The case for statecraft in education: The NDP, a recent book on governance, and the New Public Management inheritance

MARTIN GUSTAFSSON

Stellenbosch Economic Working Papers: WP16/2019

www.ekon.sun.ac.za/wpapers/2019/wp162019

December 2019

KEYWORDS: Learning outcomes, governance, monitoring systems, South Africa
JEL: C81, I21, O21

ReSEP (Research on Socio-Economic Policy)
http://resep.sun.ac.za

DEPARTMENT OF ECONOMICS
UNIVERSITY OF STELLENBOSCH

SOUTH AFRICA

UNIVERSITEIT
STELLENBOSCH
UNIVERSITY

BER
BUREAU FOR ECONOMIC RESEARCH

A WORKING PAPER OF THE DEPARTMENT OF ECONOMICS AND THE
BUREAU FOR ECONOMIC RESEARCH AT THE UNIVERSITY OF STELLENBOSCH

www.ekon.sun.ac.za/wpapers

The case for statecraft in education

The NDP, a recent book on governance, and the New Public Management inheritance

MARTIN GUSTAFSSON

DECEMBER 2019

ABSTRACT

Statecraft in the sense of building the systems of a capable state is an endeavour that should be taken seriously, yet often it is not. The paper takes issue with a position where hope in a more capable state is to some extent abandoned, or postponed, on the basis of a frustration with history. Such positions are sometimes justified through reference to alternative routes towards progress involving less reliance on a central state, and more reliance on local action and accountability. This paper argues that it is dangerous to dismiss the role of the state, especially without a careful and informed assessment of what is wrong with it. It argues that state dysfunctionality, which is clearly a reality, is so central a development problem that it warrants far more rigorous analysis than what is often found in the literature. Local accountability is also vital, but ideally as a complement to a functioning national system. The problem is not just that proponents of local action can be too quick to dismiss the role of the state. The proponents of capable states, such as the World Bank, are too often overly idealistic and impractical when they offer advice on statecraft. On some important matters, there is a mismatch between the advice and the realities planners face. The paper argues these points in the context of schooling systems, and specifically that of South Africa. It is in part a response to a recent book on governance in the South African schooling sector. It moreover makes reference to South Africa's National Development Plan.

Martin Gustafsson
Department of Economics
University of Stellenbosch
Private bag X1, 7602
Matieland, South Africa
E-mail: mgustafsson@sun.ac.za

RESEP
Research on Socio-Economic Policy

The author is based part-time at the Department of Basic Education in Pretoria, and is Associate Professor at the University of Stellenbosch.

1 Introduction

This paper uses a recently published and important book, titled *The politics and governance of basic education: A tale of two South African provinces*, as a point of departure for discussing the role of effective bureaucracies, and specific large-scale interventions, in improving learning outcomes in schools. The book serves as an interesting point of departure as it does *not* see much promise in such improvement work, at least not in the South African context. Thus, the current paper is in part a rebuttal of one message presented in the book, a book which in other respects is a valuable contribution to our stock of knowledge about schooling in South Africa.

The current paper restates the case for 'statecraft', in the South African context, though its relevance extends beyond South Africa. This term 'statecraft' is used in the more colloquial sense of building an effective state. An effective state requires a logical organisational structure, systems of incentives for officials, robust information systems, necessary financial controls, good public communication arrangements, and so on. These need to be built, and improved upon. For South Africa, especially in the 1990s, building an effective state was a very real and complex endeavour, given the political imperative to dismantle the quasi-colonial and racially stratified apartheid state, and to introduce in its place a modern liberal democracy. The more technical meaning attached to 'statecraft' here should be distinguished from what the originator of the term, the British political scientist Jim Bulpitt meant. For Bulpitt, 'statecraft' is wide enough to encompass the military, propagandistic and ideological mechanisms through which political elites attempt to remain in power[1]. The term is used in a much narrower sense here.

It should be emphasised that the current paper is not a review of the book – for this a fuller engagement with the broad range of issues examined in the book would be necessary. The paper is also more than a review insofar as it discusses policy issues clearly outside the scope of the book.

Section 2 discusses the book's position on what it sees as an optimal balance between three critical areas of action: enhancing the accountability of politicians to citizens; building a capable state (statecraft); and ensuring that schools are properly accountable to the communities they serve.

Section 3 argues the case for statecraft, and explains how a misunderstanding of South Africa's past experiences with New Public Management (NPM), and of what NPM means today in the education sector, led the book to conclude, incorrectly, that the scope for further bureaucratic reform is limited.

Section 4 explains how community involvement can complement efforts at state-building, even beyond what the book envisages. However, it is argued that community involvement is best seen as a complement to state-building, not as a replacement to the latter.

Section 5 acknowledges that while the task of state-building is relatively well defined, there are a few important pages missing in the statecraft 'recipe book'. Two critical gaps which can hold back progress in the education sector are discussed. Firstly, the guidance on universal (or censal) national assessments is weak. Secondly, confusion when it comes to the role of indicators in a developing country context bedevils planning not just in education, but also other sectors. Surprisingly, little has been written about these gaps.

Section 6 concludes.

[1] James, 2014.

2 Fixing politics, bureaucracy and community involvement

The book, edited by Levy, Cameron, Hoadley and Naidoo, and freely available online[2], grapples with questions familiar to those who worry about getting the learning outcomes of a schooling system to improve. Where does the key fault lie? Are elected officials, the politicians, sending the right signals, and emphasising the right actions? Does the fault lie in the bureaucracy, among the public servants who are meant to be the specialists? Or is the problem the actions or inaction of people 'on the ground': teachers, the school principal, parents? Obviously, there can be critical malfunctioning in all three of these spheres simultaneously. Yet it is understandable that analysts would want to rank, prioritise, and attempt to identify where investing in greater functionality is likely to produce the best returns.

The book's position in this regard can be summarised, somewhat crudely, as follows. South Africa's transition to democracy, formally cemented in 1994, was an ambitious undertaking in which people looked to the best laws and institutional arrangements around the democratic world for guidance. The basic democratic governance arrangements were enshrined in the Constitution. Moreover, New Public Management, or NPM, approaches were strongly promoted by those constructing the post-apartheid state[3]. The new Public Finance Management Act was a central pillar in the new system. In addition, a variety of education sector-specific NPM-oriented policies were introduced to hold people accountable. In basic education, a key one was the teacher performance management system, the IQMS[4]. However, after a period of initial apparent success, disillusionment crept in as unions sought to focus narrowly on increasing teacher pay at any cost, and treated senior public service positions as rewards available to distribute among union leaders[5]. Beyond the unions, corruption and factionalism among politicians led to a weakening of the nascent post-apartheid state, particularly in provinces with a strong African National Congress (ANC) majority, such as Eastern Cape[6].

According to the book, the continued pursuit of policy and bureaucratic reform, while not an entirely worthless pursuit, is likely to disappoint, at least in a context such as South Africa's[7]. Of the two South African provinces studied in the book, Western Cape has reached a ceiling as far as bureaucratic reform is concerned[8], and Eastern Cape is unlikely to fix its bureaucracy, given the turbulence in the political sphere[9]. Bureaucratic effectiveness is simply too dependent on clean politics for individual activist-bureaucrats in Eastern Cape to make much of a difference. How one cleans up politics, in general or in Eastern Cape specifically, is not pursued in the book. What is pursued, and what is portrayed as a relatively dependable solution, is encouraging local-level accountability and action. In both Western and Eastern Cape, this solution holds promise, but for different reasons. In the case of Eastern Cape, local-level action, in particular close collaboration between the school's staff and parents, is likely to make the school's services less vulnerable to mishaps in the bureaucratic and political spheres. In Western Cape, where a very different situation prevails, local-level action can energise schools and bring about school-specific innovation in a context where a rather top-

[2] Levy, B, Cameron, R., Hoadley, U. & Naidoo, V. (2016). *The politics and governance of basic education: A tale of two South African provinces*. Manchester: Effective States and Inclusive Development. Available from: <http://www.effective-states.org/wp-content/uploads/working_papers/final-pdfs/esid_wp_67_levy_cameron_hoadley_naidoo.pdf> [Accessed November 2018].

[3] Levy *et al*, 2016: 12, 61.

[4] Integrated Quality Management System.

[5] Levy *et al*, 2016: 12, 56, 66-69, 128, 142.

[6] Levy *et al*, 2016: 124.

[7] Levy *et al*, 2016: 77.

[8] Levy *et al*, 2016: 111, 224.

[9] Levy *et al*, 2016: 144.

down, though also well-organised, bureaucracy lacks the required dynamism to realise change.

The book can be said to follow a comparative education approach in understanding systemic problems. This is an approach with an excellent pedigree and its own journal, *Comparative Education Review*, among the more influential journals in education. Not only are Western Cape and Eastern Cape compared, South Africa is compared to Kenya.

If there is a message for politicians in the book, it is that they should place education more firmly at the centre of the development agenda, as has been done in Kenya[10]. This is discussed below. SACMEQ[11] data are used to compare the Grade 6 scores of Western Cape and Kenyan students of a similar socio-economic status. The conclusion is that while Western Cape may perform well in the South African context, and certainly better than Eastern Cape, it does not compare favourably to Kenya. Children with similar socio-economic backgrounds perform around 0.3 standard deviations better in Kenya than in Western Cape – this is roughly a year's worth of learning[12]. This is a vital analysis which underscores an important point about South Africa's educational under-performance. The problem is not just one of weak schooling for the poor, though this is a large part of the explanation. There are also system-wide problems which lead to under-performance, in an international context, of even the better performing segments of the system. The wealthiest quartile of the Western Cape, which we can consider middle class, performs worse than Kenya's bottom quartile in mathematics[13]. It is good that the book highlights this remarkable pattern of middle class under-performance in South Africa, a reality which few education researchers in South Africa acknowledge, and one which is absent from the policy debates. But what are the system-wide problems which appear to hold back even the middle class?

The book's conclusion is that some of the reason for Western Cape's under-performance, in an international context, lies in low levels of parent involvement[14]. This seems supported by the book's analysis of SACMEQ data[15]. However, one needs to interpret this finding carefully. The parent involvement variables in SACMEQ are about contributing funds to the school, meaning there is a risk that what may appear to be the effects of parental involvement is in fact socio-economic status (SES) and parent education effects not captured in the separate SES variable used in the models. The findings of the book in this regard are noteworthy, but are not enough to conclusively pin Western Cape's under-performance, or even a part of it, on parent inaction. Below, alternative explanations for this under-performance problem are discussed.

3 Has South Africa exhausted the bureaucratic reform options?

The term New Public Management is now seldom used, at least outside academia. Yet the policy and systems solutions put forward nowadays by organisations such as the World Bank, UNESCO and the OECD clearly evolved from NPM, and can be considered part of the same public governance tradition. There are those who have argued that there have been major ruptures in this tradition, and that one is not dealing with a single tradition. Dunleavy *et al* (2006) would be an example. For them, e-government brings about an entirely new tradition. Yet the view that there has been considerable continuity seems predominant. Here, as in the Levy *et al* book, the term NPM is used to refer to a legacy widely embraced today, at least in principle if not always in practice, which promotes a law and order-driven bureaucracy, the

[10] Levy *et al*, 2016: 283.

[11] Southern and Eastern Africa Consortium for Monitoring Educational Quality.

[12] Levy *et al*, 2016: 158-161.

[13] Levy *et al*, 2016: 169.

[14] Levy *et al*, 2016: 112.

[15] Levy *et al*, 2016: 163.

widespread use of information and data for planning, and a strong emphasis on incentivising public servants, which would include teachers.

Did a serious attempt to entrench NPM in the newly democratic South Africa in the 1990s really buckle, or at least stall, under the weight of a worsening political layer, as argued in the book? It is argued below that, firstly, NPM-type interventions have continued to be developed to the present day, often to an increasingly ambitious degree. However, it is also argued that these interventions have been less successful than they could have been, not so much because politicians squashed them, but largely because they tended to be poorly designed by the 'technicians', or those bureaucrats involved in writing the policies. Even with a turbulent political layer, the bureaucrats could have done better. Examples of what has been achieved in some developing countries outside South Africa, even in a context of political turbulence, seem to confirm this. What the book, but also many commentators on education policy reform in South Africa seem not to have appreciated, is the full range of possibilities that have been open to South Africa.

In one area, public finance reform, NPM did succeed in South Africa. Public finance accounting systems were spectacularly improved. Work starting in the 1990s explains why, for instance, South Africa shares the top position in the world with New Zealand in the Open Budget Index (OBI) (115 countries were rated in 2017)[16]. Provincial education departments use standard charts of account; relatively well annotated budgets and annual financial statements are available online; a national computerised financial management system, the Basic Accounting System (BAS), works well; and Auditor-General reports are of a high quality. These achievements were to a large degree driven by one particularly effective organisation, National Treasury. However, progress in this area stands in sharp contrast to weaknesses in other areas of government when it came to taking forward the principles of NPM. In fact, the considerable success South Africa achieved in the area of *public finance* reform is probably why the book, but also many other sources, have come to the conclusion that South Africa was an exemplary adopter of NPM. This is true for public finance reform, but not other areas, such as the reform of education systems.

A further problem seems to cloud the book's evaluation of the history of NPM in education in South Africa, namely an insufficient focus on learning outcomes, and specifically systems to gauge learning outcomes. It is a common mistake to let questions of education inputs and processes obscure what really matters, namely what children actually learn. It could be argued that at a global level this mistake is what lay behind the very strong emphasis on simply enrolling children in the Millenium Development Goals, as opposed to ensuring that children acquire skills. This mistake has been largely corrected in the Sustainable Development Goals (SDGs), which took effect from 2015. The SDGs place a large emphasis on assessing the proficiency of children, and on making progress in areas such as reading competencies. By implication, this places the emphasis on public interventions such as teacher development and national assessments.

At times, a lack of a focus on learning outcomes is linked to scepticism around whether one can know much about learning outcomes, at least in a measurable sense. Can learning outcomes really be measured? Can progress over time be tracked? At one point, the Levy *et al* book[17] does express scepticism about the measurability of learning outcomes, but without substantiating this scepticism. Below it is argued that techniques to gauge learning, and progress in this regard, are now widely available, though gaps remain.

In South Africa, a strong emphasis on inputs from 1994 to around 2007 was underpinned by the need to abolish the extremely unequal public funding of students inherited from apartheid,

[16] International Budget Partnership, 2018.

[17] Levy *et al*, 2016: 100.

a project that took almost a decade to complete[18]. Moreover, there were few systems inherited from the pre-1994 era which focussed on learning outcomes. The only one was the Grade 12 examination system, which has continued to be a centrepiece of the schooling system to this day. Many would argue it has been over-emphasised, to the detriment of other areas of system reform. An examination and qualification in Grade 10, the Junior Certificate, had existed, but had been abolished in the 1970s, although at the time the great majority of youths were not surviving school long enough to achieve the Grade 12 qualification, meaning all these youths ended up leaving the schooling system with no qualification. This problem persists to this day, though its magnitude has shrunk a bit – today around 45% of youths never obtain a national qualification[19].

The book's focus on education-specific NPM interventions is limited largely to human resource management systems, such as the IQMS, which focusses on compliance with professional development, time-on-task, and workplace rules[20]. This is of course important, but is not at the core of what is considered performance management in, for instance, UNESCO's Global Monitoring Reports[21], or the 2018 World Development Report of the World Bank, which focusses specifically on reforming schooling systems. These 'textbooks' promote strongly a focus on promoting learning outcomes. This is difficult to achieve through systems that hold individual teachers accountable. Even in the United States, where several attempts have been made to link the pay of individual teachers to performance, such accountability remains controversial and technically difficult. To illustrate the latter, one way of tackling the confounding effects of the fact that one student is often taught by many teachers, is to independently test students at the start of a year, and then again at the end of the year, and to gauge the 'value added' of the teacher. However, this is only really practical where children are taught by just one teacher during the year, a situation which often only prevails at the lower primary level[22]. In most contexts, it is impossible to link a student's academic trajectory to a teacher in this one-to-one fashion. At best, one can hold teams of teachers accountable for improvements in learning outcomes in the school as a whole. It has taken a while for this understanding to take root. Today, organisations such as UNESCO focus on 'national assessments', which may be sample-based or reach every school. Where they reach every school, information from these assessments may be used to hold school principals and the school's teachers as a team accountable. As discussed in section 5, the tendency of UNESCO and others to gloss over the complex differences between sample-based and censal (or universal) assessments has been problematic, and has contributed to poor decision-making.

Though the book barely focusses on South African systems dealing specifically with learning outcomes, it presents, three pages from the end, a particularly insightful observation on why Kenya produces such relatively good learning outcomes. The words are those of Benjamin Piper, an education expert who has worked for many years in Kenya:

> What one sees in rural Kenya is an expectation for kids to learn and be able to have basic skills … Exam results are far more readily available in Kenya than other countries in the region. The 'mean scores' for the Kenya Certificate of Primary Education (KCPE) and equivalent KCSE at secondary school are posted in every school and over time so that trends can be seen. Head teachers are held accountable for those results to the extent of being paraded around the community if they did well, or literally banned from school and kicked out of the community if they did badly.

[18] Gustafsson and Patel, 2006.

[19] See for instance 'The flaw in SA's 'real' matric pass rate figure' at https://africacheck.org/spot-check/the-flaw-in-sas-real-matric-pass-rate-figure-as-calculated-by-the-eff-da.

[20] Levy *et al*, 2016: 69, 97.

[21] In particular the GMRs of 2013-2014 and 2017-2018.

[22] Dee and Wyckoff, 2015.

While letting school communities expel school principals in this manner is hardly NPM, the accountability systems advocated by UNESCO, the World Bank and others are essentially modernised, fair and regulated versions of this. There should be processes whereby patently and persistently under-performing school principals should be removed. In South Africa, a primary school principal can only be removed on the basis of non-compliance with human resources regulations, relating for instance to attendance or proper personal conduct in the school. A principal could not be removed because of unacceptably low levels of learning among students, because there are no systems to monitor this. In fact, South Africa is the only country in the Southern African Development Community (SADC), apart from Angola, not to have some examination or national assessment at the primary level[23]. This could explain why the entire schooling system, even the part serving primarily middle class communities, under-performs in an international context.

Apart from the arbitrariness of the punitive measures described in Piper's observation, there is a further aspect of the Kenyan arrangement (at least as it is described above) which is unmodern. Examinations, the UNESCO warning goes, are generally poor gauges of progress, for instance because the difficulty of examination papers varies between one year and the next. Instead, national assessments which repeat 'secure', or secret, test questions across years, combined with statistical scoring methods using item response theory (IRT) are needed[24]. This is sound advice, but as argued in section 5, the proponents of national assessments have been vague on critical technical questions.

The work needed to build effective systems that monitor learning outcomes requires specific skills at a national level, but also astute politicians who can convince teacher unions to accept new assessments. Here it is important to disentangle the various sources of disagreement. The area of learning assessments is a highly politicised one. Unions will often oppose the introduction of new assessments for three reasons, one of which is bad, one completely understandable, and a third highly debatable.

Firstly, unions may not want the under-performance of the schooling system, and the teaching profession, to be placed under the spotlight, in part because this could weaken their bargaining power when salaries are negotiated.

Secondly, there are reasons to be suspicious of government attempts to assess students given several examples where new systems, even in developed countries, have been poorly designed, resulting in unfair judgements of under-performance, or simply confusion. System failures in the United States have been particularly well-documented and feature prominently in the resolutions of Education International, the world federation of teacher unions (and the largest labour federation in the world). A 2016 article by the Washington Post[25] listing assessment system blunders by Pearson, a company used extensively by government to deliver testing services to public schools across the United States, can be considered a useful account for education planners of what *not* to do. If testing is outsourced, far better controls over the service provider than those that have governed Pearson's work are necessary. Above all, system errors cannot be allowed to produce incorrect scores for students. If tests have real consequences for individuals, such as grade promotion, quality controls of the information must be particularly rigorous. Fortunately, and as outlined below, there are examples of assessment systems which have worked well, and have been properly documented[26].

[23] UIS database of assessments and examinations at http://uis.unesco.org/en/uis-learning-outcomes.

[24] UNESCO, 2014: 90; UIS, 2017a.

[25] Article headed 'Pearson's history of testing problems - a list' at https://www.washingtonpost.com/news/answer-sheet/wp/2016/04/21/pearsons-history-of-testing-problems-a-list/?utm_term=.5612f65a9f14 [accessed February 2019].

[26] UNESCO, 2014; OECD, 2013; OECD, 2015. PDF versions of the last two can be obtained from the author of the current report.

The third reason why unions have tended to be apprehensive about new national assessment systems is that such systems, as well as international standardised testing programmes, have been seen as part of a neoliberal package of reforms which includes the casualisation of teacher employment and the privatisation of schooling. This comes across strongly in the Education International (EI) resolutions of 2011 (formulated at a congress in Cape Town), though the EI's 2015 World Congress resolutions are noticeably less critical of assessments. One way of challenging this ideological concern is to point to the many socialist and social democratic countries which have emphasised information on learning outcomes and accountability strongly in their education policies. One example would be Cuba, by far the best performing Latin American country in the LLECE[27] international testing programme[28]. (Cuba has no representation in EI as teacher unions in the typical sense do not exist in Cuba.)

What statecraft did take place in democratic South Africa to advance the monitoring of learning outcomes? Two initiatives stand out, neither of which was a success. Understanding why is important. The first initiative was the Systemic Evaluation, a sample-based national assessment programme which was run just three times: 2001 for Grade 3, 2004 for Grade 6, and 2007 for Grade 3 again. While the actual assessment and data collation part of the programme seemed to function relatively well, what did not work well was the analysis of data, and dissemination of findings. In particular, the Grade 3 trend between 2001 and 2007 was barely made public, though what little is available[29] suggests that there was considerable improvement, a trend which would be in line with 2002 to 2011 improvements seen at the secondary level according to the international TIMSS[30] programme (these improvements continued beyond 2011)[31]. Not disseminating this information about primary-level trends, and not speculating, using the background data collected through the programme, on what the causes were for the improvement, clearly represents a missed opportunity. Not only was it a missed opportunity for the government to showcase progress. Reassuring the public that there is progress can prevent instability and unnecessary policy and curriculum change, things schooling systems are prone to. Essentially, if progress can be demonstrated, the incentive to change existing policies is reduced. Why was the information not disseminated? Discussions with relevant people suggest the problem was both limited technical capacity to analyse and report on the data in what was then the Department of Education, and limited political interest in, or understanding of, the programme.

The second initiative was a censal, or universal, national assessment. The Annual National Assessments programme ran each year between 2011 and 2014, for four years. This programme, which became almost as widely reported on in the media as the Grade 12 examinations, did enjoy considerable political and public support. However, it suffered serious design flaws which ultimately eroded support for the programme. Above all, it did not fulfil one of the requirements put forward by UNESCO and others for a national assessment. It was not designed to produce sufficiently comparable results over time to allow for reliable trends at the school, province or national level. This shortfall was explained in a few places, for instance in the 2013 national ANA report[32]. Yet many education administrators, the government, and the media treated the programme as if it is *was* producing reliable trend data. UNESCO entered ANA into its list of national assessments, though the programme lacked a key element of a national assessment, as defined by UNESCO, namely the ability to measure

[27] Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (Latin American Laboratory for Assessment of the Quality of Education).

[28] Gasperini, 1999; UNESCO, 2008.

[29] The only publicly available document seems to be a twelve-page report, untitled, which was previously available on the Department of Basic Education website but now seems only available on the blog of the education academic Nic Spaull – see https://nicspaull.files.wordpress.com/2011/04/doe-2008-systemic-evaluation-grade-3-2007se-leafle.pdf [accessed February 2019].

[30] Trends in International Mathematics and Science Study.

[31] Reddy *et al*, 2016.

[32] Department of Basic Education, 2013: 28.

the change referred to in the following (from the 2013/4 Global Monitoring Report of UNESCO[33]):

> National assessments should be a diagnostic tool that can establish whether students achieve the learning standards expected by a particular age or grade, *and how this achievement changes over time* for subgroups of the population. [Italics not in the original.]

Lack of clarity around the purpose of ANA, and the absence of technical elements which were popularly believed to exist, meant the results emerging from the programme were easily misunderstood or misused, and strengthened the hand of those opposed to ANA specifically, or standardised assessments in general[34]. In 2015, the largest teacher union, with considerable support from other stakeholders, succeeded in halting the programme. South Africa cannot be entirely blamed for this failed attempt at statecraft. Guides produced by UNESCO and the World Bank have been frustratingly vague on key technical aspects of these programmes, for instance on how 'secure anchor items', or questions repeated across years which are kept secret, can be implemented in universal national assessments that cover thousands of schools. If anything, *this* was a serious shortcoming in the NPM 'toolbox' which the Levy *et al* book could have paid attention to.

Surprisingly, the Levy *et al* book barely discusses the Annual National Assessments (the Systemic Evaluation is not mentioned at all). There is another important assessment initiative, a provincial one, whose statistics the book uses: the Western Cape's Systemic Tests programme. Unfortunately, however, the book does not discuss this programme's strengths and weaknesses as a product of statecraft, or its managerial and accountability impacts. The programme is a particularly interesting one, and arguably the most successful assessment system in South Africa. It is censal insofar as it covers all Western Cape schools. Though the tests seem to change somewhat over time, often the very same test is used in one year and the next. The way the programme achieves the required confidentiality of test items is by controlling closely, through test administrators who are external to the school, that no tests are 'leaked' in any year. School-specific results are given to schools, but not published in a 'league table', as is done in some schooling systems. Results are expressed as simple classical scores, and not IRT scores, which is not too problematic where whole tests are reused over time. Results are not communicated relative to the school's socio-economic status, something which is problematic as it is not fair to compare, say, a school catering for the children of professionals to a school serving students from a poor background. Yet the programme does to an extent seem to fulfil its purpose of getting principals and their managers in the district to pay close attention to learning outcomes. Yet, as will be argued in section 5, the Western Cape's programme represents a difficult and risky way of gauging progress across all the schools of a system. There are more effective methods, though they are not widely understood.

The ANA programme has probably not been a fruitless investment. Apart from yielding valuable lessons on how *not* to design a national assessment programme, the programme, for all its flaws, helped to place learning outcomes high on the education agenda. It raised awareness among teachers and society about the importance of worrying about what younger children can and cannot do. Very importantly, the TIMSS, SACMEQ and even PIRLS programmes have pointed to substantial test score improvements since 2002[35]. What lies behind these improvements? A stronger emphasis on monitoring learning outcomes and achieving at least comparability across schools at any one point in time, even if not over time, through ANA, were likely contributors. Better funding of books for students, and reforms

[33] UNESCO, 2014: 6.

[34] Government's own acknowledgement of these weaknesses can be seen in Department of Basic Education (2016).

[35] Department of Basic Education, 2019: 5-6; Van der Berg and Gustafsson, 2017.

which made the curriculum easier for teachers to follow, are probably further contributing factors.

While South Africa's attempts at national assessments have delivered less than they could have, other developing countries have made the attempt, and appear not to have experienced such serious problems. Why? There are indeed many other countries South Africa could be compared to. Other countries with *sample-based* national assessments, according to a UNESCO list, include: Côte d'Ivoire, Ethiopia, Kenya, Lesotho, Malawi, Mexico, Mozambique, Nigeria, Peru, Tanzania and Zambia, to name a few. Countries with *censal* national assessments are fewer, and include Botswana, Mexico and Namibia.

So where has South Africa gone wrong? The preceding discussion has suggested that the technical capacity needed for this kind of statecraft has not been what it should be. Some of the explanation probably lies in the fact that other developing countries tend to draw to a greater degree on expertise from outside the country. Ironically, this may be due to a comparative advantage in the case of South Africa: relatively healthy public finances (at least up till recently) and a low national debt. Of the 119 countries with total external debt data in the World Bank's indicator database, only two countries, China and Afghanistan, had a lower external debt, relative to the size of the economy, during the years leading up to 2011 (since 2011 South Africa's situation has worsened a bit, with its ranking slipping from 3 to 15)[36]. Moreover, external development aid to South Africa has historically been low. Since 2000, this has never exceeded 1.7% of central government spending, against over 4.0% for Thailand, Indonesia and Peru, 14% for Namibia, 19% for Botswana and Kenya, and 53% for Uganda[37]. As a consequence, South Africa has been under less pressure than many other developing countries to make use of experts from foreign aid agencies, or the World Bank. This has to some degree isolated South African experts from their non-South African counterparts. In future, South Africa's status, together with Brazil, Indonesia, India and China, as a 'key partner' of the OECD, is clearly one route towards stronger international exchanges in the development of assessment systems, and in other areas of education policy design.

What appears difficult to substantiate, is the argument that malfunction in South Africa's political sphere, at least in provinces such as Eastern Cape, is so severe that NPM-type reforms are rendered unworkable. Several countries with worse corruption indices than South Africa have been successful implementers of education reforms. Kenya has fared much worse than South Africa in terms of the Worldwide Governance Indicator (WGI) 'control of corruption'[38]. Brazil has in nearly all years in the last two decades been rated as more corrupt than South Africa against the same WGI indicator. Yet large improvements in Brazil's learning outcomes have been ascribed to the introduction of a national assessment programme viewed by World Bank analysts as 'superior to current practice in the United States and in many other OECD countries in the quantity, relevance, and quality of the student and school performance information it provides'[39].

Statecraft in the schooling sector in South Africa can indeed be considered weak. However, the sector has experimented with important innovations to a far greater extent than what is reflected in the Levy *et al* book. There is nothing to suggest that South Africa must fail in future. The National Development Plan (NDP), published in 2012, presents through its 'results oriented mutual accountability' framework[40], a credible answer to the book's core

[36] The World Bank indicator 'External debt stocks (% of GNI)'. The maximum seen per country up to 2011 was compared. All the 119 countries are developing countries.

[37] The World Bank indicator 'Net ODA received (% of central government expense)'. The maximum annual value seen per country during 2000 to 2017 was compared.

[38] Gustafsson and Nuga Deliwe, 2017: 16.

[39] Bruns *et al,* 2012: 7.

[40] South Africa: National Planning Commission, 2012: 311.

question on how to balance hierarchical and horizontal approaches in governance. The NDP's chapter on education is very much in line with global thinking, and the NPM inheritance, on how to strengthen educational quality. Crucially, at the heart of the NDP's envisaged accountability system is information on learning outcomes. What learners learn is what people must ultimately be accountable for. But just as crucial is the NDP's emphasis on having a clear framework on what this information can be used for, and what it cannot be used for.

To conclude, for those interested in statecraft and NPM in the education sector, there is a wealth of guidance and experience to draw from, though there are gaps in the 'toolbox', in particular in relation to censal national assessments. Unfortunately, the range of options is often poorly understood. The Levy *et al* book makes the common mistake of under-estimating the importance of monitoring learning. How this monitoring differs across South Africa and Kenya is barely explored, yet it seems likely that these differences explain much of South Africa's relative under-performance. The book's conclusion that NPM-type reforms have largely failed in South Africa is in part correct, but not for the reasons put forward in the book. The central issue is not that human resources management systems have failed to deliver. The central issue is rather that systems to monitor learning outcomes have been poorly designed, or suffered from missing parts. These problems can be fixed. NPM remains a worthwhile pursuit.

4 Community involvement as a complementary measure

Levy *et al*, in describing the value of the 'short route' in education sector accountability, specifically accountability of schools to the communities they serve, draw from the widely quoted 2004 World Development Report of the World Bank, titled *Making services work for poor people*[41]. This report makes the point commonly made when opportunities associated with school accountability to communities, and school autonomy, are discussed: 'schools cannot be given autonomy unless they are given clear objectives and regular assessments of progress'[42]. Without sufficiently good information on learning outcomes, communities will not be in a position to know whether their schools are truly working for them. The World Bank report acknowledges risks associated with assessments, such as 'teaching to the test', but also makes the point that if learning outcomes are weak, as they often are in schools in developing countries, then even such teaching can constitute an improvement[43].

The Levy *et al* book acknowledges that the empirical evidence on the efficacy of reform focussing on accountability to parents is mixed[44]. The literature they draw from deals largely with the experimental and quasi-experimental evidence, in other words evidence emerging from groups of control and 'treatment' schools. What is also worth considering is the evidence based on across-country differences using data from the international testing programmes. Wößmann (2005: 162) finds that both TIMSS and PISA data reveal the same important pattern at the secondary level[45]. School autonomy, in other words a reduced reliance on traditional vertical lines of control and accountability, is only associated (in the sense of conditionally correlated) with higher learning outcomes if external examinations exist. Put differently, empowering schools is good, but it requires there to be good and comparable information on what students learn. This is understandable. Communities and committed school principals need to know whether what is occurring in the classroom is leading to learning outcomes which are satisfactory, relative to how well the schooling system as a whole performs.

[41] Levy *et al*, 2016: 88.

[42] World Bank, 2003: 114.

[43] World Bank, 2003: 120.

[44] Levy *et al*, 2016: 8.

[45] It is clear that the values appearing in Figure 3(b) in Wößmann (2005) are not the right ones, and are not in line with the vertical axis of the graph. This basic error was corrected in a reproduction of this graph in Hanushek and Wößmann (2007: 18).

Effective central examinations are among the most important manifestations of a capable education authority. Devolving powers to schools and communities requires relevant information systems to already be in place. In fact, South Africa has a well-established national examination in Grade 12, and this is used extensively by the authorities, and to some extent communities, to hold secondary schools accountable. On the other hand, at the primary level, there has been an almost complete absence of standardised information on learning outcomes outside of the short four-year life of the Annual National Assessments (as explained above, Western Cape is an exception due to the existence of a provincial assessment system). This stark contrast between the primary and secondary levels in South Africa represents an interesting opportunity for researching how accountability works, or does not work, in the schooling system. It is an opportunity which has hardly been explored by researchers. In fact, too often the South African research generalises about schools, when the primary and secondary levels display rather different dynamics. Levy *et al*, it should be noted, are essentially examining *primary* schools in South Africa, though conclusions are couched as if schools in general are covered.

Of course the problem, already mentioned above, that examination results are often poor measures of relative school performance or progress over time must be taken into account. In South Africa, Grade 12 examination indicators, in particular the ‘pass rate’, meaning candidates who achieve the certificate over all candidates entering the examination, is easily manipulated. Schools or whole provinces can and do control who enters the examination to manipulate the indicator. But does this mean examinations should play no role in accountability? Should they be completely discarded as measures of institutional quality or systemic progress? Probably not. Extracting more reliable indicators from examination is technically complex, and can result in indicators which are difficult to explain to the general public. But the task is not impossible, as can be seen from Gustafsson’s (2016) analysis of South Africa’s Grade 12 results. The lack of guidance in this regard in the manuals accessible to education planners can be considered another missing page in the statecraft ‘recipe book’.

To conclude this section, if the complementarity of, on the one hand, centrally run systems, such as examinations or standardised assessments, and, on the other, local action are considered, the scope for community involvement is wider than what Levy *et al* envisage. In the case of Western Cape, for instance, the province Levy *et al* consider to have hit a ceiling as far as hierarchical innovations are concerned, key questions should be asked around the role of the province’s systemic tests in promoting horizontal accountability. Historically, the province has not required results to be shared with parents. But do schools perhaps share the results on a voluntary basis? It is very likely that this occurs in school governing body meetings, which parents attend. How do these results influence the conversation between parents and school staff? And so on. Naturally, assessment systems are not the only interventions mandated through the hierarchical route which are of interest when considering accountability to communities. School funding systems, provincially run building programmes, and even personnel performance management systems are of interest to parents. But assessments and learning outcomes should play a central role in horizontal accountability.

5 Gaps in the typical policymaker toolbox

The statecraft ‘recipe book’ for a schooling sector, particularly one in a developing country where planning capacity is often weak, is by no means complete. For a number of key questions, planners have little in the way of manuals to turn to, and must often learn by doing, which can imply costly mistakes. If one speaks to education planners, where the gaps lie quickly becomes clear. One gap is in the area of teacher supply and demand. A relatively comprehensive manual for this was produced half a century ago by the International Institute for Educational Planning (IIEP)[46]. No-one appears to have produced something more recent,

[46] Williams, 1979.

taking into account recent developments in terms of EMIS[47], payroll systems, and the capabilities of spreadsheet tools such as Excel.

Below the emphasis falls on two gaps. In each instance the gap is explained, and how one might fill this gap then discussed. The first gap is the paucity of guidance on censal assessment systems, in particular guidance on how to make the results of such systems comparable over time. The need for such systems is clear in the literature, but how to design them is unclear. This is dangerous, as there are many pitfalls, as witnessed in the case of South Africa's Annual National Assessments (ANA) programme. The second gap affects not just schooling systems, but public services in general. A planner will easily find manuals on how to organise sector planning. Often these manuals come in the form of instructions emanating from an organisation such as the Ministry of Finance. There are manuals that talk about the importance of 'SMART'[48] performance indicators, and of linking inputs to processes and outcomes. However, as will be argued below, these manuals tend to be weak in the sense that they are overly theoretical and have not taken into account institutional realities, such as data quality and human capacity. The consequences have been serious, and include malicious compliance on a grand scale. Malicious compliance can be described as conforming to the rules with anger, perhaps to the point of wishing for overall failure, due to a belief that the rules get in the way of success, even though officially they are said to facilitate success. How better guidance and rules aimed at planners might be crafted is discussed.

5.1 How to make censal assessments sufficiently comparable

The World Bank, the International Institute for Educational Planning (IIEP) and UNESCO Institute for Statistics (UIS) have all published advice of some kind to guide developing countries in the area of assessments. All have glossed over important technical details which planners should be aware of. One can assume that this has slowed progress in the design of national assessments, and led to costly mistakes. Here the focus falls on insufficient clarity on how censal assessments can bring about sufficient comparability over time.

Between 2008 and 2015, the World Bank released a five-volume set of manuals on assessment, coming to just under a thousand pages in total. The first volume mentions that '[s]ome national assessments have used both census- and sample-based approaches', Costa Rica and France being among the examples of this. The details of this are not explained, but it is suggested that sample-based national assessments are preferable, while it is also acknowledged that censal national assessments tend to be popular with politicians. The methods presented in the manuals are applicable to sample-based assessments, and not necessarily censal assessments. But this is not very clear, and the dangers of applying some of the methods to censal assessments are not explained. The differences between sample-based and censal assessments are presented largely as a matter of scale and cost[49], when the differences are in fact more complex. But even for a planner working on a sample-based assessment, the manuals lack vital details. Crucially, the volume on test design, consisting of 190 pages, has nothing on anchor items linking tests across years. In this respect, the manual actually only provides guidance on how to run a national assessment once, not how one designs a system to track progress over time. The same criticism can be made of an IIEP guide on test design[50].

A 2017 guide by the UIS at least raises a lack of sufficient comparability over time as a major risk in a national assessment[51]. However, why this problem often arises is not explained, nor

[47] Education Management Information System.
[48] Specific, measurable, accepted, relevant and time-bound.
[49] Greaney and Kellaghan, 2008: 31.
[50] Izard, 2005.
[51] UIS, 2017b.

are ways of mitigating the risks, which would be different for sample-based and censal assessments. The technical complexities associated with choosing between the two types of assessments are glossed over in the UIS guide too.

The basic challenge that is not addressed is the following. First, it should be made very clear that repetition of whole tests or specific test items is essential if sufficient comparability over time is to be assured, and if small improvements are to be detected. It is important to remember that historical trends suggest that annual improvements across an entire system, if they occur at all, are small, meaning fine-tuned monitoring is required. Fortunately, if one uses a representative sample of schools, very few schools are needed to provide statistically reliable trends. To illustrate, to obtain nationally representative statistics on the reading skills of nine-year-olds, the National Assessment of Educational Progress (NAEP) in the United States has sampled fewer than 200 schools a year, out of a total of around 90,000 primary schools[52]. High levels of inequality push the requirement up somewhat, for instance to around 300 schools in the case of highly unequal South Africa. With so few schools, it is not difficult to repeat tests or items, year after year. External test administrators can come to the school, hand out tests to students, monitor the process, and then collect all tests from students at the end. Ensuring that tests are not 'leaked' is feasible. If tests are leaked, that can be disastrous for the comparability of results over time. Even if only a few schools access and make use of the leaked tests, the fact that there is uncertainty over how seriously the leaked tests distorted the results would be devastating for trust in the trends seen in the results.

Clearly, applying the same level of security if one is dealing with 90,000 schools would be prohibitively costly, and probably impossible. It becomes very difficult to avoid 'leaks'. One compromise would be to run both a sample-based assessment, to gauge national trends to a high degree of accuracy, and a separate censal system where it was accepted that at times school-level values would be weakly comparable over time. If there is too much uncertainty in the national trend, then this trend is not able to guide national policy debates as it should. However, some uncertainty in a school's trend, especially if one makes it clear to everyone what the likely margin of error is, has less serious consequences. Yet the trends of individual schools need to be sufficiently reliable to be meaningful for basic management and accountability purposes, and such reliability is not achieved through tests which are completely different every year, even if test designers do their best to achieve equivalent levels of difficulty. The challenge seems almost insurmountable.

Before solutions are discussed, it is interesting to note that the lack of clarity on these technical aspects of assessments could lie behind what appear to be 'isomorphic mimicry' in several developing country national assessment systems. Isomorphic mimicry, a term coined by Lant Pritchett, and employed in Levy *et al*, refers to a tendency in countries with weak institutional capacity to mimic the capacity of the most developed countries, for instance through employment of the accepted jargon, and then to conceal the fact that whole systems are dysfunctional or not operating as they should. According to Pritchett[53], this type of mimicry has been exacerbated by donor funders in regions such as Africa.

To illustrate, a 2016 report by the Kenyan authorities on the sample-based National Assessment System underlines that item response theory (IRT) was employed, and that results from two waves of the national assessment, 2010 and 2015, are comparable to each other, with no improvement seen over time for the country[54]. IRT is crucial when results from tests

[52] Estimate of the school sample derived from tables A-2 and A-4 in United States: National Center for Education Statistics (2005). The number of schools, as opposed to the number of students, sampled in NAEP's long-term trend (LTT), seems not to be made explicit in any public source. A figure of just under 200 schools is not too far from the number of schools sampled in the United States for PIRLS Grade 4, which was 369 schools in 2016 (own analysis of the PIRLS microdata).

[53] Pritchett *el al*, 2012.

[54] Kenya National Examination Council, 2016: 115, 133.

from different points in time with a mix of repeated and non-repeated test items are placed on a single scale, to allow for comparability over time. The Kenyan report might initially create the impression that IRT was used for this purpose, but the details suggest that IRT was used only to make comparisons across students *within* 2015. In fact, the report does not provide reassurance that the 2010 and 2015 results are comparable. Perhaps the necessary technical work for this was performed, but then the question is why one would not report on this, given that other aspects of the technical work were reported on.

Turning to Brazil, materials publicly available on the web relating to the censal assessment system, Prova Brasil, raise serious questions[55]. Prova Brasil is said to make use of secure anchor items to achieve comparability over time across all schools. The question is how this security would be achieved, in a developing country context, for over 30,000 schools per grade. Either a spectacularly successful and extremely costly logistical operation has essentially not been noticed, for instance by the World Bank, and is not reported on in official Prova Brasil reports, or school-level results are not as comparable over time and across schools as they are made out to be.

In South Africa, the official report for the 2012 Annual National Assessments programme indicated that anchor items were used to make the 2011 and 2012 tests comparable[56]. ANA tests from all years were widely available as they stayed with schools after schools administered them. Had questions been repeated, this would have been widely discussed, and would certainly have been used by the detractors of the programme to discredit it, as it could have been argued that in preparation for the 2012 testing teachers drilled identical questions from the 2011 tests. Moreover, had anchor items been used to bring about comparability over the two years, 2012 results would have had to be rescaled, using a technique such as IRT. It is clear from the 2012 report that such rescaling did not occur, and that results appearing in the report were unadjusted classical scores. Anchor items could not have been used to equate 2011 and 2012 scores, yet someone with just a basic familiarity with assessment techniques might not have understood why this was not possible.

The problem of a veneer of technical sophistication concealing what is really occurring is not limited to developing countries. The global PISA and TIMSS reports are impressive as far the details around the processing of data submitted by countries are concerned. However, what has remained to a large extent a 'black box' is the processes undertaken by data collection agencies in each country. The details on this are seldom published. Jerrim (2013), in trying to understand why England's TIMSS and PISA trends were pointing in completely different directions, uncovered irregularities in the England sampling and governance processes which seemed to explain the discrepancies.

There is one place where the solutions to the challenge of comparable censal assessments can be found: the technical documentation of Australia's National Assessment Programme (NAP). NAP tests students across all schools in Australia, using a completely different set of tests each year. Every student in the same grade and year writes exactly the same test in a specific subject, and on the same day. After the test day, it does not matter if tests are widely accessible. Year-on-year comparability is achieved through an 'equating sample'. Some days before the censal test, the approximately 1,000 students in the equating sample, which spans around 40 schools, write a secure test with secure items repeated from year to year, as well as items from the tests applied censally. Thus, within each grade and subject, around 1,000 students provide a link between the current year and previous years, and this link is used to adjust the scores of all students across the system.

[55] Brazil: Ministério de Educação, 2011, 2015a; 2015b; De Andrade *et al*, 2000.

[56] Department of Basic Education, 2012: 11.

The benefits of this over the approach followed by Western Cape should be clear. In the Western Cape, the authorities must devote resources to securing tests across all of the province's approximately 1,500 schools. Australia needs to do the same for just 40 schools per grade and subject. This comes to around 250 schools for the country when all tested grades and subjects are considered. But arguably more serious than the higher cost, is the higher risk Western Cape runs that security will be breached and comparability compromised.

5.2 Not seeing the wood for the indicators

Not seeing the wood for the trees refers to an inability to grasp the overall picture due to an excessive attention to individual details, and a lack of attention to the inter-connectedness of the details. This appears to be a useful metaphor to describe a problem found in government planning and reporting systems across the world, but in particular in developing countries.

Crudely, and with a dose of poetic licence, the problem can be characterised as follows. Planners in a national Ministry of Education are told to produce medium-term plans for the sector, using templates and formats prescribed by some government-wide agency. These templates contain grids where each row is an activity, and columns cover things like purpose, expected outcome, risks, responsibility, indicators, and targets. The planners are instructed to include a lot of detail, on matters ranging from teacher training, to school building programmes, to materials development, and so on, the assumption being that a wealth of details is needed if everyone is to be held accountable. No-one should escape the accountability net. This is a completely noble goal. However, by focussing on a multitude of 'trees', and given a format which does not lend itself to thinking about inter-connections and the whole, the work becomes a compliance-oriented 'fill in your cell in the template' exercise. There is a vague sense that the exercise is not a terribly productive one, or proper planning, but no-one wants to challenge the process. These are the rules which must be followed. This is in part because capacity in areas such as statistics and economic or educational development is limited, meaning no-one is prepared to challenge the templates, which become the de facto science. The templates require the maintenance of a large set of indicators, because the details must all be measured. There may also be a requirement that each indicator have a formal technical description, and for that there is yet another template.

As inputs come in from various sections of the Ministry, planners get ready to do a lot of copying and pasting. Difficulties relating to indicators result in a large proportion of time being spent on these. Excel becomes the predominant medium, because it seems amenable to the task. Many meetings and phone calls are devoted to resolving a range of disputes. Is an indicator on whether policy X has been promulgated yet counted as a SMART indicator (this is not clear in the manuals)? Is it acceptable to insert an indicator on number of meetings held per quarter, to comply with the requirement that reporting and targets must follow a quarterly cycle (this is after all an easy indicator to report against, even if it not very relevant)? Does the planning branch understand how much time is going to be taken off core work for the maintenance of more complex indicators? For baseline indicator values, can rough estimates based on the available, and not entirely consistent, data be used? And then there is the question of targets. Officials who are close to the work are continuously pushing back against what they see as overly ambitious targets. The more removed the senior official is from the actual work, the more ambitious the target, with politicians insisting on the most ambitious targets of all.

The result of this process tends to be long plans, with many 'trees' in the form of indicators and lists of tasks to be done, but with limited coherence across these parts. The 'wood', such as improving educational quality, is often lost in the maze. This confusion is costly. As argued by Greaves and Achterstraat (2015: 107), 'we cannot afford to become lost in a morass of information – we need to identify the important qualitative and quantitative indicators'.

It is remarkable that the problem has not been better covered in the literature. Planning, especially in a developing country context, lacks a firm empirical and theoretical basis. Pritchett has made compelling arguments about the magnitude and nature of the problem. More analysts need to take this work forward, and examine solutions. The above characterisation is based on observations made within South Africa, yet an examination of national plans available in the IIEP's online repository[57] suggests that similar problems are common across many developing countries. And they manifest themselves not just in official plans, but also reports.

Below, the problem, and possible solutions, are discussed more formally, with reference to some of the literature that exists on the topic. It should be emphasised that sector plans and reports are used here as a lens for viewing the larger problem of weak governance. Such a focus comes with limitations. For instance, organisational capacity, structures or culture are not directly dealt with. Yet plans and reports seem to present a particularly useful window onto the broader problem. All education administrations produce such documents, meaning they provide an easy basis for comparison. The obvious caveat, however, is that producing better looking plans and reports, for instance by using external consultants, cannot be equated with better governance. Yet better governance is likely to be manifested through better quality documents.

The discussion that follows centres around five pitfalls: forgetting outcomes; excessive simplification; indicators set in stone; indicator overload; and the targets problem.

Forgetting outcomes. Giving greater prominence to outcomes is relatively easy to realise, because everyone agrees this should happen (the same cannot be said about the other four pitfalls). Much of the challenge lies in deciding how to measure progress over time through learning assessments with sufficient reliability (and in grappling properly with technical issues discussed in section 5.1). But how should one cut through the typical silo effects and ensure that curriculum, finance, physical infrastructure and human resources specialists increasingly see their work as contributing to better learning outcomes? One way of doing this is to make explicit what the assumed theory of change for the sector is[58]. What is the assumed impact of teacher time-on-task, classroom pedagogy, educational materials, school nutrition, and class size on learning outcomes? What is the assumed impact of various cultural factors: professionalism, accountability, rule of law, constitutionalism, respect for human rights? Hardly any education plans make their theory of change explicit.

Excessive simplification. Despite the length of South Africa's education plans and reports, they simplify complexities relating to both the way education service delivery occurs, and the data available to planners. This compromises the value of the documents. A 2017 review of data use in South Africa's basic education system, commissioned by National Treasury, confirms that South Africa's documents are indeed long. Annual reports from two South African provinces were found to be about twice as long as those of New South Wales (in Australia) and New Zealand[59]. The review concludes that while guides for planning produced by National Treasury between 2007 and 2011 were an important step forward in creating a common 'planning language' across government, there is a need to update them, on the basis of lessons learnt. To illustrate what should be avoided, KwaZulu-Natal's plan from one year uses a high Grade 12 'pass rate', or percentage of examination candidates obtaining a national certificate, achieved in the last year, as a basis for arguing that quality is on the rise and that certain interventions are working well. The following year's plan, however, using the following year's pass rate, which is much lower, argues that the system is struggling and that

[57] https://planipolis.iiep.unesco.org.

[58] Pritchett *et al,* 2012: 14.

[59] South Africa: National Treasury, 2017a: 43.

people need to try harder[60]. This is bad planning from a number of angles. Firstly, real quality trends are unlikely to fluctuate in this manner, from one year to the next. Secondly, planners did not take into account the fact that nationally the pass rate also dropped in the last year. What should have appeared is some kind of crude 'difference in difference' analysis.

It is useful to distinguish between production and consumption problems when it comes to statistics. The South African plans display production problems, for instance in the form of inconsistent pupil-teacher ratios. There are many ways of calculating such ratios, and producers of these statistics need to be explicit about what method they used. To think that a decline in an indicator value, such as that of the pass rate, automatically reflects a deterioration, points to a problem where statistics are being 'consumed'. Often those who produce statistics and those who consume them are different people, meaning the distinction can be useful when designing training interventions.

What the literature tends to warn against is placing too much trust in the levels and trends displayed by typical service delivery indicators. Especially in developing countries (but not only there), measurement systems suffer from a host of production problems: there are often incentives to fake better numbers; where statistics are sample-based, samples may not be comparable; and data analysts often lack all the skills, or equipment, needed to analyse microdata properly. More comparisons across different data sources, so-called triangulation, acceptance that monitoring systems are not perfect, and careful judgement, are all needed[61].

One way of deepening the understanding of actual trends is to bring in more long-range trends. South Africa's reports suggest that planning occurs without considering the implications of such trends. KwaZulu-Natal's report had 85 indicators with on average two years of historical data, against New South Wales's 50 indicators with an average of six past years. How trends are displayed can also strengthen a report. New South Wales's report contained 34 graphs, against none in KwaZulu-Natal's[62].

Indicators set in stone. In recognition of the fact that monitoring systems in developing countries should be undergoing a continual process of maturation and development, as lessons are learnt, indicator definitions should not be set in stone. This is not easy for national agencies such as the Auditor-General, whose staff have been drilled into working with rigid financial accounting standards, to accept. Of course, when it comes to non-financial statistics, which measure hundreds of things, not just money, and draw from samples, student testing, opinion surveys, ratings by inspectorates, and so on, and enjoy very little in the way of good overriding accounting principles, things are very different. Documentation on how best to gauge progress on the non-financial side is needed. However, technical indicator descriptions as required by South Africa's National Treasury, and promoted by the UNESCO Institute for Statistics (UIS)[63], are not really the answer. Such descriptions put forward a highly theoretical method, with much of the focus falling on a formula. A particularly glaring example are the UIS indicators relying on a 'reconstructed cohort method'. The theory of the method is described in a key UIS indicator guide directed at education planners, with no evidence that the method works with the actual data planners in typical developing countries have access to. In fact, the method often provides wildly inaccurate statistics in such contexts, because grade repetition data are weak. Yet many developing country planners have attempted to follow the method, and some would have put forward the resultant statistics as fact, when they should not. In South Africa, such a process contributed to confusion and a public enquiry into supposedly high levels of dropping out[64]. In short, existing guides tend not to take into

[60] South Africa: National Treasury, 2017a: 38.
[61] Jackson, 2011: 23-4; Dunleavy, 2015: 29.
[62] South Africa: National Treasury, 2017a: 43.
[63] UIS, 2009.
[64] South Africa: Department of Education, 2008.

account how people have succeeded, or not succeeded, in measuring a particular concept in the past, what the typical data quality problems are, data collection costs, and so on. A better approach would be to require organisations such a Ministry of Education to state, using policy and available or realistically planned monitoring systems as a point of departure, what trends should and can be measured. An ideal measurement approach can be proposed, with some sense of the time and money it would take to reach it. A government-wide set of principles governing the detection of trends would then be used determine whether a trend was reliable or not. The key technical documentation should thus emerge *with* the reported trend, not *before* the analysis began. That way one would be able to take into account difficulties relating to, for instance, data quality. To illustrate, in gauging the trend with regard to class size, one might measure change between 2005 and 2010 using one method, and change between 2010 and 2015 using a different method, if monitoring systems had changed over time, which is likely in a developing country. The 2005 and 2015 values may not be strictly comparable, but at least one would have a good sense of what the change was over the longer term.

Indicator overload. Indicator overload is something government officials often complain about. National Treasury's review seems to suggest there is a trade-off between the quantity and quality of indicator information. One thing that seems to encourage a proliferation of indicators is a planning approach based strongly on tables. Such a format lends itself to the demand that each row, each activity, must come with its own indicator. South Africa's plans, but also those of many other developing countries, rely excessively on such tabular planning. Forty-four pages in the Western Cape report is taken up by tabular descriptions of what was done. In contrast, the New Zealand and New South Wales reports provide nearly all such descriptions in the form of regular paragraphs. Tables tend to reinforce silo effects and limit discussions about the inter-connectedness of different activities. Moreover, beyond the indicator proliferation problem, they also result in other forms of 'padding', as well as repetition, as planners think of things to put in each table cell, the rule often being that no cell is to be left blank.

While it could be argued that overall there should be *fewer* indicators, it could also be argued that there could be *more* indicators of a certain kind. The South African plans examined in the Treasury review tend to use many indicators reflecting internal processes, but few reflecting processes the public are likely to find interesting. Thus, training workshops for teachers receive considerable attention, but not the provision of information on schools to parents and communities. Tellingly, the 'customer satisfaction survey' of the Western Cape monitors not the satisfaction of parents with the schooling system, but of school principals with the administration, though principals and administrators are both employees of the education department[65].

The targets problem. Setting targets in a schooling system may seem like an obvious thing to do to if progress is to be encouraged. This assumption seems to underpin many monitoring systems within countries, but also the education targets of the UN's Sustainable Development Goals. It therefore comes as a surprise to many that targets have been found to be counter-productive. In the United States, the federal government moved away from a strong emphasis on achieving specific targets from around 2009. Targets had been found to undermine management in a number of ways. Non-attainment of a target could easily make teams of people doing as best as they could under the circumstances, appear to be failing. Conversely, target attainment could make people appear to be succeeding, even if it was clear they could have done a lot more, but did not because the target had been reached. In short, even in a technocratic environment targets are difficult, because it is so difficult to predict what should be achieved one, two or three years into the future. But in developing countries, planning environments tend not to be technocratic, with politicians frequently pushing through highly

[65] South Africa: National Treasury, 2017b: 14.

aspirational targets into the plans of bureaucracies. In the case of the United States, the emphasis shifted away from reaching hard and very specific targets, and towards reporting that some improvement had occurred, and producing evidence that one had done one's best. To put this in a more developing country context, a manager would be required to demonstrate, firstly, that a school building programme had become more efficient than in the past and, secondly, that in the current circumstances, it would have been difficult to do any better. There would thus be no target of X schools had to be built. The binary and overly simplistic notion of having reached or missed a specific target is thus avoided[66].

The rise and fall of 'Deliverology', an attempt to create a more rigorous and empirically-focussed planning approach, offers a sobering warning of how the targets problem can undermine governance. Deliverology originated in the United Kingdom in around 2000. Apart from intensifying the focus on achieving targets in the public service, Deliverology meant establishing a 'delivery unit' of highly skilled people, with strong links to the head of state, to mediate matters such as target-setting[67]. The approach achieved some success in the United Kingdom. A focus on simple indicators which matter for the public, such as waiting times at public hospitals, seems to have had a positive impact. In schooling, success appears to have been limited. Much of the focus here fell on indicators tracking the use of standard lesson plans, something which teachers were opposed to on professional grounds. There is no clear evidence that this emphasis contributed to better learning outcomes. Around 2010, the United Kingdom turned to *de*-emphasising targets, along the lines of what was occurring in the United States[68].

However, at about this time Deliverology, often under a different name, but with the original strong focus on targets, was taken up by several developing countries, including, to a limited extent, South Africa. How this occurred in Pakistan was widely reported on and controversial. In that country, targets at a local level in the schooling system received much attention, with a particular focus on attendance and test scores. Several analysts have concluded that the approach incentivised a faking of the numbers. Moreover, the fairly typical problem of moving very quickly from almost no monitoring to excessively intense monitoring made poor quality data almost inevitable. To illustrate the excesses, testing of children by external monitors occurred monthly, the expectation being that a certain improvement should be seen every month at the level of the system[69]. If methods for detecting improvements in student performance at a systemic level from one year to the next are under-developed (as discussed previously), the problem is far more serious for anyone attempting to track month-on-month progress for a whole system. Jishnu Das, senior economist at the World Bank, has essentially argued that the faking of the evidence in Pakistan extended to the narrative around Deliverology itself, with enrolment trends being used incorrectly to attribute a causal link between the approach and service delivery improvements[70].

The review of the National Treasury manuals concludes that they need updating, based on lessons learnt. It is of course important to view optimal planning methods as an evolving body of knowledge. There seems to be a real risk that methods which do not work well become entrenched, and difficult to change. The history of Deliverology moreover suggests that developing countries easily adopt approaches from developed countries even after these developed countries have moved on to something better. Implicit in Pritchett's arguments is that different levels of country development require rather different approaches. Emulating developed countries is a part of the problem for developed countries.

[66] Metzenbaum, 2015: 47.
[67] Barber *et al*, 2011.
[68] Richards and Chegus, 2018.
[69] Naviwala, 2016.
[70] Das, 2013.

In South Africa, but also other countries, a critical factor is the Auditor-General, which has increasingly become a judge of non-financial performance, using non-financial indicators. The problem with this is that Auditor-General (AG) staff typically have a financial accounting background and are not familiar with, for instance, the opportunities and risks associated with measuring learning outcomes. Even if the AG's office hires, say, education experts, the financial auditing environment, with its low tolerance of uncertainty and margins of error, is not conducive to the monitoring of outcomes. Obviously, non-financial performance information should be quality-controlled. The question is how. One solution can be found in New South Wales. Unlike in South Africa, or New Zealand, in New South Wales the AG does not express an opinion on non-performance data as part of its auditing of the annual report of the education department. The AG does occasionally monitor performance, but through a completely separate process, and through separate reports[71]. This helps to remove from the annual auditing process overly simplistic discussions around whether certain performance information are 'clean' or not. Such discussions are essential for financial auditing, but often unsuitable on the non-financial side. In South Africa, there are under-explored options for the quality control of non-financial data. The Statistics Act allows Statistics South Africa to pronounce on the reliability of official government statistics. Umalusi, an independent body which has been highly successful in quality assuring the Grade 12 examinations, could assume further functions relating to the monitoring of at least educational outcomes information at other levels.

6 Conclusion

This paper has used an interesting recent book, *The politics and governance of basic education: A tale of two South African provinces*, as a point of departure in analysing what constitutes a capable state equipped to promote progress in educational quality. The book is somewhat pessimistic about the potential for the state to make a difference. This paper has argued against pessimism, for a more rigorous analysis of the incapable state, and for a better understanding of technical solutions in areas such as the monitoring of learning outcomes. The book is right in arguing that the political context can impose serious constraints on development. However, experiences in other countries demonstrate that building specific capacities among bureaucrats and those who work with bureaucrats can serve as a bulwark against ills such as corruption and populist policymaking.

Like the book, this paper finds Pritchett's warnings against mimicry and faking in the area of statecraft highly relevant. But the problem of mimicry extends beyond developing country states. The manuals and literature intended to guide the building of better institutions do themselves at times not live up to the promise of their covers and titles. Planners should use existing guides critically, and should advocate for better and updated ones in areas such as the monitoring of learning outcomes, sectoral performance indicators, and teacher supply and demand.

References

Unless otherwise indicated, all sources, other than articles in journals requiring a subscription, were obtained off the web.

Barber, M., Kihn, P. & Moffit, A. (2011). *Deliverology: From idea to implementation.* New York: McKinsey.

Brazil: Ministério de Educação (2011). *PDE|SAEB: Plano do Desenvolvimento da Educação.* Brasilia.

[71] See for instance New South Wales: Audit Office, 2019.

Brazil: Ministério de Educação (2015a). *Resumo técnico: Resultados do Índice de Desinvolvimento da Educação Básica 2005-2015.* Brasilia.

Brazil: Ministério de Educação (2015b). *Avaliação Nacional da Alfabetização: Relatório 2013-2014.* Brasilia.

Bruns, B., Evans, D. & Luque, J. (2012). *Achieving world class education in Brazil: The next agenda*. Washington: World Bank.

Das, J. (2013). *A data guide to Sir Michael Barber's "The good news from Pakistan".* Washington: World Bank.

De Andrade, D.F., Tavares, H.R. & Valle, R. (2000). *Teoria de resposta ao item: Conceitos e aplicações*. Fortaleza: Universidade Federal do Ceará.

Dee, T.S. & Wyckoff, J. (2015). Incentives, selection and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis of Management,* 34(2): 267-297.

Department of Basic Education (2012). *Report on the Annual National Assessment 2012: Grades 1 to 6 & 9.* Pretoria.

Department of Basic Education (2013). *Report on the Annual National Assessment of 2013: Grades 1 to 6 & 9.* Pretoria.

Department of Basic Education (2016). *The development of a National Integrated Assessment Framework*. Pretoria.

Department of Basic Education (2019). *National Senior Certificate 2018 examination report.* Pretoria.

Dunleavy, P. (2015). Public sector productivity: Puzzles, conundrums, dilemma's and their solutions. In Wanna, J., Lee, H-A. & Yates, S. (eds.), *Managing under austerity, delivering under pressure: Performance and productivity in public service.* Acton: ANU Press.

Dunleavy, P. & Margetts, H., Bastow, S & Tinkler, J. (2006). New public management is dead - long live digital-era governance. *Journal of Public Administration Research and Theory,* 16(3): 467-494.

Gasperini, L. (1999). *The Cuban education system: Lessons and dilemmas.* Washington: World Bank.

Greaney, V. & Kellaghan, T. (2008). *Assessing national achievement levels in education.* Washington: World Bank.

Greaves, P. & Achterstraat, P. (2015). Reviewing performance to improve delivery: Key insights from two auditors-general. In Wanna, J., Lee, H-A. & Yates, S. (eds.), *Managing under austerity, delivering under pressure: Performance and productivity in public service.* Acton: ANU Press.

Gustafsson, M. (2016). *Understanding trends in high-level achievement in Grade 12 mathematics and physical science.* Pretoria: Department of Basic Education.

Gustafsson, M. & Nuga Deliwe, C. (2017). *Rotten apples or just apples and pears? Understanding patterns consistent with cheating in international test data.* Stellenbosch: Stellenbosch University.

Gustafsson, M. & Patel, F. (2006). Undoing the apartheid legacy: Pro-poor spending shifts in the South African public school system. *Perspectives in Education,* 24(2): 65-77.

Hanushek, E.A. & Wößmann, L. (2007). *Education quality and economic growth.* Washington: World Bank.

International Budget Partnership (2018). *Open Budget Survey 2017.* Washington. Available from: <https://www.internationalbudget.org> [Accessed November 2018].

Izard, J. (2005). *Trial testing and item analysis in test construction*. Paris: IIEP.

Jackson, P.M. (2011). Governance by numbers: What have we learned over the past 30 years? *Public Money & Management,* 31(1): 13-26.

James, T.S. (2014). Neo-statecraft theory, historical institutionalism and institutional change. *Government and Opposition,* 51(1): 84-110.

Jerrim, J. (2013). The reliability of trends over time in international education test scores: Is the performance of England's secondary school pupils really in relative decline? *Journal of Social Policy,* 42(2): 259-279.

Kenya National Examination Council (2016). *Monitoring learner achievement at Class 3 in literacy and numeracy in Kenya.* Nairobi. [Not publicly available on the web.]

Levy, B., Cameron, R., Hoadley, U. & Naidoo, V., eds. (2018). *The politics and governance of basic education: A tale of two South African provinces*. Oxford: OUP.

Metzenbaum, S.H. (2015). Measuring and improving government performance: Learning from recent US experience. In Wanna, J., Lee, H-A. & Yates, S. (eds.), *Managing under austerity, delivering under pressure: Performance and productivity in public service.* Acton: ANU Press.

Naviwala, N. (2016). *Pakistan's education crisis: The real story.* Washington: Wilson Center.

New South Wales: Audit Office (2019). *Ensuring teaching quality in NSW public schools.* Sydney.

OECD (2013). *Synergies for better learning: An international perspective on evaluation and assessment*. Paris.

OECD (2015). *Education policy outlook 2015: Making reforms happen*. Paris.

Pritchett, L., Woolcock, M. & Andrews, M. (2012). Looking like a state: Techniques of persistent failure in state capability for implementation. *Journal of Development Studies,* 49(1): 1-18.

Reddy, V., Visser, M., Winnaar, L. & Arends, F. (2016). *TIMSS 2015: Highlights of mathematics and science achievement of Grade 9 South African learners*. Pretoria: HSRC.

Richards, G. & Chegus, M. (2018). *Does 'Deliverology' deliver?* Ottawa: Institute on Governance.

South Africa: Department of Education (2008). *Ministerial Committee on learner retention in the South African schooling system*. Pretoria.

South Africa: National Planning Commission (2012). *National development plan 2030: Our future - make it work*. Pretoria.

South Africa: National Treasury (2017a). *Towards better generation and use of data within the basic education sector*. Pretoria.

South Africa: National Treasury (2017b). *Generation and use of data in the Western Cape in the delivery of basic education*. Pretoria.

UNESCO (2008). *Primer reporte: SERCE: Los aprendizajes de los estudiantes de América Latina y el Caribe*. Santiago: OREALC/UNESCO. Available from: <http://unesdoc.unesco.org/images/0016/001606/160660s.pdf> [Accessed March 2010].

UNESCO (2014). *Education for All global monitoring report 2013/4: Teaching and learning: Achieving quality education for all*. Paris.

United States: National Center for Education Statistics (2005). *NAEP 2004 trends in academic progress: Three decades of student performance in reading and mathematics*. Washington.

UIS (2009). *Education indicators: Technical guidelines*. Montreal.

UIS (2017a). *Principles of Good Practice in Learning Assessment*. Montreal.

UIS (2017b). *Quick guide no. 3: Implementing a national learning assessment*. Montreal.

Van der Berg, S. & Gustafsson, M. (2017). *Quality of basic education: A report to Working Group 1 of the High Level Panel on the Assessment of Key Legislation.* Cape Town: Parliament.

Williams, P. (1979). *Planning teacher demand and supply*. Paris: IIEP.

Wößmann, L. (2005). The effect heterogeneity of central examinations: Evidence from TIMSS, TIMSS-Repeat and PISA. *Education Economics,* 13(2): 143-169.

World Bank (2003). *World Development Report 2004: Making services work for poor people*. Washington.