Oxford Review of Education

Special Issue 2018, 44, 5 (October)

The consequences of metrics in education: gaming, malpractice and cheating

Editorial Metrics in education - control and corruption

Neoliberalism is a form of governance that has free market economics at its heart: competition, deregulation and macro-level quantification are its founding blocks. The term was coined in 1938 by Ludwig von Mises and Frederich Hayek and was a reaction to forms of governance involving collectivism. The economist Milton Friedman is an iconic figure in the neoliberal field, as are some political leaders - Mikhail Gorbachev, Margaret Thatcher, Ronald Reagan, Tony Blair, David Cameron and Donald Trump. Adherents to neoliberalism subscribe to the belief that economic and individual well-being is best served by liberating individual entrepreneurial skills within a framework of strong property rights, free markets and free trade. In societies that have come to be governed by neoliberal regimes, few people even recognise the term. This is the equivalent of living under communism or democracy without having heard such terms. In October 2017, when giving a keynote address at an international conference of examination boards, Baird asked the audience whether they had come across the term neoliberalism. Only one third of the participants put up their hands. Yet neoliberalism affects our lived realities as individuals, including through the metrics for which we are objects. Through those metrics we are coerced and controlled and at times rational actors seek to corrupt the system for a range of reasons.

Neoliberalism is important for the conceptualisation of what is happening to education because neoliberal views have become such an accepted part of the discourse that they are presented as though there is no alternative. This is reminiscent of George Orwell's dystopian novel, *Nineteen Eighty-Four*, where the character Winston Smith explains that the regime's aim is to narrow the range of thought such that "there will be no words in which to express" certain thoughts. In this context, this Special Issue looks at a range of issues associated with the implications of neoliberal metric regimes in education. The Editorial provides a context for the individual cases in the articles. We discuss the foundational ideas behind metrics such as neoliberalism, performativity, the audit society, managerialism and workfare. Some of the metrics in education are outlined, including known effects. We then turn to the focus of the Special Issue: gaming, malpractice and cheating consequences of metrics. In the Florez et al. paper (this issue) and at the end of the Editorial, we ask what the future looks like for metrics systems.

This Special Issue looks at a range of ways in which metrics have been introduced in education and how they have provoked unintended effects. The term 'metrics' is used for a quantitative measure of educational practice. Quantification brings the power of numbers, in all its forms: numbers can be more persuasive to certain audiences; through aggregation metrics are used to

represent large-scale systems in an apparently coherent manner; calculations can be carried out upon them; and they can be compared. Assigning numbers to qualitative features of education gives these benefits, but the richness of life is lost in the process. The context in which the behaviour arose, the people involved and the institutional arrangements are hard to capture in all of their colour and diversity through numbers, even with a range of variables. There is a well-developed literature on these issues in educational research methods, which we cannot rehearse here due to the constraints of space. However, it is noteworthy that the extensive work on qualitative versus quantitative data (and even mixed methods research) is relevant to these debates but rarely referred to in policy circles. School effectiveness researchers have developed mixed methods techniques to a large extent, as they have faced these problems directly. However, it would seem that there are corners in which numbers are still seen as a dominant form of knowledge.

Governing volume *and* complexity in any system is devastatingly difficult and can lead to system failure. Therefore simplifying the world to deal with large-scale systems is a logical response to the situation that governments and their agents find themselves in. Sticking with the public sector for now, the need for scrutiny of how public money is spent is often cited as a rationale for the introduction of metrics, and likewise efficiency and quality are desirable. After all, without public scrutiny, organisations can run themselves for the benefit of their own staff. Measurement in itself does not produce governance, but is used as part of the governance process which involves the characterisation of what is noteworthy, the quantification of these characteristics, setting of targets in relation to those quantities, monitoring progress against them and putting in place incentives and penalties to reinforce the targets. Setting out what the norms of performance should be for all provides a standardised metric that can be used across the system, rather than having to grapple with the idiosyncrasies that would be encountered if each case were taken on its own merits and context. In the face of complex systems, what else is a government to do?

Norms are set out for us in many aspects of our lives and are so woven into the fabric of our social existence that it can be hard for us to distinguish these norms from natural order. As Foucault (1975, p.308) put it,

> The judges of normality are everywhere. We are in the society of the teacher-judge, the doctor-judge, the 'social-worker'-judge; it is on them that the universal reign of the normative is based; and each individual, wherever he may find himself, subjects to it his body, his gestures, his behavior, his aptitudes, his achievement.

Metrics articulate neatly with this social order and help to shape it. Whilst there is much written about the imposition of metrics and the effect on the professional (Ball, 2003; Perryman et al., 2011; Muller, 2018), the insidious thing about them is that they do not just control us externally. We do not only accept these norms as individuals, but we also internalise them as part of our identities and we can define ourselves in relation to them. Take the example of the seven-year-old girl who said "I'll be a nothing" when preparing to take a Key Stage test

in England (Reay & Wiliam,1999). Or, more personally, how we see our own intelligence following all of the test results that we have had in examinations. Christian Ydesen and Frederik Forrai Ørskov (this issue) consider the IQ test itself, and in particular its implementation in England and Denmark in the mid-20th century. In their paper they show how IQ test scores were used to condemn certain groups of children as 'ineducable'- with consequences for their life chances. They use specific cases from the two countries - both shocking and fascinating for educational professionals - to critique the concept of objectivity in relation to IQ scores and to demonstrate the fundamental game-ability of metric assessments of ability. Although dealing with historical contexts, they draw parallels with contemporary neoliberal concepts such as 'workfare', in which jobseekers are expected to take on low-skilled and low-status work as volunteers to qualify for unemployment benefits.

Roger Goodman and Chinami Oka (this issue) provide the first comprehensive account in English of a particular metric in secondary education: the Japanese *hensachi*, which has become something of a cultural phenomenon. Rankings from single high-stakes examination taken at the end of secondary school, the *hensachi* have the power to determine the life courses of individuals and therefore their estimates of their own worth: when the stakes are this high, the incentives to gain whatever advantage you can become obvious, and opportunities for outside agencies to sell products claiming to give advantage multiply. Goodman and Oka explore interlinked issues of reputation, education and systemic as well as individual responses to the *hensachi* and 'examination hell'. The consequence is a fundamental questioning of the concept of meritocracy when it comes to educational achievement.

Notwithstanding the above, metrics do not always adhere to our identities; they can have very different effects. Young people feel alienated with metric systems they do not recognise as valuable, or offer no opportunities for them. As professionals, rejection of the overarching system of governance involving metrics is a tricky act to pull off if you want to keep your job. Therefore, we can end up in the very uncomfortable position of cognitive dissonance in which we hold two contradictory views (Festinger, 1957). In this case, we have to behave as though we believe in the metrics even if we think they are misplaced, irrelevant or even have deleterious effects upon education. Lived realities within such systems are alienating as our own behaviours and interactions with each other have less meaning. As Peck (2002, p.350) put it, we can find ourselves on "treadmills of someone else's making, coerced and constrained by a combination of competitive relations and superordinate rule systems". Teachers have been known to become disenfranchised and even cynical under school accountability systems, for example.

The philosopher Baroness Onora O'Neill (2002) argued in her Reith Lectures that accountability systems produce *less* trust in society because of the externalised targets. With imposed evaluation systems, professionals have active disincentives from thinking for themselves about how to improve things. In any case they are typically kept very busy with the imposed targets, so time for professional engagement is scarce. Further, if our social milieu is permeated

with the idea that we have to monitor one another and measure one anothers' performances in relation to targets, it is self-evidently one that is not built upon trust and which reinforces mistrust. Many have argued that the accountability system in education is an effective panopticon - because you can never know when inspection will come (from an external agency or from internal monitoring) then teachers are constantly in a state of 'inspection readiness', with books demonstrating written marking and lessons demonstrating the approved methods and structure (e.g. Bushnell, 2003; Perryman, 2006). These are the consequences of metricisation: both intended and unintended.

The concept of trust in public life and what happens when it proves to be unwarranted was the central theme of Germany's doctoral plagiarism crisis a few years ago. Hubert Ertl (this issue) analyses two high-profile cases of alleged plagiarism in doctoral theses which led to the career downfalls of two government ministers. His paper uses these cases to debate academic malpractice in the German context, and to examine the structures that potentially enable or explain that malpractice. He describes the intense public interest these cases generated, and the resulting emergent phenomenon of the plagiarism hunter. This paper also shows the potential thinness of the line between individual accountability and the systemic issues that promote, allow or incentivise what can be termed gaming, malpractice or simply 'cheating'.

Ideas of metrics came to the fore in the 80s and 90s, with the rise in new public management (Hood, 1991), which was a drive to make public management more business-like. On the back of this came what Power (1997) termed 'the audit society', in which we are all going through the mechanisms of checking on each other. As testament to this, witness how university academics in England currently log in a ritualistic, bureaucratic fashion interactions with students who have a visa to study in the country. Flagging lack of engagement with their studies would not be enough, as the point is to make sure that the paperwork is correct for the auditors. Equally, there have been examples of Assessment for Learning activities designed not only for pupils, but for the inspector to see (Usher, 2013), or the now infamous 'verbal feedback given' stamp to be used on pupils' books. Working within audit systems, which go well beyond economic activity, people are coerced to meet the targets, even if they are considered perverse. Knowledge of performance is signified by the metrics. Questions about quality and truth are replaced by discussions of the metrics.

Understanding what the metrics tell us is not a property of the data itself; it must be interpreted. In 2008, the global economic crisis came as a shock because the banking sector did not understand the derivation of the macroeconomic figures that they were dealing with. Mortgage lending to US citizens who could not afford the payments had been aggregated and manipulated quantitatively in so many ways that banking executives could not see what they were buying and selling. Knowing what the numbers signify is just as crucial to the endeavour in education as in economics. In the banking sector, the profit motive, competitive structures, short-termism, weak regulation, conflicts of interest and globalisation provided a context for the 2008 global economic crisis. Here we see that important facts of the situation got lost in macroeconomics that are difficult for

state actors such as politicians to control. Education is to a far larger extent within political and state control.

Lyotard (1984) introduced the notion of 'performativity' to express how we behave in such systems, by trying to meet the criteria without believing in their utility or truth. Professionals are going through the motions, ticking the boxes as best they can, to avoid the penalties that come with non-compliance. Performativity has an unreal quality for people because the signifiers of quality and efficiency have come to have a life of their own. The signifiers have become more important than the signified (Hanson, 2000). Examination data, for example, have a social currency going well beyond the content of the tests themselves. Indeed, educational assessment has been a very useful tool for the metrics agenda. Test scores are supposed to signify an individual student's knowledge, teacher quality, school effectiveness and even the performance of education systems at the level of nation state. Challenges to the utility of the data for all of these purposes are manifold (e.g. Grek, 2012; Leckie & Goldstein, 2009; Nuttall, 1987; Popham, 2007). These debates are important counters to performativity, as it is not simply that our *behaviour* is controlled by accountability mechanisms, but that the very definition of our work is altered by them. In this case, metrics control what counts as learning and teaching (Ball, 2003).Channelled through management structures, people experience a 'managerialist' culture in which metric targets are prioritised over professional values.

Metrics in education

Measurements in education inhabit different levels of the education system. Assessment has a long history of use as a metric for individuals, with Imperial China first creating examinations in 124 BCE for selection into the civil service (Roberts, 2006, p.31). School inspections also have a long history, going back over 100 years in England. Pre-school inspections came later. Higher education has increasingly been audited in many countries in a range of ways. Additionally, supranational metrics have come into play this century, with international tests such as the Organisation for Economic Cooperation and Development (OECD)'s PISA or the International Association for the Evaluation of Educational Achievement (IEA)'s TIMSS studies being used to rank-order countries' secondary school systems.

Following the discussion above regarding the signifier and the signified, it is crucial that there is an understanding of what the metrics mean, but this is often absent. Politicians and other commentators regularly point to the PISA results to show that change is urgently needed in their education systems (Baird et al., 2016), but it is highly questionable whether they understand international tests and what they measure. Indeed, given the complexities of the international testing, it is impossible to interpret the results in any case. They suffer from a range of issues, from sample differences across countries, translation effects, statistical modelling controversies and curriculum alignment discrepancies, to name a few (Hopfenbeck et al., 2017). However, a surface reading of the results tells us that Shanghai's education system produces 15-year-olds with much higher levels of reading Chinese than other countries' 15-year-olds can read their

own languages. Since PISA results are highly correlated with intelligence (*r*=.88; Baumert et al., 2009), this is an extraordinary finding and one that has been heavily questioned, raising the question of whether PISA results can be gamed.

Bernadetta Brzyska's paper (this issue) moves the question of gaming onto the international stage, presenting data from five cycles of PISA to consider the differing exclusion rates of students with special educational needs from the testing panel in the participating countries. She considers variations in the definition of special educational needs across countries and how those variations and consequent removal of students from the sample have the potential to skew - intentionally or otherwise - the overall outcome of an international survey which has become so important on the world stage that it has its own systemic syndrome named after it: PISA-shock (Gruber, 2006). As with other papers in this issue she demonstrates that the weight we place on a score should only be as great as our confidence in that score - and that we may be misplacing our confidence.

No less in higher education have the effects of metrics made themselves felt. Australia and the UK have both undergone several cycles of assessment of research excellence in universities: the 'Excellence in Research for Australia' or ERA programme and the Research Assessment Exercise (RAE) which became the Research Excellence Framework (REF) in the UK. The last REF in 2014 had the effect of turning the academic jobs market into something which rather more resembled the transfer window for footballing superstars, while those who could not afford to compete advertised one-day-a-week posts for full professors who could bring with them a raft of 4* publications, and yet others had retirees hanging on just long enough on similar contracts. The perverse incentives for this have been removed in the new rules for the REF in 2021, but instead the new rule of non-portability (that publications can only be submitted at the institution in which they were generated) has left early career researchers wondering how best they can play their first few publications into employment as well as 'REF-ability'. Nevertheless, the impact of research assessment has been to introduce new routines into academic life across the board: mock REFs, REF strategy groups, external reviews of published research papers and others are all becoming embedded in university life in order to find the edge in the competition for research reputation and funding (Oancea, 2018).

The incentives in both the REF and ERA are largely institutional, but in other places the researcher is rather more on the front line. Quan, Chen and Shu (2017) explored the financial rewards offered by some Chinese universities for international publications; one Chinese professor published 279 times in the same journal between 2004 and 2009. 'Payment by results' means finding the quickest possible route to a publication, rather than investing time and energy in what has the potential to be a slow-burn groundbreaking work. Peter Higgs, who won the Nobel Prize in 2013 for his work in the 1960s on the Higgs boson, a theoretical particle which was finally demonstrated to be correct in 2012 by the large hadron collider at CERN, has queried whether he would ever have had an academic career in the era of the REF. Research metrics more generally have been challenged, particularly those that rely on citation: simple counts do not

reveal whether citation is framed positively or critically, some journals artificially inflate their impact factor by requiring authors to cite previous articles, and some authors form friendly 'citation circles' to inflate their own and one another's scores (Muller, 2018).

Various world university rankings, such as those produced by the Times Higher Education Supplement, or the QS World University Rankings, draw on a variety of metrics such as research volume and income, citations, industry income and reputational measures from other universities or from domestic employers. Muller (2018) described an oddly placed advert for a small university in the *Chronicle of Higher Education*; he eventually worked out it was an attempt to influence the influencers of the university reputation metric.

Neoliberalism also requires metrics to promote the marketisation of higher education, demonstrating particular concerns which may be masked under the guise of concern for students. The 'scorecard' system for ranking universities in the United States relies heavily on the earning expectations of graduates to make its judgements. In the UK on 26th June 2018 Sam Gyimah, the Minister for Universities, Science, Research and Innovation tweeted a call for tech companies to use existing data on graduate earnings for different courses and at different institutions to create a 'MoneySupermarket' (SamGyimah, 2018) of universities, that would enable prospective students to make informed choices - that is, choices informed by financial considerations. Meanwhile the new Teaching Excellence Framework, touted as an objective metric for English universities, has been analysed to demonstrate that the narrative submission has the potential to substantially change outcomes away from the 'objective' metric score (Gillard, 2018). The same themes emerge throughout the discussion of metrics in any context: the fallibility of their objectiveness, the ability to manipulate outcomes in various ways, and the unexpected consequences of measurement.

Even if metrics are a device of neoliberal regimes, they are certainly not the only kind of societies happy to have them at their disposal. Medieval monarchs used metrics to conduct censuses of the population so that they could set taxes, the fascists in Nazi Germany collated a wide variety of statistics about the population, but it is the communist regimes that have outclassed the other forms of governance in their use of metrics and in their inventiveness regarding the forms they might take (Bevan & Hood, 2006). In fact, the history of communist regime centrally set targets and their effects provides a rich set of case studies for the effects of metrics. When overly ambitious targets were set centrally by politicians, who had an eye on international public relations, factories and farms recognised that they could not meet them. Since the consequences of failure could be death, Siberian prison camps or if you were lucky, simply losing your job, the incentives were high to appear to meet the targets. Cheating was the only rational option for actors within such a system, though many people starved as a consequence of the food production target metric systems. Of course not all metric systems are as draconian. Nonetheless, gaming, cheating and malpractice are incentivised in metric systems and regulatory bodies are often set up to monitor the targets and even to some extent to audit the ways in which people

reach them (Bevan & Hood, 2006). The extent to which the regulators are independent from government varies.

Gaming, malpractice and cheating

'Gaming' is working within the rules to your own advantage. It includes behaviours which fall within the letter but not the spirit of the law. Many of the unintended consequences of metrics fall within the category of gaming, as they encourage people to find ways to play the game in order to maximise their own benefit (which educationally may mean that of their students).

Malpractice meanwhile has a professional bent. Legally it is defined as 'the negligent rendition of professional services' (Mallen, 1979, p. 205), specifically without intent, and usually with a harm created to someone else by that negligence. Within contexts other than that of law and medicine it might work better to think of professional 'obligations', that is, adherence to the code of your profession. Does this mean that teaching to the test and therefore neglecting a key area of the curriculum (perhaps because it came up in the exam last year) or a key purpose of education (because it cannot be tested) is malpractice? Potentially. Recognising malpractice is made more difficult by the adoption of the somewhat euphemistic term 'academic malpractice' which can cover both being negligent in the application of referencing and attribution norms, but also deliberate and sustained plagiarism.

This plagiarism, we would argue, falls more clearly within the definition of 'cheating' which is the most problematic of these terms, given its pejorative and accusatory use. Cheating is not only against the spirit but also the letter of the law. It is deliberate defiance of the rules to gain advantage: it is intentional, thus differentiating it from malpractice. Cheating is buying knowledge of the questions that will be on the exam ahead of time, or lying on your expenses claim. The behaviours described within the papers in this Special Issue fall into all three of these categories, and often into more than one in relation to any particular data set. While the strict definitions might propose a hierarchy of increasing reprehensibility, behaviours categorised in these terms are considered to be more or less problematic on an individual basis by participants in the studies described.

An extreme case of outright cheating occurred in Atlanta, US, in 2014. Nine teachers were convicted of racketeering. The Superintendent, Beverley Hall, had won awards and been given bonuses for improving school results through her data-driven approach. She died of cancer before her trial came to court, but nine others who worked with her were given prison sentences. On an institutional level, the English school's inspectorate, Ofsted, has recently announced that it has identified 300 schools with higher than expected levels of 'off-rolling' - that is, students who disappear from the school register between years 10 and 11, when they would sit their national examinations (Bradbury, 2018). There are parallels with case selection bias in surgery incentivised by pay for performance: surgeons will not operate on the difficult cases with the chances of higher mortality rates, which are then filtered off to less careful locations. Some 19,000 students in England are no longer on the register in year 11 for the school they

were registered at for year 10; half of them never reappear on any school's register. One assumption is that they are encouraged to leave because they are the students who would bring down the school's overall results. Is this cheating, malpractice or simply gaming?

'Off-rolling' is one of the activities considered by Michelle Meadows and Beth Black (this issue). While they note that malpractice is rare within the English examination system, given the scale of activity (six million qualifications awarded each year), they relate data from an anonymous online survey of the teaching community asking about activities that were undertaken to maximise qualification results, and the acceptability of those activities. Removing pupils from the school roll was reported as being experienced by a quarter of their respondents and was ranked as one of the more unacceptable activities. Meadows and Black also provide some extraordinary first-hand accounts by teachers of activities they or their schools took part in, and the ethical dilemmas they faced as a consequence. By drawing on what teachers had experienced at first-hand they attempt to differentiate evidence of ethically questionable behaviours in response to accountability measures from the inevitable rumour that surrounds such behaviours.

This difference appears in a different guise in the lead article in the issue. Jenni Ingram, Victoria Elliott, Caroline Morin, Ashmita Randhawa and Carol Brown's paper speaks to the question of unintended consequences in secondary education. Drawing on data from an interview study that explored the perspectives of teachers before and after recent examination reforms in England, focusing specifically on mathematics and English, they consider the ways in which schools responded to possibilities in examination entry to maximise achievement both for individuals and overall. Their respondents move from metaphors of 'playing the system' to 'levelling the playing field' when speaking of the reforms' effect on schools. The participants speak differently about decisions concerning examination entry, re-sits, early entry or tiering, depending on whether those decisions happen in their own schools or in those of others, perhaps reflecting again the idea of cognitive dissonance in relation to responses to education metrics. Both Meadows and Black's and Ingram et al.'s papers recall the 'pressure cooker' (Perryman et al., 2011) in which teachers find themselves because of accountability measures, and which affects their feelings and behaviours.

The future of metrics

The final paper in this issue, by María Teresa Flórez Petour, Tamara Rozas Assael, Jacqueline Gysling Caselli and José Miguel Olave Astorga, takes an ambitious look at the wider question of whether metrics in education are achieving what they are intended to achieve, and argues that they are in fact inhibiting the promotion of social justice through education. The authors set out the case for a wide re-examination of educational assessment set within the context of metrics and their consequences. To those caught up in the cycle of performativity and control by metrics their vision might seem utopian, but perhaps the time has come for us to move beyond metrics in their current state.

Flórez Petour et al. ask in whose interests educational metrics are: and we echo that question at the end of this editorial. While ensuring that public money is well spent is an important matter, many metrics in educational contexts have spawned a variety of activities which can be both costly in themselves and of questionable value to the larger enterprise of education.

If gaming, cheating and malpractice are features of metrics systems, then, as mentioned above, regulation is surely an important way of combating these undesirable side-effects. Essentially, the 'market' conditions and actors' behaviours in it can be influenced by regulatory bodies, such as the new Office for Students in England, which will regulate higher education institutions, deciding which bodies are fit to practice, fining institutions for poor performance and so on. Regulation may be necessary, but it is not a magic bullet and unintended consequences of revised metrics are sure to arise. Within the healthcare 'market', England has recently seen a case[1] in which it appears that all of the regulatory bodies involved failed and, as mentioned above, we know that this happened with the global economic crisis of 2008.

Regulation as the answer certainly falls within the thought territory of the neoliberal. Stepping beyond that, we might question whose interests metrics serve. Whilst educational professionals are busy record-keeping, what is being neglected that might be more important? What metrics are not being given attention? Metrics at state level for issues that are under political control are not the focus of scrutiny in such systems. Metrics systems push the onus away from the centre, onto individuals and institutions, even if they do not have the resources to address the key issues facing their professions.

Teacher shortages are a key example. Teachers are widely recognised as the key factor in educational improvement, but there are widespread teacher shortages in England. Individual schools can do little to alter the labour market conditions when such shortages exist and none of the metrics systems address this problem (NAO, 2017). As Muller (2018, p.18) noted, 'measuring only a few aspects creates incentives to neglect the rest.' As such, political accountability is not strengthened by these systems – an issue identified by Ozga (2012) in relation to metrics from international tests. Metrics, their implementation, effects and their regulation are highly costly: the opportunity costs go uncalculated.

One approach could be to try to limit the life of their own that metrics have taken on, to reduce the distance between the signifier and the signified, in Hanson's (2000) terms. This would ensure that we focused more upon what was valued and considered important, reduce performativity, disenfranchisement and possibly even managerialism. However, it is well recognised that measuring what matters is not as easy as it sounds. In all likelihood though, this is where attention needs to focus because monitoring and metrics will blossom in the ever-more digital future. Artificial intelligence and other computational techniques are already flourishing to provide information about individuals,

[1] https://www.bbc.co.uk/news/uk-england-44547788

including in their working lives.[2] If we are not careful, we will spend all of our time tracking the trackers. Metrics systems of the future are likely to be less transparent, controlled by the private sector or unknown agencies, cross-sector and highly creative. As we keep being reminded, we are the commodities for these systems. This Special Issue has been in the planning for over two years, since well before the furore over Facebook, Cambridge Analytica, Aggregate IQ, Brexit and the Trump election. If Winston Smith's thought control in George Orwell's *Nineteen Eighty-Four* seemed relevant to the metrics agenda then, it is now even more so. As the articles in this issue show, we need better ways of governing our education systems that enhance fairplay and professionalism, rather than undermine them.

Jo-Anne Baird
Professor of Educational Assessment, Oxford University

Victoria Elliott
Associate Professor of English and Literacy Education, Oxford University

References

Baird, J., Johnson, S., Hopfenbeck, T.N., Isaacs, T., Sprague, T., Stobart, G. & Yu, G. (2016). On the supranational spell of PISA in policy. *Educational Research*, 58(2), 121–138. Special Issue: International Policy Borrowing and Evidence-based Educational Policy Making: Relationships and Tensions.

Ball, S.J. (2003). The teacher's soul and the terrors of performativity. *Journal of Education Policy*, 18(2), 215–228.

Baumert, J., Ludtke, O., Trautwein, U., & Brunner, M. (2009). Large-scale student assessment studies measure the results of processes of knowledge acquisition: Evidence in support of the distinction between intelligence and student achievement. *Educational Research Review*, 4, 165–176.

Bevan, G. & Hood, C. (2006). What's measured is what matters: targets and gaming in the public health care system. *Public Administration*, 84, (3).

Bradbury, J. (2018). *Off-rolling: using data to see a fuller picture.* Available online at: https://educationinspection.blog.gov.uk/2018/06/26/off-rolling-using-data-to-see-a-fuller-picture/.

Bushnell, M. (2003). Teachers in the schoolhouse panopticon: Complicity and resistance. *Education and Urban Society*, 35(3), 251–272.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Foucault, M. (1975). *Discipline and punish: The birth of the prison*. New York, NY: Random House.

Gillard, J. W. (2018). An initial analysis and reflection of the metrics used in the Teaching Excellence Framework in the UK. *Perspectives: Policy and Practice in Higher Education*, 22(2), 49–57.

Grek, S. (2012). What PISA knows and can do: studying the role of national actors in the making of PISA. *European Educational Research Journal*, 11(2), 243–254.

[2] https://www.economist.com/leaders/2018/03/28/the-workplace-of-the-future

Gruber, K.H. (2006). The German 'PISA-Shock': some aspects of the extraordinary impact of the OECD's PISA study on the German education system. In H. Ertl (ed.) *Cross-national Attraction in Education: accounts from England and Germany.* Oxford: Symposium Books.
Hanson, F.A. (2000). How tests create what they are intended to measure. In A. Filer (ed.) *Assessment. Social Practice and Social Product*. London, Routledge, 67–81.
Hood, C. (1991). A public management for all seasons? *Public Administration*, 69, 3–19.
Hopfenbeck, T.N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J. & Baird, J. (2017). Lessons Learned from PISA: A Systematic Review of Peer-Reviewed Articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research*, 62(3), 333–353.
Leckie, G. & Goldstein, H. (2009). The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(4), 835–51.
Lyotard J.-F. (1984). The Postmodern Condition: A report on knowledge. *Theory and History of Literature*, Volume 10. Manchester: Manchester University Press.
Mallen, R.E. (1979). Recognizing and Defining Legal Malpractice. *SCL Rev*., 30, 203–212.
Muller, J.Z. (2018). *The Tyranny of Metrics*. Princeton, NJ: Princeton University Press.
NAO (2017). *Department for Education. Retaining and developing the teaching workforce*. HC 307 Session 2017–19. 12 September. Available online at: https://www.nao.org.uk/wp-content/uploads/2017/09/Retaining-and-developing-the-teaching-workforce.pdf.
Nuttall, D.L. (1987). The validity of assessments. *European Journal of Psychology of Education*, 2(2), 109–18.
O'Neill, O. (2002). *A question of trust: The BBC Reith Lectures 2002*. Cambridge University Press.
Oancea (2018, forthcoming). *Research governance and the future of research assessment*. Palgrave Communications.
Ozga, J. (2012). Introduction:. Assessing PISA. *European Educational Research Journal,* 11(2), 166–171.
Peck, J. (2002). Political economies of scale: fast policy, interscalar relations, and neoliberal workfare. *Economic Geography*, 78(3), 331–360.
Perryman, J. (2006). Panoptic performativity and school inspection regimes: Disciplinary mechanisms and life under special measures. *Journal of Education Policy*, 21(2), 147–161.
Perryman, J., Ball, S., Maguire, M. & Braun, A. (2011). Life in the pressure cooker–School league tables and English and mathematics teachers' responses to accountability in a results-driven era. *British Journal of Educational Studies*, 59(2), 179–195.
Popham, W.J. (2007). Instructional sensitivity of tests: accountability's dire drawback. *Phi Delta Kappan*, 89(2), 146–50.
Power, M. (1997). *The Audit Society. Rituals of Verification*. Oxford: Oxford University Press.

Quan, W., Chen, B. & Shu, F. (2017). Publish or impoverish: An investigation of the monetary reward system of science in China (1999-2016). *Aslib Journal of Information Management*, 69(5), 486–502.
Reay, D. & Wiliam, D. (1999). 'I'll be a nothing': structure, agency and the construction of identity through assessment. *British Educational Research Journal*, 25(3), 343–354.
Roberts, J. (2006). *A history of China* (2nd ed.). Houndmills: Palgrave Macmillan.
SamGyimah (2018). We want students to be better informed about degree choices & the returns - today, we're officially launching a competition for tech companies to take graduate data & create a MoneySuperMarket for students, giving them real power to make the right choice. [Tweet]. 26 June. Available online at : https://twitter.com/SamGyimah/status/1011530092965715968.
Usher, N. (2013). *Assessment for learning in secondary writing classrooms*. Unpublished Master's thesis, University of Oxford.