Developing Management: An expanded evaluation tool for developing countries

Renata Lemos*
World Bank
CEP-London School of Economics

Daniela Scur†
Blavatnik School of Government
University of Oxford

First draft: March 2016
This draft: July 2017

[Click here for the latest version]

Abstract In recent years new striking evidence has emerged showing a large tail of badly managed schools and hospitals in developing countries across a number of management areas such as operations management, performance monitoring, target setting and people management. But where exactly along the process of setting their management structures are these organizations failing? This paper describes the development of a survey tool based on an existing instrument to measure management quality – the World Management Survey (WMS) – but tailored to research in the public sector of developing countries: the Development WMS. We collected detailed data from pilots in India, Mexico, and Colombia using face-to-face interviews in settings where weak management practices prevail and observe more variation in the left tail of the distribution. Using this data, we present a brief discussion of the type of data that can be collected and explored with the expanded tool, including the three activities used to systematically measure the strength of each management area in the WMS: (1) implementation, (2) usage, (3) monitoring.[1]

*Email: rlemos@worldbank.org
†Email: daniela.scur@bsg.ox.ac.uk

[1]We thank Kerenssa Kay and Raissa Ebner for excellent research assistance. We thank Karthik Muralidharan for making the pilot of this project possible, Rafael de Hoyos and Ciro Avitabile for the use of the Mexican data and Arthuro Harker Roa for use of the Colombian data. We also thank Morten Bennedsen, James Fenske, Clare Leaver, Kalina Manova, Lant Pritchett and Justin Sandefur for very helpful comments and discussions.

"If the system does not add up to a functional whole, the causal impact of augmenting individual elements is completely unpredictable."

— Lant Pritchett, *RISE Working Paper 15/005*

1 Introduction

Although there has been much progress in improving school enrolment around the world, there is still striking heterogeneity in the distribution of student learning outcomes across countries. This is particularly true for the developing world, and researchers and policy makers are paying increasing attention to addressing this "learning crisis". The traditional economics literature that considers the effect of an individual input on output has provided us with great insights into the individual effect of inputs such as teacher salaries, school infrastructure, school financing, extra teachers, different curriculums, and more textbooks, among many. However, variation in these inputs has not been able to explain a substantial share of the variation in student learning (Glewwe & Muralidharan (2015)). Thus, a new research agenda is urging a more holistic view of education systems in a "systems framework" that includes a series of interconnected types of relationships between different actors and stakeholders, outlined in Pritchett (2015).

This paper makes a methodological contribution by taking lessons from private enterprise and applying to the "public sector" (broadly defined). We develop a feasible tool to measure management practices in schools in developing countries, based on the well-established World Management Survey tool.[2] Here we describe the *Development WMS*, a survey tool based on the original WMS but tailored to measuring management practices in the public sector of developing countries. We discuss each innovation in detail below, but in short:

[2]Since 2008, we have worked alongside Nicholas Bloom, Raffaella Sadun and John Van Reenen to significantly expand the original WMS data collection project and systematically measure management practices within and across countries and it is using this experience that we developed this tool in a comparable way.

1. We identified three management *activities* - implementation, usage, and monitoring - taken into consideration when measuring the strength of each management practice covered by the WMS but which could not be extricated ex-post from a score in the original methodology.

2. We expanded the survey "vertically" by disentangling and mapping these activities to each question of the 20 management practices, creating 60 items to score.[3] In this new survey, however, the responsibility of weighting the importance of each process does not lie with the interviewer, thereby reducing measurement error and allowing the data user to know precisely what led the score for a particular practice to be higher or lower.

3. We expanded the survey "horizontally" to allow for greater variation of scores and allow interviewers to differentiate at a finer level between the strength of processes in place at these schools and hospitals.

While we have strived to keep the essence of the WMS in terms of the questions and practices being measured and the spirit of the scoring grid, we also ensured that the adapted version was applicable in the development setting by addressing three main challenges to using the original WMS in developing countries.

First, the distribution of scores in the education sector in the two developing countries surveyed in the original WMS, India and Brazil, was tight around the scores for weak management practices. Although the global context of the WMS project allows for a useful comparison of world-class and poorly managed organizations across a number of countries, the very thick — almost truncated — left tail for developing countries makes it harder to explore the variation of managerial practices in the less well managed organizations. For example, Lemos & Scur (2012) points out the thick left tail in both schools and hospitals in India and Bloom et al. (2015) show that there is evidence of truncation at the lower bound score of 1, with 82% of the schools in the WMS Indian sample having an overall management score between 1 and 2 that

[3]We did this based on our eight years of training interviewers to conduct the WMS interviews, such that the questions asked related to types of activities are comparable to previous years of surveys.

and no schools have a score above 3 on the original WMS scale. During the data collection for these countries, we often heard interviewers wishing they could "give a 0" to those schools and hospitals that had no process whatsoever to differentiate those from schools and hospitals that had minimal processes, but not enough of an informal process to warrant a score of 2 in the scoring grid.[4]

Second, in terms of implementation, the WMS original methodology uses available sampling frames from established organizations and phone calls to carry out the interviews. Although this was less of a barrier in the manufacturing survey, it was a massive barrier in the public sector surveys in developing countries. For instance, sampling frames in India were difficult to acquire and build, and, when available, they often had names of schools and hospitals but no phone numbers. Unfortunately a common reason for the lack of phone number was that schools simply did not have a physical phone line available. We often ran interviews through managers' cell phones, and a handful of times through payphones located near these organizations as cellphones or landlines were not available. When we were able to reach them, the connection itself was sometimes problematic and several calls had to be placed to complete the interview.[5]

Finally, when thinking about policy implications, we did not have much information in the WMS to pinpoint precisely what part of the process these organizations were failing at the most. Although useful experiments such as Bloom et al. (2013) and Fryer (2014) have substantially helped us learn about the large effect that improvements in whole sets of management practices can afford, we do not yet have a systematic picture of what particular *types of processes* matter the most across different settings in developing countries.[6] The 20 management practices covered by

[4]The reason we refrained from stretching the scoring grid to 0 and instead added half points was to preserve comparability of the ordinal scale and increase specificity equally across all score categories.

[5]The higher the number of calls that have to be made, the lower the probability of completing an interview.

[6]Focusing on charter schools in the US, Dobbie & Fryer (2013) run a similar exercise where they collect a large amount of information on the inner-workings of 35 charter schools to investigate the practices that matter the most for school effectiveness.

the WMS are scored based on a set of processes which are systematically triangulated by the skilled interviewer and facts are evaluated based on the survey grid to determine higher or lower scores. However, we argue that it becomes important to understand the marginal importance of each type of process when considering the type of policy interventions that are feasible, especially in the context of countries facing limited budgets and institutional constraints.

We have also developed accompanying field paper forms to facilitate the interview process as the Development WMS is meant to be run face-to-face by enumerators who visit the schools and hospitals. These forms were carefully designed to ensure that the information collected during the interviews would be sufficient for the post-interview scoring. In the phone interviews, the enumerators are able to consult the grid to ensure they have enough information, but in the face-to-face interviews they are not allowed to take the grid along as it would undermine the double-blind exercise. The importance of providing a useful field-friendly data collection tool is often underestimated. The enumerators are often not researchers by training and may fail to record important information or even record wrong information during survey interviews if not properly prompted by their field tool.[7]

With a set of individual project partners, we are in the process of collecting data using this new expanded survey tool in schools in Andhra Pradesh-India (completed), Mexican schools (ongoing, pilot completed), Colombian schools (completed), Chinese hospitals (ongoing) and Indian hospitals (pilot completed).[8] This survey tool has often been used as an additional module in larger projects, and sampling frames of

[7] A website with instructional videos and interactive calibration tools to minimize the fixed costs of training and implementation will be made freely available to the research community in mid-2017.

[8] We have partnered with Karthik Muralidharan and the APSC project for Indian schools, Arturo Harker Roa and the Colombian Ministry of Education for Colombian schools, Rafael de Hoyos and Ciro Avitabile from the World Bank and the Mexican Ministry of Education for Mexican schools, Winnie Yip and the Ministry of Health for Chinese hospitals and Raffaella Sadun for Indian hospitals. We are immensely thankful to Raissa Ebner and Kerenssa Kay for training the Mexican school pilot teams, Raissa Ebner for training the Mexican and Colombian school teams, and Kerenssa Kay for running the Indian hospital pilot. For an initial look at the Colombian data, see Bermudez & Harker (2016).

these projects were not always necessarily representative random samples and thus are not directly comparable. While these samples were not formally designed to be representative of all schools in these countries, collectively they paint a useful picture of selected public sector organizations in low- and middle-income countries and allow us to validate our new survey tool.[9]

2 Measuring processes in developing countries

The original public sector WMS covers 20 questions across two main areas: operations management and people management. The original survey sub-divides operations management into lean operations, monitoring and target management, as follows:

1. *Operations management*

   (a) *Lean operations* in schools covers practices including whether the school has meaningful processes that allow pupils to learn over time; teaching methods that ensure all pupils can master the learning objectives; whether the school uses assessment to verify learning outcomes at critical stages and makes data easily available and adapts pupil strategies accordingly.

   (b) *Monitoring management* covers practices of continuous improvement, performance tracking, review and dialogue, and consequence management. It measures whether the school has processes towards continuous improvement and lessons are captured and documented, whether school performance is regularly tracked with useful metrics, reviewed with appropriate

[9]The samples are as follows: the Andhra Pradesh data is a random sample of public and private primary schools in 5 districts from the APREStt project; the Mexican data is a combination of samples from primary schools that are part of PEC (Programa Escuelas de Calidad) in Durango, Guanajuato, Estado de Mexico and Tabasca, marginalized primary schools in Puebla, and primary and junior high schools in Tlaxcala and Morelos; the Colombian data is a random sample from the lowest performing public schools in the country (approximately 4,000 of the 22,000 schools in Colombia); the Chinese hospital data is a random sample of hospitals and the Indian hospital data is from a pilot of 25 hospitals in Andhra Pradesh.

frequency, quality, and follow-up, and communicated to staff.

(c) *Target management* covers practices in the balance and interconnection of targets, the time-horizon and difficulty of the targets, as well as their clarity and comparability. It measures whether the school, department, and individual targets cover a sufficiently broad set of metrics; whether these targets are aligned with each other and the overall goals.

2. *People management* covers practices in handling good and bad performance, measuring whether there is a systematic approach to identifying good and bad performance, rewarding school teachers proportionately, dealing with under-performers, and promoting and retaining good performers.

As mentioned before, we preserve the practices and areas covered in the original WMS and identify three key activities used to systematically measure these practices, and expand it both "vertically," by further dividing each of the 20 practices into the three activities we are looking to measure and "horizontally," increasing the granularity of scores by allowing half points.

2.1 Identifying processes behind management practices

In the Development WMS, we identify three key activities that are captured to measure the strength of each management practice within an organization. Each process consists of:

1. Activity 1. Implementation: formulating, adopting and putting into effect management practices.
2. Activity 2. Usage: carrying out and using management practices frequently and efficiently.
3. Activity 3. Monitoring: monitoring the appropriateness and efficient use of management practices.

More specifically, in the original WMS, each of the overall management, operations and people management indices is made up of a set of the 20 practices, and each practice is measured through several structured questions. Each one of the 20 management practices contains a large amount of information about how that specific practice being carried out at the establishment. For example, when measuring "data-driven planning and student transitions" at a school, the WMS interviewer evaluates the practice based on three activities: (1) what type of data is available (test scores, attendance, etc), (2) leaders understand critical points of transition for students (when to change learning levels), (3) leaders have a data-driven approach to decisions (principal and teachers use data to determine transitions). The combined responses to this practice are scored against a grid which goes from 1 - defined as "School may be aware of critical transitions for students, but little or no effort is made to match support services to students; data is often unavailable or difficult to use." up to 5 - defined as "Student transitions are managed in an integrated and proactive manner, supported by formative assessments tightly linked to learning expectations; data is widely available and easy to use."

In the original WMS instrument, the interviewer triangulates the activities herself and assigns one single score taking all the activitites into account. This task requires a high cognitive ability from the interviewer as well as consistent monitoring of the interviewing process by supervisors to ensure comparability.[10] It is not possible, however, to extricate from the final data ex-post how each process weighed in the interviewer decision. In the Development WMS, each process is evaluated separately and ex-post averaged out to get the practice's score. This is useful in a practical sense because it removes the triangulation responsibility from the interviewer, which then lowers the cognitive threshold required in hired interviewers and facilitates the deployment of the survey in low-capacity contexts.

[10]This is one of the reasons for the high per-interview cost of the WMS. Interviewers are generally masters students from top UK schools and experienced supervisors monitor over 80% of the interviews.

Furthermore, in an academic and policy research sense we can now disentangle precisely where the process is failing and be much more specific in targeting of interventionist policies. For example, in one of the pilot school visits we carried out in Andhra Pradesh, when asked the first question in the earlier example the principal promptly pulled out examples of report cards that they used to track student performance (Figure 1). The report cards had plenty of detail on student achievement and behaviour over time, and were signed by the teacher, principal and parent. This would certainly warrant a score of 3.5 or 4 on the implementation process part of the topic being measured. When we then asked the subsequent questions of how the data is used and how it relates to student transitions, we received the unsatisfying answer that the report cards were simply stacked in the corner of the principal's office and if the teachers were curious they could go and find an individual student's card whenever they wanted.

Figure 1: Report card from a rural school in Andhra Pradesh

In short, there was no process of compiling the data to be useful more generally, and there was certainly no process to use the data to help guide the transition between

levels of learning. Crucially, the scores for the usage and monitoring parts would have been low, in the 1-1.5 range and the overall score might have been around a 2-2.5, and we would correctly interpret that there is not a very good formal system of data-driven student transitions, *but* we would have missed important information that it is not the data collection part of the process that is failing, but rather the usage of the data already collected. If we think in policy terms, a policy that targets giving schools best practices for report card development would be relatively useless in this context, whereas one that builds a system that they can use the data already collected would be much more effective.

2.2 Expanding the instrument vertically: higher dimensionality

Operationally, we develop the extended grid by mapping each of the three key activities back to the questions asked for measuring each WMS practice. "Implementation" is broadly related to question 1, "usage" is broadly related to question 2, and "monitoring" is broadly related to question 3 in each management practice. Thus, beyond looking at the average score of each practice, we can also dig deeper to understand what part of the process is driving the results. This increases the number of scores from 20 to 60.

With the increased number of variables, we can create a new set of indices to test whether they are any more informative than the original survey. We construct four sets of indices. For the first set, we follow a similar methodology to the original WMS and use the information referring to all three activities very simply. First we take an average of the three sub-questions to build a single score for each of the 20 original practices, analogous to how a WMS interviewer would assign a single score to each practice. We then take the z-score of each practice and creating indices for overall management (average of 20 practices), operations management (average of fourteen operations practices), people management (average of six people management practices). This can be interpreted in the same way as the original WMS, but

with lower measurement error.[11]

The main innovation in our survey is in the second, third and fourth set of indices. To build these, we skip the first step of averaging across the three activities for each practice and re-organize the dataset into three new sets of 20 practices along the lines of each process. We take the z-score of all the sub-questions and build average indices for overall management, operations management and people management using the 20 sub-practices for each of the process types. For example, we can build an implementation management index by taking the average of the 20 implementation sub-practices, and again do the same to create a usage management index and a monitoring management index. Further, we could build implementation operations and implementation people management indices using only the first sub-questions within each management practice.

In short, we first produce a set of overall management, operations management and people management indices using a similar methodology to the original WMS (ie. using all the information given for a particular question), and also produce three "finer" sets of indices, broadly referring to (1) process implementation of overall, operations and people management, (2) process usage of overall, operations and people management, and (3) process monitoring of overall, operations and people management.

While we broadly follow the original WMS convention for building the comparable indices (overall management, operations and people management), we have conducted a factor analysis of our new school survey tool with the data from Andhra Pradesh, Mexico and Colombia to validate this. We find that factor analysis on the 20 management practices as well as the more granular 60 processes yields generally similar results to those found in the manufacturing sector in Bloom et al. (2014), though considering the factors within each of the processes is rather telling. We present the results of the factor analysis in Table 1. The first two factors across all the four sets of variables analyzed explained all the variation across the variables, so we present only

[11]The original WMS excludes the leadership questions from its overall management score, so we follow the same convention here.

the first two factors here for each set of variables analyzed. The first two columns use the management practice scores that are comparable to the original WMS; that is, each of the 20 management scores is an average of the three respective activities in the D-WMS. The following six columns break down the use of the dataset instead of taking an average of the three activities: columns (3) and (4) use only the first activity for each management practice —implementation, columns (5) and (6) use only the second activity for each management practice — usage, and columns (7) and (8) use only the third activity for each management practice — monitoring.

Starting with the original WMS practices, the first factor pattern suggests the largest pattern of relationships in the data, and we see that there is one principal factor that explains 87% of the variance and loads positively on all practices, and the loadings are high.[12] Similar to the result in manufacturing in (Bloom et al. 2014), this suggests that there is a "common factor of good management" (Factor 1). In the context of the D-WMS data, this suggests that schools that are well managed on one practice are more generally also likely to be well managed across all practices. The second factor pattern suggests the second largest pattern of relationships that is uncorrelated with the first, and this second factor that explains about 13% of the variance and loads negatively on nearly all of the operations management practices, but positively on all the people management practices. The loadings on the second factor are, however, much smaller than those seen in the first factor — especially those in operations. In the manufacturing results, the second factor loaded positively on operations and negatively on people management. Here we see the opposite loadings but a similar pattern of specialization. In our context, it seems there is a second factor of "good people management" suggesting schools specialize in people management versus operations management.

Moving on to examining the results for each individual activities, the first pattern we note is that across all three activities we find a common factor of "good management" as the first factor. The second factor, however, is not the same across activities. In the implementation activity questions across the full survey, we see that the second factor

[12] A general rule of thumb tends to be factor loadings of 0.4 and above are relevant.

Table 1: Factor analysis of management practices and processes (unrotated)

	Original WMS		Implementation only		Usage only		Monitoring only	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2
Ops 1: Standardisation of instructional processes	0.7257	**-0.1389**	0.5831	0.1985	0.5490	**-0.1507**	0.5713	**-0.0255**
Ops 2: Data driven planning and student transition	0.7064	**-0.1008**	0.6287	0.1769	0.4864	0.0132	0.6315	**-0.0934**
Ops 3: Personalization of instruction and learning	0.7524	**-0.1228**	0.5909	0.1253	0.6221	**-0.1599**	0.5671	**-0.1140**
Ops 4: Adopting educational best practices	0.7535	**-0.1421**	0.5994	0.1226	0.5608	**-0.1406**	0.6699	**-0.1680**
Ops 5: Continuous Improvement	0.7106	**-0.1761**	0.5959	0.1915	0.5717	**-0.2039**	0.6038	**-0.1100**
Ops 6: Performance Tracking	0.6791	**-0.2727**	0.5248	0.0144	0.4044	**-0.3834**	0.6354	**-0.1892**
Ops 7: Review of Performance	0.6956	**-0.3445**	0.4165	0.2374	0.5881	**-0.1080**	0.6343	**-0.2265**
Ops 8: Performance Dialogue	0.7585	**-0.2517**	0.5737	0.2257	0.5998	**-0.2800**	0.7036	**-0.1505**
Ops 9: Consequence Management	0.7165	**-0.2725**	0.6511	0.2714	0.4102	**-0.3295**	0.6065	**-0.1642**
Ops 10: Type of Targets	0.7547	**-0.0392**	0.5964	0.1142	0.6412	0.0441	0.6596	**-0.0059**
Ops 11: Interconnection of Goals	0.7479	**-0.0989**	0.5504	0.2372	0.5761	0.2020	0.5988	**-0.0543**
Ops 12: Time Horizon	0.5292	**-0.0727**	0.4730	0.0943	0.4216	**-0.1601**	0.3600	0.0766
Ops 13: Goals are Stretching	0.6800	0.0236	0.4763	0.1107	0.5682	**-0.0392**	0.4505	0.1552
Ops 14: Clarity of Goals	0.7215	0.1141	0.5911	0.0362	0.4672	0.2729	0.6181	0.0421
People 1: Instilling a talent mindset	0.6910	0.5543	0.6783	**-0.4340**	0.7096	0.4230	0.4986	0.4205
People 2: Incentives and Appraisals	0.7462	0.2712	0.7367	**-0.4433**	0.5863	0.1609	0.4146	0.2262
People 3: Making room for Talent	0.7059	0.2156	0.6778	**-0.4421**	0.6488	0.0233	0.3597	0.0895
People 4: Developing Talent	0.5544	0.2292	0.5896	**-0.3115**	0.5290	0.0854	-0.0934	0.2893
People 5: Distinctive Emp Value	0.6163	0.3884	0.5018	**-0.1004**	0.4966	0.2886	0.5143	0.3385
People 6: Retaining Talent	0.4507	0.5486	0.3737	**-0.1714**	0.1990	0.5648	0.5074	0.4272
Eigenvalue	9.513	1.3882	6.6605	1.1278	5.9000	1.2148	6.1177	0.8507
% total variance	0.8712	0.1271	0.8639	0.1463	0.8600	0.1771	0.9532	0.1326
Cumulative variance	0.8712	0.9983	0.8639	1.0102	0.8600	1.0371	0.9532	1.0858

loads negatively across all people management practices, while it loads positively across the same practices under the usage and monitoring activities.

In all, it is reassuring to see that the general patterns that have been found in the literature using the WMS hold with our new instrument as well. It is also reassuring to see that we find some different patterns across different activities within each management practice. The analysis suggests that we can, indeed, learn something new in terms of the patterns of management within schools from the new survey. In the next section we will show how each of the activities are correlated with the outcome of interest in this context — student and teacher outcomes — but for now we resume the discussion of the other survey changes.

2.3 Expanding the instrument horizontally: greater score variation

The horizontal expansion of the instrument is more straightforward. In the original WMS, interviewers are allowed to score values of 1, 2, 3, 4 or 5. No half points are allowed and no "2 or 3" values are accepted. If interviewers are unsure of whether the practice warrants a 2 or a 3, they discuss it with their colleagues and their supervisors to make a final decision. This scoring guideline worked well in developed countries as there was wide range of scores, with some schools or hospitals being very well managed and some being very badly managed, but most schools or hospitals had at least *some* practice in place, even if rudimentary. In the India and Brazil waves, however, we found several schools that had absolutely no practices in place and some that had very minimal practices in place. To score a 2 in the WMS, there must be a reasonable practice in place that is informal (if it were a formal practice it would be awarded a 3 or higher). Thus, both schools with no practices and minimal practices were awarded 1, whereas in the Development WMS the interviewer would be able to distinguish and score 1 for no practices and 1.5 for minimal practices.

Crucially, however, we follow the same gradual scoring scheme as the original WMS,

which allows us to easily re-cast the scores into what they would have been in the original survey. For example, if an interviewer gave a practice in a school a score of 1.5, it is because it did not reach a high enough level to be a 2, and thus in the original WMS it would have been a score of 1. We implement these adjustments to create original WMS-comparable scores, and plot both distributions in Figure 2.

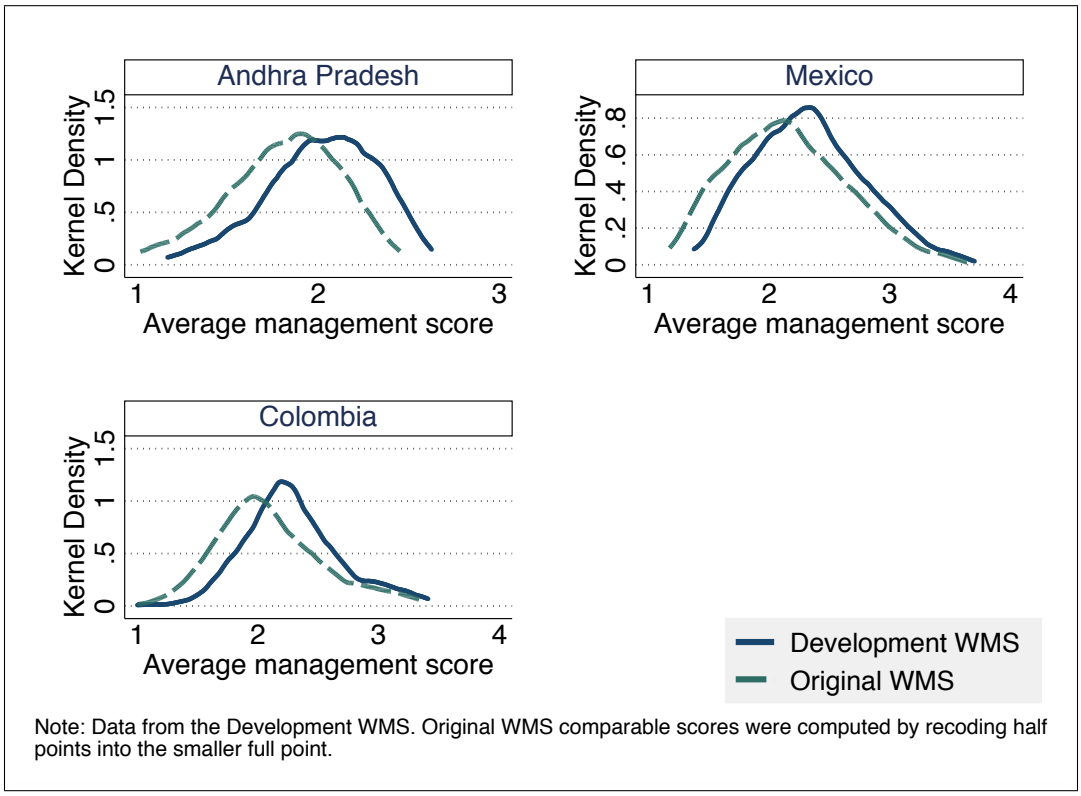
The figure shows an example of the distribution of scores using data from Andhra Pradesh, Colombia and Mexico. The main goal of the new scoring guide was to allow for a systematic distinction between schools with absolutely no structures in place — a score of 1 — and schools with very minor structures in place that could not yet be considered informal processes, but were also not completely nonexistent — a score of 1.5. Allowing for 0.5 extra points in all scores has the expected effect of shifting the distribution to the right as analysts are allowed to score higher points, but crucially we now observe a longer tail between the score of 1 and 2 despite still seeing distributions skewed to the left.

2.4 Interpreting the management index and sub-indices measures

Before we move on to providing an overview of the data collected thus far, it is important to emphasise a few key points when interpreting the management index and sub-indices.

The D-WMS (as well as the WMS) does not measure the skills of the *manager* but rather measures the processes embedded in each managerial practice in place within the establishment. Thus, the methodology requires that interviews be conducted with managers who have been in the establishment long enough to become acquainted with the practices in place at that establishment. If the interview is conducted with a manager who has recently taken a post in the establishment in question (that is, less than one year), the manager might refer to practices that were in place in her

Figure 2: Difference in distribution of scores between WMS and D-WMS



Note: Data from the Development WMS. Original WMS comparable scores were computed by recoding half points into the smaller full point.

previous post rather than the particular establishment she is currently working in.[13] For example, a principal who has been at a school for only two months might not have gone through a review process with their teachers and cannot speak directly about the appraisal systems in place in that particular school. Although they possibly bring in new and different managerial practices into the school, it becomes difficult to discern whether these practices have truly been implemented in the new school or whether it is a current "wish list" of the new principal.

Considering that we are measuring the management practices currently in use, in general the management indices can be interpreted as follows:

- A score between 1 to 2 refers to an establishment with practically no structured management practices or very weak management practices implemented;
- A score between 2 to 3 refers to an establishment with some informal practices implemented, but these practices consist mostly of a reactive approach to managing the organization;
- A score between 3 to 4 refers to an establishment that a good, formal management process in place (though not yet often or consistent enough) and these practices consist mostly a proactive approach to managing the organization;
- A score between 4 to 5 refers to well-defined strong practices in place which are often seen as best practices in the sector.

[13] In fact, this does happen during interviews and those conducting the interviews are instructed to continuously check that the examples provided are from the current establishment rather than any previous post.

3 Does D-WMS provide any new meaningful variation for data analysis?

3.1 Observing within-practices and between-practice variation

As mentioned in the previous section, the expanded D-WMS instrument allows us to improve the quality of data collection in a number of practical ways, but is this new way of collecting data also helpful in terms of data analysis? That is, do we observe any *within-practice* and *between-practice* variations in the data which can be further explored?

Within-practice variation indicates whether organizations emphasize one process over the other within each management practices such as scoring highly in process implementation but poorly in process usage or process monitoring. For example, in order to track their performance, schools may formulate and put into effect a system of metrics to monitor performance but not use this system frequently and efficiently. Alternatively, some schools may define perhaps only one or two indicators to monitor performance but use this indicators appropriately and frequently. *Between-practice variation* indicates if the scores for the three types of processes vary systematically across all management practices. For example, schools may be able to formulate and put into effect systems for performance monitoring, target setting as well as people management. But while process implementation scores may be high across the board for some organizations, they might not be able to effectively use or monitor all systems in place.

We present the correlation matrix for activities within each practice in Figures 3 and 4. We observe that all correlations are positive and significant at the 1% level but of varying coefficients, ranging from 0.04 to 0.66: 14.1% of correlated pairs present a coefficient of equal or lower than 0.25, 65.0% present a coefficient between 0.25 and 0.50, while 21% present a coefficient of equal or above 0.50.

Figure 3: Management process: correlations

		Andhra Pradesh Schools			Mexico Schools			Colombia Schools		
		implementation	usage	monitoring	implementation	usage	monitoring	implementation	usage	monitoring
2. Standardization of Instructional Planning Processes	implementation	1.00			1.00			1.00		
	usage	0.52	1.00		0.37	1.00		0.44	1.00	
	monitoring	0.42	0.37	1.00	0.45	0.38	1.00	0.42	0.44	1.00
3. Personalization of Instruction and Learning	implementation	1.00			1.00			1.00		
	usage	0.11	1.00		0.36	1.00		0.26	1.00	
	monitoring	0.04	0.52	1.00	0.52	0.57	1.00	0.45	0.39	1.00
4. Data-driven Planning and Student Transitions	implementation	1.00			1.00			1.00		
	usage	0.34	1.00		0.44	1.00		0.39	1.00	
	monitoring	0.39	0.38	1.00	0.42	0.43	1.00	0.37	0.45	1.00
5. Adopting Educational Best Practices	implementation	1.00			1.00			1.00		
	usage	0.47	1.00		0.47	1.00		0.56	1.00	
	monitoring	0.43	0.32	1.00	0.48	0.55	1.00	0.48	0.53	1.00
6. Continuous Improvement	implementation	1.00			1.00			1.00		
	usage	0.41	1.00		0.42	1.00		0.44	1.00	
	monitoring	0.30	0.57	1.00	0.42	0.56	1.00	0.52	0.61	1.00
7. Performance Tracking	implementation	1.00			1.00			1.00		
	usage	0.26	1.00		0.25	1.00		0.28	1.00	
	monitoring	0.12	0.26	1.00	0.44	0.27	1.00	0.32	0.35	1.00
8. Performance Review	implementation	1.00			1.00			1.00		
	usage	0.29	1.00		0.21	1.00		0.30	1.00	
	monitoring	0.32	0.41	1.00	0.26	0.44	1.00	0.34	0.41	1.00
9. Performance Dialogue	implementation	1.00			1.00			1.00		
	usage	0.36	1.00		0.46	1.00		0.48	1.00	
	monitoring	0.33	0.31	1.00	0.43	0.52	1.00	0.57	0.56	1.00
10. Consequence Management	implementation	1.00			1.00			1.00		
	usage	0.26	1.00		0.16	1.00		0.32	1.00	
	monitoring	0.18	0.23	1.00	0.47	0.17	1.00	0.33	0.34	1.00
11. Balance of Targets/Goal Metrics	implementation	1.00			1.00			1.00		
	usage	0.35	1.00		0.60	1.00		0.54	1.00	
	monitoring	0.48	0.41	1.00	0.54	0.59	1.00	0.26	0.46	1.00

equal or below 0.25
equal or above 0.50

Note: All correlations are significant at the 1% level.

Figure 4: Management process: correlations

		Andhra Pradesh			Mexico			Colombia		
		implementation	usage	monitoring	implementation	usage	monitoring	implementation	usage	monitoring
12. Interconnection of Targets/Goals	implementation	1.00			1.00			1.00		
	usage	0.40	1.00		0.31	1.00		0.29	1.00	
	monitoring	0.31	0.47	1.00	0.31	0.54	1.00	0.34	0.52	1.00
13. Time Horizon of Targets/Goals	implementation	1.00			1.00			1.00		
	usage	0.08	1.00		0.58	1.00		0.64	1.00	
	monitoring	0.18	0.20	1.00	0.42	0.50	1.00	0.48	0.35	1.00
14. Stretch of Targets/Goals	implementation	1.00			1.00			1.00		
	usage	0.15	1.00		0.36	1.00		0.29	1.00	
	monitoring	0.11	0.14	1.00	0.28	0.34	1.00	0.14	0.32	1.00
17. Clarity and Comparability of Goals	implementation	1.00			1.00			1.00		
	usage	0.30	1.00		0.49	1.00		0.47	1.00	
	monitoring	0.40	0.30	1.00	0.47	0.41	1.00	0.39	0.38	1.00
18. Building a High Performance Culture/ Rewarding High Performers	implementation	1.00			n.a.			1.00		
	usage	0.46	1.00		n.a.	n.a		0.34	1.00	
	monitoring	0.57	0.66	1.00	n.a.	n.a	n.a	0.36	0.26	1.00
19. Making Room for Talent/ Removing Poor Performers	implementation	1.00			1.00			1.00		
	usage	0.27	1.00		0.51	1.00		0.13	1.00	
	monitoring	0.07	0.12	1.00	0.34	0.37	1.00	0.36	0.30	1.00
20. Promoting High Performers	implementation	1.00			1.00			1.00		
	usage	0.57	1.00		0.42	1.00		0.41	1.00	
	monitoring	0.42	0.57	1.00	0.24	0.38	1.00	0.10	0.20	1.00
21. Managing Talent	implementation	1.00			1.00			1.00		
	usage	0.10	1.00		0.35	1.00		0.52	1.00	
	monitoring	0.27	0.07	1.00	0.15	0.02	1.00	0.45	0.65	1.00
22. Retaining talent	implementation	1.00			1.00			1.00		
	usage	0.44	1.00		0.43	1.00		0.43	1.00	
	monitoring	0.36	0.57	1.00	0.42	0.56	1.00	0.32	0.40	1.00
23. Creating a Distinctive Employee Value Proposition	implementation	1.00			1.00			1.00		
	usage	0.55	1.00		0.44	1.00		0.39	1.00	
	monitoring	0.45	0.53	1.00	0.52	0.37	1.00	0.33	0.35	1.00

equal or below 0.25
equal or above 0.50

Note: All correlations are significant at the 1% level.

3.2 Validation of the new survey

At its core, the relevance of this research project relies on how much of the variation in the outcomes we are concerned about can be picked up by our management measure. Bloom et al. (2015) show that the original WMS measure is correlated with student outcomes across a range of countries, so we can expect that our measure will also likely be correlated with school-based student outcomes. What we can explore further is whether any one of the process types explain more of the variation compared to the other processes. To explore this, we are currently matching the new data with various performance datasets from the countries where we have access to such data. In Andhra Pradesh we have already conducted this analysis and found a positive correlation between management and teacher value added across the indices.[14]

4 Conclusion

Over the past decade the research agenda on the economics of management practices has been moving forward in exciting ways. As development economists, we see and hear about the missed opportunities in our field visits and in hundreds of interviews when it comes to "good management" practices. As suggested in Pritchett (2015), management practices are important facet in understanding public service delivery from a systems framework view. This new measurement tool is only the first step, and we are building a training platform that will allow individual research teams to include the full survey or individual modules in their own field work. This will be crucial for building a large-scale *comparable dataset* and start to uncover how schools across the world are managed, and which levers are more important in which contexts.

We have two main avenues where we plan to take this work. The first is to conduct a more rigorous analysis of identifying the patterns of management processes that

[14] Lemos et al. (n.d.)

are correlated with student outcomes, and also expand these to other important outcomes such as teacher behaviour and value added. We plan to use new methods such as machine learning to tackle these new questions. A second avenue of work is considering theoretically what might be behind this relationship between management and student outcomes. Bloom et al. (2016) develops a model for the manufacturing sector, but the education sector is fraught with issues of interdependent relationships of accountability and deals with different types of workers — such as intrinsically motivated teachers and principals. We hope to use the stylized facts we have learned from this new picture of management across different school systems in different countries to help guide a starting point for a theoretical framework.

5 Appendix

Collecting data using the Development WMS In order to collect the data in developing countries, rigorous training on the Development WMS for schools was provided to 15 interviewers in India, 30 interviewers in Colombia, 70 interviewers in Mexico, and training on the Development WMS for hospitals was provided to 40 interviewers in China.

The training consists of thorough explanations of the scoring grid in an interactive environment, and multiple group scoring sessions of mock interviews to correct any inconsistent interpretation of responses and to ensure consistency across interviewers.[15] This one-week training session and subsequent routine data and calibration checks are crucial for data quality, and we have developed a process to standardize both the training and the supervisory follow up.

The Development WMS uses the same open-ended questions used in the original WMS methodology, seeking both comparability and to follow best practices in eliciting truthful responses from respondents. Continuing with the example on the management practice of "Performance Tracking," the interviewer starts by asking the open question "What kind of main indicators do you use to track school performance?", rather than a closed ended question such as "Do you use class-room level test scores indicators [yes/no]." The first question is then usually followed up by further open-ended questions such as "how frequently are these indicators measured?", "Who gets to see this data?" and "If I were to walk through your school what could I tell about how you are doing against your indicators?" Such open-ended questions avoid leading responders towards a particular answer and produce higher quality data. As mentioned above, the interviewer knows the information she is seeking and will continue to ask follow up questions if necessary.

[15] During the training week for the school survey in India, we also piloted the Development WMS in 5 schools (a mix of private and public) to ensure the detailed questions and scoring grid appropriately captured the information provided during the interview. Travel expenses were generously covered by J-PAL.

In order to ensure the interviews are consistent within interviewer groups and non-biased, all interviews were "double-scored" and "double-blind," following the WMS methodology but adapting it to face-to-face interviews. Double scored means that the first interviewer was accompanied by a second interviewer whose main role was to monitoring the quality of the interview being conducted by taking notes and separately scoring the responses after the interviews had ended. The first and second interviewers would then discuss their individual scores to correct for any misinterpretation of responses. We mixed pairs of interviewers as much as possible throughout the survey, conditional on geographic limitations. Double-blind means that, at one end, interviewers conducted the face-to-face interview without informing school principals or hospital managers that their answers would be evaluated against a scoring grid.[16] At the other end, our interviewers did not know in advance anything about the school or hospital's performance.

As detailed in Bloom et al. (2014), the original WMS is an expensive survey to run and requires highly skilled interviewers to conduct the interviews and consistently score establishment practices. The WMS has primarily employed masters and PhD students from top European and North American universities to conduct the interviews over the past 10 years of the project. With the Development WMS instrument the level of skill of the interviewers is relatively lower considering that the decision of "weighting" the quality of the processes to decide on a single score for each practice is taken away. To be sure, the interviewers still need to be skilled enough to understand the training session and the practices being measured, but in general the new tool allows for greater flexibility in recruitment of interviewers and facilitates local capacity building by hiring from local institutions.

[16]None of the forms used by both the first and the second interviewers contained the detailed scoring grid. The interviewers would score the interviews based on their notes after the interviews had been completed and, therefore, the scoring grid was not shared with the principal.

References

Bermudez, N. & Harker, A. (2016), Factors associated with the quality of school management practices: an empirical analysis for colombia, Working paper series: Documentos de trabajo egob, Universidad de los Andes.

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D. & Roberts, J. (2013), 'Does management matter? evidence from india', *The Quarterly Journal of Economics* **128**, 1–51.

Bloom, N., Lemos, R., Sadun, R. & Reenen, J. V. (2015), 'Does management matter in schools?', *The Economic Journal* **125**, 647–674.

Bloom, N., Lemos, R., Sadun, R., Scur, D. & Reenen, J. V. (2014), 'The new empirical economics of management', *Journal of the European Economics Association* .

Bloom, N., Sadun, R. & Van Reenen, J. (2016), Management as a technology?, Working paper series, NBER.

Dobbie, W. & Fryer, R. G. (2013), 'Getting beneath the veil of effective schools: evidence from new york city', *American Economic Journal: Applied Economics* **5**(4), 28–60.

Fryer, R. G. (2014), 'Injecting charter school best practices into traditional public schools: evidence from field experiments', *Quarterly Journal of Economics* **129**(3), 1355–407.

Glewwe, P. & Muralidharan, K. (2015), Improving school education outcomes in developing countries, Working Paper 15/001, RISE.

Lemos, R., Muralidharan, K. & Scur, D. (n.d.), School management in india. Manuscript.

Lemos, R. & Scur, D. (2012), Could poor management be holding back development?, Working paper, International Growth Centre.

Pritchett, L. (2015), Creating education systems coherent for learning outcomes, Working Paper 15/005, RISE.