more thorough adoption of the practice of rewarding high performers has not been found to be unequivocally 'better'. Three recent studies illustrate. A performance pay scheme for mathematics teachers in Uganda had no effect on attendance, achievement, or attainment in schools without mathematics books—although in schools with the appropriate books, it raised teacher attendance slightly and improved student performance on test items covered in the books (Gilligan et al., 2019). A test-score based performance pay scheme in Pakistan raised student test scores, but skewed lesson time toward test preparation and negatively affected student socio-emotional development (Andrabi and Brown, forthcoming). And the introduction of high-stakes teacher evaluation in US states raised the quality of new teachers (as measured by the selectivity of their undergraduate institutions), but also raised the likelihood that schools would have unfilled teacher vacancies (Kraft et al., 2020).

Recent research has also shown that there are other management practices, beyond the formal incentives emphasised in the WMS for schools, that may be important in driving better performance: namely, responding to—and shaping—social and professional norms.

In some cases, norms can make school management harder. For example, social status hierarchies in India and Indonesia have influenced teachers' perceptions of who can legitimately evaluate their work (Broekman, 2015; Gaduh et al., 2020; Narwana, 2015). In an insider study, Mizel (2009) finds that teachers in Bedouin schools in Israel tended to prioritise accountability to the tribal sheikh over accountability to the education ministry, even though the schools were part of the state education system. Divergence between norms and official expectations can vary considerably across contexts:, Sabarwal and Abu-Jawdeh (2018) find that over 75 percent of teachers surveyed in Argentina, Senegal, and Tajikistan believe that it is acceptable to be absent from class when the teacher has completed the curriculum, providing they leave work for the students to do, and/or if they are engaged in tasks that serve the community; whereas fewer than 25 percent of teachers in Myanmar and Pakistan expressed similar beliefs.

On the other hand, norms can also constructively reinforce school management practices. In Vietnam, which is currently an outlier with high student learning levels despite relatively low income levels, officially mandated processes for monitoring teacher performance are complemented by strong professional ethics among teachers, as well as high levels of societal attention to education (McAleavy, Tran, and Fitzpatrick, 2018). Crucially, different norm orientations may support different management configurations. Singapore's socio-cultural context of top-down management and meritocracy supports the use of highly structured teacher career ladders and performance bonuses; whereas Finland's sociocultural context of egalitarianism and individual civic responsibility supports the deployment of carefully selected, expertly trained, and highly autonomous teachers—such that some Finnish teachers say that introducing people management practices that score highly on the WMS coding would have a negative impact on teaching and learning (Hwa, 2019). This variability in teacher norms and, accordingly, in appropriate management practices implies a need for caution in assuming a positive relationship between WMS people management practices and school performance in a given context.

3.2.4. Moving forwards

This brief review shows that the WMS for schools remains a highly relevant tool today, more than a decade after its inception. We fully support its use in schools (ideally alongside similar surveys in the education bureaucracy) but have emphasised three issues for researchers to bear in mind when doing so.

The first issue is the importance of alignment. The principle that unites the management practices selected for inclusion in the operations domain is alignment and, if anything, this is seen as more important today, particularly the alignment of classroom instruction with students' learning needs (personalisation of instruction). Researchers may want to think more about how to capture alignment *across* management practices—e.g. by probing thoroughly whether practices that standardise instructional strategies and ensure consistency across classrooms also allow for personalisation to student learning needs (recognising that this may not always be feasible in low-capacity settings).

The second issue is the challenge of measuring student learning. Although the practices included under the domains of monitoring and target setting are well supported by recent evidence, the WMS for schools appears to take the quality of student assessments, on which 'balanced targets' should be based, for granted. Researchers may want to probe whether there are management

practices in place to ensure that the assessments used to measure student learning for target setting are valid, reliable, and appropriately calibrated.

The final issue is the context-specificity of people management. There are two points here. The first is the systems-perspective point: in some school and country contexts, the principal being surveyed may not have decision-making discretion to adopt high-scoring (on WMS coding) people management practices. The second is that for some practices under this domain, notably rewarding high performing teachers with performance pay, 'more' may be 'better' in some school and country contexts but not in others. Researchers may want to reflect on whether *in their particular context* it makes sense to administer the questions in the people management domain in their current form: do they adequately capture management *by* school principals and, further, is it appropriate to code the answers on a monotonic 'more is better' scale?

4. Concluding Remarks

We began this article by summarizing the objective of the WMS for schools. Then, drawing on analysis of the original WMS data, secondary sources, and a review of the recent literature on school management, we took stock of its usefulness as a tool for future research and policy.

Our view is that the WMS for schools remains a highly useful tool today for its stated purpose: the standardised measurement of (a subset of) management practices within schools. The practices selected for inclusion are well-supported by the evidence that has emerged over the past 11 years, especially for the operations, monitoring, and target-setting domains. We did, however, make two sets of recommendations for its use going forwards.

First, we encourage researchers and policymakers seeking to benchmark management practices in schools to take a systems perspective. We showed that, in four of the eight WMS countries, management scores were lower and less dispersed in regular government schools than in other schools with more autonomy, particularly in the people management domain. In these four countries, decision-making on key issues (e.g. teacher duties, salaries, professional development, and dismissals) is more centralised, and teacher unions negotiate more pervasive collective bargaining agreements on pay and conditions, relative to the other WMS countries. Hence, the lower and less dispersed management scores in regular government schools may have been a reflection of the system, rather than choices made by individual school principals. Extending the WMS approach 'upwards' into the education bureaucracy—to capture the nature of the management relationship between education authorities and their schools, as well as the influence of other stakeholders—is an important agenda for future research. Such work seems particularly timely in view of the fact that recent high-profile, cross-cutting management interventions in schools show no impact on student learning (Bedoya et al., 2020; Muralidharan and Singh, 2020). These interventions followed global 'best practice' but may have failed to account for pressures and constraints arising in their specific system context.

Second, turning to the measurement of management practices *within* schools, we encourage researchers to think about how best to assess alignment across practices in the operations domain and the challenge of measuring student learning for monitoring and target-setting. For the domain of people management, where there is less consensus, researchers may want to reflect on whether, in their particular county context, the questions in the people management

domain adequately capture management by school principals, and whether it is in fact appropriate to code the answers to all questions on a monotonic 'more is better' scale.

The WMS for schools was not conceived to provide causal explanations or to make practical policy prescriptions, but rather as a device to benchmark management practices in schools over time and countries. We feel it remains useful for this purpose, particularly with the recommendations suggested in this article. Such measurement can help as part of the process of understanding which management practices improve classroom teaching and student learning outcomes, alongside (or embedded in) careful quantitative and qualitative studies of school management in specific settings.

References

AERA (American Educational Research Association). (2015). AERA Statement on Use of Value-Added Models (VAM) for the Evaluation of Educators and Educator Preparation Programs. *Educational Researcher*, *44*(8), 448–452. https://doi.org/10.3102/0013189X15618385

Amrein-Beardsley, A. (2014). *Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability*. Routledge.

Andrabi, T., & Brown, C. (forthcoming). Subjective versus Objective Incentives and Employee Productivity. RISE Working Paper Series.

Angrist, J. D., Battistin, E., & Vuri, D. (2017). In a Small Moment: Class Size and Moral Hazard in the Italian Mezzogiorno. *American Economic Journal: Applied Economics*, *9*(4), 216–249. https://doi.org/10.1257/app.20160267

ASA (American Statistical Association). (2014). ASA Statement on Using Value-Added Models for Educational Assessment. https://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf

Atuhurra, J., & Kaffenberger, M. (2020). System (in)coherence: Quantifying the alignment of primary education curriculum standards, examinations, and instruction in two East African countries. RISE Working Paper 20/057. https://doi.org/10.35489/BSG-RISE-WP_2020/057

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukherji, S., Shotland, M., & Walton, M. (2016). *Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India*. National Bureau of Economic Research. Working Paper No. 22746. https://doi.org/10.3386/w22746

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., & Walton, M. (2017). From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. *Journal of Economic Perspectives, 31*(4), 73–102. https://doi.org/10.1257/jep.31.4.73

Bates, M. A., & Glennerster, R. (2017). The Generalizability Puzzle. *Stanford Social Innovation Review*. https://ssir.org/articles/entry/the_generalizability_puzzle

Bedoya, J., de Hoyos, R., Romero, M., & Yanez-Pagan, M. (2020). School management, grants, and test scores: Experimental Evidence from Mexico. Manuscript in preparation.

Bergbauer, A., Hanushek, E. & Woessmann, L. (2018). Testing. RISE Working Paper 18/025. https://doi.org/10.35489/BSG-RISE-WP_2018/025

Berkhout, E., Pradhan, M., Rahmawati, Suryadarma, D., & Swarnata, A. (2020). From Cheating to Learning: An Evaluation of Fraud Prevention on National Exams in Indonesia. RISE Working Paper 20/046. https://doi.org/10.35489/BSGRISEWP_2020/046

Berry, J., Kannan, H., Mukherji, S., & Shotland, M. (2020). Failure of frequent assessment: An evaluation of India's continuous and comprehensive evaluation program. *Journal of Development Economics*, *143*, 102406. https://doi.org/10.1016/j.jdeveco.2019.102406

Bitler, M., Corcoran, S., Domina, T., & Penner, E. (2019). *Teacher Effects on Student Achievement and Height: A Cautionary Tale* (Working Paper No. 26480). National Bureau of Economic Research. https://doi.org/10.3386/w26480

Bloom, N., Lemos, R., Sadun, R., & Van Reenen, J. (2015). Does Management Matter in Schools? *The Economic Journal*, 125, 647-674.

Bloom, N., & Van Reenen, J. (2007). Measuring and Explaining Management Practices Across Firms and Countries. *The Quarterly Journal of Economics,122*(4), 1351–1408. https://doi.org/10.1162/qjec.2007.122.4.1351

Bold, T., Filmer, D., Martin, G., Molina, E., Rockmore, C., Stacy, B., Svensson, J., & Wane, W. (2017). What Do Teachers Know and Do? Does It Matter? Evidence from Primary Schools in Africa. World Bank Policy Research Working Papers. No. 7956. https://doi.org/10.1596/1813-9450-7956

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

Breeding, M. E., Béteille, T., & Evans, D. K. (2020). *Teacher Pay-for-Performance (PFP) Systems: What works? Where? And how?* Manuscript in preparation.

Broekman, A. (2015). The Effects of Accountability: A Case Study from Indonesia. In J. Evers & R. Kneyber (Eds.), *Flip the System: Changing Education from the Ground Up*. Taylor & Francis. https://doi.org/10.4324/9781315678573-7

Buckner, E., & Hodges, R. (2016). Cheating or cheated? Surviving secondary exit exams in a neoliberal era. *Compare: A Journal of Comparative and International Education*, *46*(4), 603–623. https://doi.org/10.1080/03057925.2015.1088379

Burdett, N. (2016). The Good, the Bad, and the Ugly - Testing as a Key Part of the Education Ecosystem. RISE Working Paper 16/010. https://doi.org/10.35489/BSG-RISE-WP_2016/010.

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

Cartwright, N., & Hardie, J. (2012). *Evidence-Based Policy: A Practical Guide to Doing It Better*. Oxford University Press.

Cilliers, J., Dunford, J., & Habyarimana, J. (2020). What do local government education managers do to boost learning outcomes? Manuscript in preparation.

Coffield, F. (2012). Why the McKinsey reports will not improve school systems. *Journal of Education Policy*, *27*(1), 131–149. https://doi.org/10.1080/02680939.2011.623243

Deci, E. L., & Ryan, R. M. (1980). The Empirical Exploration of Intrinsic Motivational Processes. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 13, pp. 39–80). https://doi.org/10.1016/S0065-2601(08)60130-6

Dee, T. S., Dobbie, W., Jacob, B. A., & Rockoff, J. (2019). The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations. *American Economic Journal: Applied Economics*, *11*(3), 382–423. https://doi.org/10.1257/app.20170520

Diamond, R., & Persson, P. (2016). *The Long-term Consequences of Teacher Discretion in Grading of High-stakes Tests*. National Bureau of Economic Research. Working Paper No. 22207. https://doi.org/10.3386/w22207

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Education International. (2013a, July 16). *Brazil: A 2012 EI country study report*. Retrieved September 2, 2020, from https://www.worldsofeducation.org/en/woe_homepage/woe_detail/4749/brazil-a-2012-ei-country-study-report.

Education International. (2013b, July 25). *Sweden: A 2012 EI country study report*. Retrieved September 2, 2020, from https://www.ei-ie.org/en/woe_homepage/woe_detail/4764/sweden-a-2012-ei-country-study-report

Ericsson, K., Krampe, R., & Tesch-Römer, C. (1993). The Role of Deliberate Practice in the Acquisition of Expert Performance. *Psychological Review*, *100*(3), 363–406. https://doi.org/10.1037/0033-295X.100.3.363

Eurydice. (2020, April 16). *Italy: Conditions of Service for Teachers Working in Early Childhood and School Education*. Retrieved September 3, 2020, from https://eacea.ec.europa.eu/national-policies/eurydice/content/conditions-service-teachers-working-early-childhood-and-school-education-38_en.

Frey, B. S. (1997). *Not just for the money: An economic theory of personal motivation* Edward Elgar.

Gaduh, A., Pradhan, M., Priebe, J., & Susanti, D. (2020). Scores, Camera, Action? Incentivizing Teachers in Remote Areas. RISE Working Paper 20/035. https://doi.org/10.35489/BSG-RISE-WP_2020/035.

Gilligan, D. O., Karachiwalla, N., Kasirye, I., Lucas, A. M., & Neal, D. (2019). Educator Incentives and Educational Triage in Rural Primary Schools. *Journal of Human Resources* 1118-9871R2. https://doi.org/10.3368/jhr.57.1.1118-9871R2

Glewwe, P., Kremer, M., & Moulin, S. (2009). Many Children Left Behind? Textbooks and Test Scores in Kenya. *American Economic Journal: Applied Economics*, 1(1), 112–135. https://doi.org/10.1257/app.1.1.112

Glewwe, P., & Muralidharan, K. (2015). Improving School Education Outcomes in Developing Countries. RISE Working Paper 15/001. https://doi.org/10.35489/BSG-RISE-WP_2015/001.

Goodhart, C. A. E. (1984). *Monetary Theory and Practice: The UK Experience*. Macmillan Education.

Guha, R., Wagner, T., Darling-Hammond, L., Taylor, T., & Curtis, D. (2018). The promise of performance assessments: Innovations in high school learning and college admission. Learning Policy Institute. https://learningpolicyinstitute.org/product/promise-performance-assessments-report

Gustafsson, M., & Nuga Deliwe, C. (2017). Rotten apples or just apples and pears? Understanding patterns consistent with cheating in international test data (Working Paper No. 17/2017). Stellenbosch University, Department of Economics. https://econpapers.repec.org/paper/szawpaper/wpapers293.htm

Hinnerich, B. T., & Vlachos, J. (2017). The impact of upper-secondary voucher school attendance on student achievement. Swedish evidence using external and internal evaluations. *Labour Economics*, *47*, 1–14. https://doi.org/10.1016/j.labeco.2017.03.009

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

Huang, A.R., Revina, S., Fillaili, R., & Akhmadi. (2020). The Struggle to Recruit Good Teachers in Indonesia: Institutional and Social Dysfunctions. RISE Working Paper 20/041. https://doi.org/10.35489/BSG-RISE-WP_2020/041.

Hwa, Y.-Y. (2019). *Teacher accountability policy and sociocultural context: A cross-country study focusing on Finland and Singapore* [Doctoral dissertation, University of Cambridge]. UoC Repository. https://doi.org/10.17863/CAM.55349

International Labour Organization. (2020). Collective bargaining coverage rate (%) - Annual. ILOSTAT [database]. Available from https://ilostat.ilo.org/data/.

Jacob, B. A., & Levitt, S. D. (2003). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *The Quarterly Journal of Economics*, *118*(3), 843–877. https://doi.org/10.1162/00335530360698441

Johnson, D., & Parrado, A. (2020, July). Assessing the Assessments: Taking Stock of Learning Outcomes Data in India. RISE Online Presentation Series. https://riseprogramme.org/sites/default/files/inline-files/Johnson_Assessing%20the%20assessments_RISE%20version_v2.pdf

Kingston, N., & Nash, B. (2011). Formative Assessment: A Meta-Analysis and a Call for Research. *Educational Measurement: Issues and Practice*, *30*(4), 28–37. https://doi.org/10.1111/j.1745-3992.2011.00220.x

Koretz, D. M. (2008). *Measuring up: What educational testing really tells us.* Harvard University Press.

Kraft, M. A., Brunner, E. J., Dougherty, S. M., & Schwegman, D. J. (2020). Teacher accountability reforms and the supply and quality of new teachers. *Journal of Public Economics, 188*, 104212. https://doi.org/10.1016/j.jpubeco.2020.104212

Leaver, C., Lemos, R., & Scur, D. (2019). Measuring and Explaining Management in Schools: New Approaches Using Public Data. RISE Working Paper 19/033. https://doi.org/10.35489/BSG-RISE-WP_2019/033

Leaver, C., Ozier, O., Serneels, P., & Zeitlin, A. (2020). Recruitment, effort, and retention effects of performance contracts for civil servants: Experimental evidence from Rwandan primary schools. RISE Working Paper 20/048. https://doi.org/10.35489/BSG-RISE-WP_2020/048

Lemos, R., & Scur, D. (2016). Developing Management: An Expanded Evaluation Tool for Developing Countries. 16/007. https://doi.org/10.35489/BSG-RISE-WP_2016/007

Martinelli, C., Parker, S. W., Pérez-Gea, A. C., & Rodrigo, R. (2018). Cheating and Incentives: Learning from a Policy Experiment. *American Economic Journal: Economic Policy*, *10*(1), 298–325. https://doi.org/10.1257/pol.20150066

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

Mbiti, I., & Schipper, Y. (2020). *Teacher and Parental Perceptions of Performance Pay in Education: Evidence from Tanzania.* Research on Improving Systems of Education (RISE). https://doi.org/10.35489/BSG-RISE-WP_2020/037

McAleavy, T., Tran, T. H., & Fitzpatrick, R. (2018). Promising practice: Government schools in Vietnam. Education Development Trust. https://files.eric.ed.gov/fulltext/ED588856.pdf

McAlpine, K., Kagucia Omesa, N., Semwene, J., & Mbise, A. (2018). A twist on performance theory. How rewarding individual teachers may promote whole school improvement. Twaweza; Dar es Salaam, Tanzania.

Mizel, O. (2009). Accountability in Arab Bedouin Schools in Israel: Accountable to Whom? *Educational Management Administration & Leadership*, *37*(5), 624–644. https://doi.org/10.1177/1741143209339654

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

Muralidharan, K., & Singh, A. (2020). Improving Public Sector Management at Scale? Experimental Evidence on School Governance in India. RISE Working Paper 20/056. https://doi.org/10.35489/BSG-RISE-WP_2020/056

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India. *American Economic Review*, *109*(4), 1426–1460. https://doi.org/10.1257/aer.20171112

Muralidharan, K., & Sundararaman, V. (2011a). Teacher Performance Pay: Experimental Evidence from India. *Journal of Political Economy, 119*(1), 39–77. https://doi.org/10.1086/659655

Muralidharan, K., & Sundararaman, V. (2011b). Teacher opinions on performance pay: Evidence from India. *Economics of Education Review, 30*(3), 394–403. https://doi.org/10.1016/j.econedurev.2011.02.001

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. (2000). *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. The National Academies Press. https://doi.org/10.17226/9853

OECD. (2012). *Education at a Glance 2012: OECD Indicators*. OECD Publishing. http://dx.doi.org/10.1787/eag-2012-en

OECD. (2015). *Education at a Glance 2015: OECD Indicators, Annex 3 - Indicator D7*. OECD Publishing. http://www.oecd.org/edu/EAG2015-Annex3-D7.pdf

OECD. (2019). *TALIS 2018 Results (Volume I): Teachers and School Leaders as Lifelong Learners*. OECD Publishing. https://doi.org/10.1787/1d0bc92a-en

Paglayan, A. S. (2019). Public-Sector Unions and the Size of Government. *American Journal of Political Science*, *63*(1), 21–36. https://doi.org/10.1111/ajps.12388

Patrick, B., Plagens, G. K., Rollins, A., & Evans, E. (2018). The Ethical Implications of Altering Public Sector Accountability Models: The Case of the Atlanta Cheating Scandal. *Public Performance & Management Review*, *41*(3), 544–571. https://doi.org/10.1080/15309576.2018.1438295

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. Sage.

Pham, L. D., Nguyen, T. D., & Springer, M. G. (2020). Teacher Merit Pay: A Meta-Analysis. *American Educational Research Journal*. https://doi.org/10.3102/0002831220905580

Piper, B., Destefano, J., Kinyanjui, E. M., & Ong'ele, S. (2018). Scaling up successfully: Lessons from Kenya's Tusome national literacy program. *Journal of Educational Change, 19*(3), 293–321. https://doi.org/10.1007/s10833-018-9325-4

Pritchett, L. (2015). Creating Education Systems Coherent for Learning Outcomes. RISE Working Paper 15/005. https://doi.org/10.35489/BSG-RISE-WP_2015/005.

Pritchett, L. (2018). The Risks of Dangerous Dashboards in Basic Education. RISE Insight. https://doi.org/10.35489/BSG-RISE-RI_2018/006.

Pritchett, L., & Beatty, A. (2012). The Negative Consequences of Overambitious Curricula in Developing Countries (SSRN Scholarly Paper ID 2102726). Social Science Research Network. https://papers.ssrn.com/abstract=2102726

Pritchett, L., & Sandefur, J. (2017). Girls' Schooling and Women's Literacy: Schooling Targets Alone Won't Reach Learning Goals. RISE Working Paper 17/011. https://doi.org/10.35489/BSG-RISE-WP_2017/011.

Rasul, I., Rogger, D., & Williams, M. J. (2020). Management, organizational performance, and task clarity: Evidence from Ghana's civil service. *Journal of Public Administration Research and Theory,* 1-19. https://doi:10.1093/jopart/muaa034

Reynolds, D., Teddlie, C., Chapman, C., & Stringfield, S. (2015). Effective school processes. In C. Chapman, D. Muijs, D. Reynolds, P. Sammons, & C. Teddlie (Eds.), *The Routledge International Handbook of Educational Effectiveness and Improvement:Research, Policy, and Practice* (pp. 77–99). Taylor & Francis.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Sabarwal, S., & Abu-Jawdeh, M. (2018). *What teachers believe: Mental models about accountability, absenteeism, and student learning* (No. WPS8454). World Bank. https://openknowledge.worldbank.org/handle/10986/29883

Scur, D., Sadun, R., van Reenen, J., Lemos, R., & Bloom, N. (2021). The World Management Survey at 18: Lessons and the way forward. *Oxford Review of Economic Policy* forthcoming.

Sellar, S., & Lingard, B. (2013). The OECD and global governance in education. *Journal of Education Policy*, *28*(5), 710–725. https://doi.org/10.1080/02680939.2013.779791

Singh, A. (2020). Myths of Official Measurement: Auditing and Improving Administrative Data in Developing Countries. RISE Working Paper 20/042. https://doi.org/10.35489/BSG-RISE-WP_2020/042

Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K., Conover, L. A., & Reynolds, T. (2003). Differentiating Instruction in Response to Student Readiness, Interest, and Learning Profile in Academically Diverse Classrooms: A Review of Literature. *Journal for the Education of the Gifted,* 27(2–3), 119–145. https://doi.org/10.1177/016235320302700203

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner & E. Souberman., eds.) (A. R. Luria, M. Lopez-Morillas and M. Cole [with J. V. Wertsch], trans). Harvard University Press. (Original manuscripts [ca. 1930-1934]).

Walter, T. F. (2018). State management of education systems and educational performance. Evidence from a management survey at district education offices in Zambia. IGC Working Paper S-89454-ZMB-2.

Williams, M. J. (2020). Beyond 'context matters': Context and external validity in impact evaluation. *World Development*, *127*, 104833. https://doi.org/10.1016/j.worlddev.2019.104833

World Bank (2004). *World Development Report 2004: Making Services Work for Poor People*. World Bank. https://openknowledge.worldbank.org/handle/10986/5986

World Bank (2018). *World Development Report 2018: Learning to Realize Education's Promise*. World Bank. https://openknowledge.worldbank.org/handle/10986/28340