



Education International
Internationale de l'Éducation
Internacional de la Educación

Education International Discussion Paper by Tore Bernt Sørensen

Value-added measurement or modelling (VAM)

Summary

This discussion paper reviews the policy instrument of value-added measurement or modelling (VAM) and the implications the instrument has for teaching and learning in a global context. VAM is based on the assumption that it is possible to create adequately complex statistical models that capture the essential and universal factors in what makes some schools and teachers more effective than others without sacrificing the complexity of education, teaching and learning.

VAM is related to some of the dominant trends in education policy globally, and the origins and spread of the policy instrument should be understood within the context of the rise of the school effectiveness movement in education research, policy and practice. VAM is currently incorporated into market-based teacher accountability systems in a number of countries. While in the US the use of VAM as a policy instrument to evaluate schools and teachers has been taken exceptionally far in the last 5 years, most other high-income countries

remain cautious towards the use of VAM, as reflected in OECD reports on the policy instrument.

The paper unfolds the debates and critique raised against VAM. After a brief account of the origins, basic ideas and current use of VAM globally, four particular concerns related to VAM are discussed:

- 1) a technical critique of the statistical modelling underlying VAM;
- 2) a broader critique on the constitutive effects of VAM on education and its objectives;
- 3) the sidelining of teachers in the debate on evaluation of school and teacher performance; and
- 4) the promotion of VAM by private enterprises and major development agencies in low- and middle income countries.

On the basis of literature reviews, the paper points out that the alleged promises of VAM are undermined by deep flaws in terms of

reliability, validity, bias, and fairness. The effects of VAM include biased and unfair assessment of schools, heads and teachers, misdirection of resources, and the provision of misleading information to the public, parents and students. However, politically the reductionism of VAM has proved to have some appeal as a simple solution to fix complex realities. Combined with the media's hunt for headlines and the prospects for market- and profitmaking that comes with VAM, teacher unions should be aware of the characteristics of the policy instrument. In particular, education systems in low-income

countries might prove vulnerable in the coming years as international donors and for-profit enterprises appear to be endorsing VAM as a means to raise school and teacher quality in spite of the lack of evidence and the extensive critique raised against the instrument.

On the basis of literature reviews, the paper points out that the alleged promises of VAM are undermined by deep flaws in terms of reliability, validity, bias, and fairness

Value added measurement in education

In the last forty years, the school effectiveness movement has been highly successful in setting an agenda for education research and policy. The school effectiveness paradigm is fundamentally positive in its outlook, dispensing with what some might see as a structuralist determinism of the past in asserting that schools and teachers make a difference that can help to overcome background and contextual characteristics of students.

With school effectiveness, great hopes become linked with education. It is a distinctive hope based on a belief in progress through measurement, rational choice and management of incentives. Maximising the efficiency of education systems is meant to contribute to equality of opportunities, and in this sense school effectiveness to some extent aspires to shake up the social fabric and its inherited patterns of privilege.

(...) labelling education as the great equaliser comes across as a politics of distraction or, perhaps, resignation

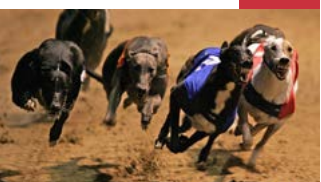
The school effectiveness outlook is thus a distinctive politics of hope, and positive visions of the power of education can only be welcomed, discussed and tried out. Moreover, much of the resulting political and public attention directed towards education is positive.

Yet, the belief in school effectiveness can be taken too far. When this happens, the politics of hope verge into a politics of distraction.

When Horace Mann (1848) heralded education as 'the great equaliser' in the mid-19th century he did so in a very particular context. Today, there is research consensus that there is only so much that education can do in terms of ensuring equality of educational opportunities. While teachers have been identified as important 'in-school factors' for student outcomes, this should be understood within the context that ever since the seminal Coleman report (Coleman et al. 1966) was released fifty years ago factors beyond the control of schools and teachers have been found to be much more important for student outcomes.

Therefore, it appears misleading when the allegory 'the great equaliser' is taken up by major figures in contemporary education policy like OECD Secretary General Ángel Gurría and former US Secretary of Education Arne Duncan (Duncan 2011; International Summit on the Teaching Profession 2016). In light of increasing inequalities, manipulation of national tax regimes by global corporations, and widespread dismantlement of welfare states, labelling education as the great equaliser comes across as a politics of distraction or, perhaps, resignation (Figazzolo and Harris 2011; Nolan et al 2014; Piketty 2014).

This discussion paper concerns the statistical policy instruments known as value-added measurement or modelling (VAM), a child of the school effectiveness movement. There is by now a considerable literature on VAM, its applications and limitations in the evaluation of schools and teachers. Increasingly, there is criticism, also among researchers specialising in the instruments.



VAM has been in the spotlight also in Education International. In her review of teacher feedback and appraisal systems mainly in OECD countries, Figazzolo (2013) noted that VAM has raised serious concerns among teachers and their unions, especially in the US, where VAM has increasingly been incorporated into the evaluation of teachers. The critique includes that the calculated VAM scores are highly unstable and unpredictable, and this means that a teacher who appears to be ineffective in one year might appear effective in the following year with a high associated risk for unfair measures taken towards school staff. A broader critical point is that by tying the perception of school and teacher effectiveness to student scores on standardised tests, the use of VAM promotes a culture of competitiveness among teachers centred on teaching to standardised tests and a narrowing of the curriculum. In addition, the incorporation of VAM into evaluation frameworks might discourage teachers from wanting to work in schools with more disadvantaged students.

(...) VAM sidelines educators from having a voice in the preparation of school and teacher evaluation frameworks. VAM closes down the debate on education

The paper unfolds the critical debates on VAM. It makes the case that the use of VAM shows that the high hopes attached to education may turn into misdirected efforts to drive up standards. VAM leaves a trail of distorting and distracting noise in the system and among its users and the public. Schools and teacher evaluation frameworks

should be proportionate and tempered by the fact that there are indeed limitations to how much education can 'compensate for society'.

Moreover, the use of VAM sidelines educators from having a voice in the preparation of school and teacher evaluation frameworks. VAM closes down the debate on education rather than opens it up.

On this basis, it seems perfectly reasonable to suggest that VAM – in direct opposition to the original intentions – undercuts innovation and effective reform in education. When VAM is incorporated into market-based accountability systems operating according to a Fordist mode of production, there is a real risk that the eternal quest for raising standards hollow out the meaning of education as an individual pursuit and collective good.

The paper first provides a brief account of the origins, basic ideas and current use of VAM globally. Subsequently, four particular concerns related to VAM are discussed in separate sections:

- 1) a technical critique of the statistical modelling underlying VAM;
- 2) a broader critique on the constitutive effects of VAM on education and its objectives;
- 3) the sidelining of teachers in the debate on evaluation of school and teacher performance; and
- 4) the promotion of VAM by private enterprises and major development agencies in low- and middle income countries.

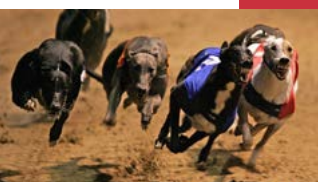
VAM and the search for the teacher effect

The school effectiveness movement in many ways started out with a progressive vision to address inequality of educational outcomes and school segregation (see for example Rutter et al. 1979). A child of the movement, the debate on VAM has been going on since the 1970s, mainly centred on the US and England where the school effectiveness movement emerged and gradually came to set the research agenda on educational issues. Spurred on by human capital theory, the research and political interest in the returns of the investment in the education sector increased from the 1960s. In the following decades and in accordance with the rise of new public management, student

attainment has gradually become the touchstone for assessing and comparing the performance of schools and teachers (Normand 2008).

This raised the question of how well schools overcome differences between the socio-economic background of their intakes, and whether some schools are more effective than others with students of similar backgrounds.

Since the performance of schools and teachers cannot be fairly assessed on the basis of the raw-score attainment of their students due to the considerable variation



in school intakes (in terms of realities like poverty, home language and parents' educational background), the value-added approach to judging school performance therefore appeared to be a perfectly reasonable idea. Instead of the absolute levels of attainment of students, schools and teachers were to be judged by the progress that their students make during attendance at the school (Gorard 2013).

Underlying VAM are thus the notions of 'school effect' and 'teacher effect'. With VAM effectiveness translates into the added value in terms of student attainment in standardised tests over a period of time.

Accordingly, school effectiveness researchers have since the 1970s developed increasingly sophisticated statistical models with the objective to identify school and teacher effects based on the claim that while much of the variation in school outcomes is due to school intake characteristics, the residual variation, that is, the difference in raw-scores unexplained by student intake, is evidence of differential school and teacher effectiveness (e.g. Chetty et al. 2013a; Chetty et al. 2013b; Gray and Wilcox 1995; Hanushek 1973; Hanushek 1979; Hanushek 2011; Hanushek and Rivkin 2010; Kyriakides, 2008; Meyer 1997; Woessmann 2011).

While VAM in principle appears superior to raw attainment score models, this does not imply that VAM delivers on its promises. Moreover, major studies - based on much of the same data - have continued to argue that school and teacher effects are minimal or non-existent once student intake differences have been taken into account (Coleman et al. 1966; Gorard 2000; Lubienski & Lubienski, 2006).

Indeed, a fundamental difference in view remains:

Is the variation in school outcomes unexplained by student background just the messy stuff left over by the process of analysis?

Or is it large enough, robust and invariant enough over time, to be accounted a school 'effect'?

Can we promote, reward and reprimand schools and teachers on this basis?

(Citation from Gorard 2010, p.746)

Reflecting on this issue, we should note that while some VAM studies have estimated

teachers to be the most significant in-school factor in student achievement scores, VAM studies overall find that the aspects of educational effectiveness that are measurable and within teacher control represent a small part of the total variation in student test scores or growth. Most VAM studies thus attribute between 1-14% of the variation in scores overall to teachers (ASA 2014; Holloway-Libell and Collins 2014).

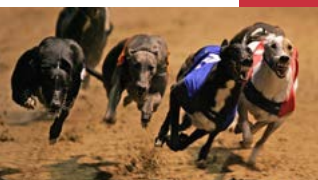
Politically, school effectiveness research has two appealing features: it offers solutions to complex issues, and those solutions appear technical, scientific and based on 'objective' results. These features owe much to the discipline of economics

Conversely, the majority of the variation in test scores is attributable to factors outside of teacher control such as student and family background, poverty, curriculum, and unmeasured influences. Such out-of-school factors are up to three times more powerful than school and teacher factors. These numbers are put in perspective by the more general observation that by the time students turn 18, they have spent about 90 percent of their lives outside formal learning settings (ASA 2014; Berliner 2009, 2013).

It should be emphasised that these findings do not mean that teachers overall have little effect on students. However, the *variation* among teachers accounts for a small part of the variation in student scores (ASA 2014; Gorard et al. 2013).

Regardless of these numbers, there has been remarkably little opposition to the adoption of school effectiveness as the guiding light for educational research, policy and practice. From the 1980s, the school effectiveness paradigm was embraced globally, and the focus of much educational research turned to attempts to identify 'school effects' and 'teacher effects' on the basis of different varieties of VAM (Gorard 2010; MacBeath 2012; Normand 2008).

Politically, school effectiveness research has two appealing features: it offers solutions to complex issues, and those solutions appear technical, scientific and based on 'objective' results. These features owe much to the discipline of economics. Like other public policy sectors, education has been the subject of an 'economic imperialism'.



Economics have proven successful with its objective to unify thought by providing a language that can be used to understand a wide range of social phenomena (Fourcade et al. 2014; Lazear 2000).

Economics provide bold assertions of universal truth claims on efficiency-maximisation based on allegedly scientific procedures of testing and revising refutable theories. The economic outlook is narrowing by design, seeking to strip away complexity to identify what is essential. Abstract modeling comes at a price though because

when the simplifying assumptions are taken too far, it might undercut the analysis and narrow the focus of the researcher. On the one hand, this allows for the formulation of solutions. On the other, it also neglects larger features of the problem at hand (Lazear 2000, 99-100). This trade-off between abstract modeling and allowing for complexity sums up many of the issues related to VAM, and we will turn to these further below.

VAM, GERM and the geographical spread of VAM

VAM is closely related to what Sahlberg (2011) has termed the *Global Educational Reform Movement* (GERM) which since the 1980s has radically altered education sectors throughout the world with an agenda of evidence-based policy based on the school effectiveness paradigm. Governments, international development agencies, consultant firms, some bilateral donors, venture philanthropy and major transnational policy actors such as the OECD and the European Commission have to some extent all bought into GERM (see figure 1).

GERM is centred on the creation of educational market places, with central authorities exercising 'steering from a distance' and schools competing to attract parents and students. GERM thus combines the centralised formulation of objectives and standards, and monitoring of data, with the decentralisation to schools concerning decisions around how they seek to meet

standards and maximise performance in their day-to-day running.

Sahlberg (2011) puts forward a profound critique of GERM by suggesting that the introduction of GERM policies is likely to have perverse effects in terms of promoting an excessive focus on performance leading to a narrowing of the curriculum, undermining cooperation in and between schools, and de-motivating teachers and students in the process. Ultimately, Sahlberg asserts that GERM therefore undercuts the very system innovation it was meant to further.

While the GERM concept and critique might be criticised for being somewhat general, it provides a powerful lens for examining more closely the clusters of policies in place in specific locations. The paper shows in more detail below that the use of VAM indeed epitomises the perverse effects of GERM. It appears perfectly reasonable to suggest that

the use of VAM in school and teacher evaluation frameworks promotes a most unhelpful short-termism in educational policy and practice that distracts debate from the real issues and hinders long-term reform efforts to ensure educational quality, effectiveness and efficiency.

VAM being directly tied to the evaluation of schools and teachers has mainly taken place in the US and England, both hubs for the conception and spreading of GERM and the school effectiveness movement. In both entities, VAM has been introduced as an integral part of a cluster of policies seeking to create market mechanisms and encourage customer behaviour in the education sector, including centralisation of educational

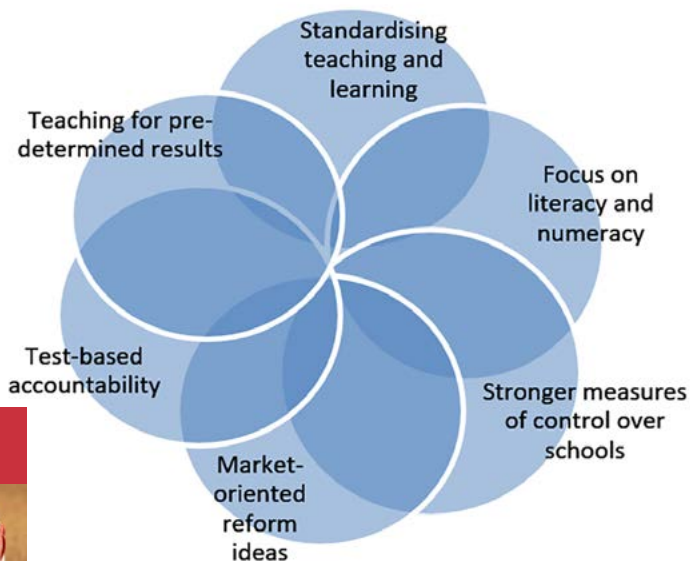
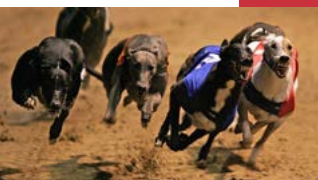


Figure 1. Six features of GERM (based on Sahlberg 2011)



objectives, curriculum and assessment frameworks, school choice, publication of attainment scores and school league tables, and decentralisation of budget decisions to schools (Holloway-Libell and Collins 2014; Stevenson and Wood 2013). Overall, the restructuring of education along these lines has had immense negative implications for educators and their unions in terms of de-professionalisation, performance-related pay, and fragmenting efforts to reach collective agreements for teachers (Carter et al. 2010; Robertson 2000).

The US warrants particular attention. By 2014, a variety of VAM models had been adopted in 44 states and the District of Columbia (Collins and Amrein-Beardsley, 2014). In these locations, VAM scores feed into the evaluation of individual head teachers and teachers and thus have direct consequences for them, in terms of advancement, pay, and termination of employment. In addition, VAM models are in some instances incorporated into the evaluation of teacher training programmes (AERA 2015). Prompted by federal financial incentive programs such as Race to the Top and the Teacher Incentive Fund grants program, many states, school districts, and administrators, are now for the first time in history evaluating teachers by methods that are up to 50% based on their VAM scores (Holloway-Libell and Collins 2014). These initiatives have proved extremely controversial among educators across the US (Tareen 2012).

It is remarkable that the American Statistical Association and American Educational Research Association have responded to the political trenchancy of VAM in the US by issuing statements calling for more reflection in the use of the policy instrument

In addition, the major research project *Measures of Effective Teaching* (MET Project), funded by the Bill and Melinda Gates Foundation, in its investigation of which teaching practices, and which teaching effectiveness measures, best predict future achievement gains, applied VAM and subsequently advocated the use of VAM as a component in teacher evaluation frameworks (Measures of Effective Teaching 2013).

Finally, it is remarkable that the American Statistical Association (ASA 2014) and

American Educational Research Association (AERA 2015) have responded to the political trenchancy of VAM in the US by issuing statements calling for more reflection in the use of the policy instrument.

VAM has taken a firm hold in policy and practice also in England, embedded in a 'pincer movement' of marketisation and managerialism that asserts increased control over teachers' work (Stevenson and Wood 2013). 'School Performance Tables' are published by state authorities, and the individual school results are used in setting targets and development plans. The national school inspection system in England, run by OFSTED, starts with VAM and the results partly pre-determine the judgment of school quality. VAM scores have been used to close schools down and reward or caution individual teachers or departments (Gorard 2010; Gorard et al. 2013).

It should be stressed that from a global perspective the US and England represent exceptions in how far VAM as a policy instrument has been taken. This reflects the particular trajectories of educational research and policy in the US and England. In other parts of the world, considerable caution has been exercised towards adopting VAM as a policy instrument for school and teacher evaluation.

Yet, with GERM and the school effectiveness paradigm being accommodated on a global scale, VAM has also been on the research and policy agenda elsewhere. We also need to take into account that with the thickening of global education governance, policy instruments like VAM might be promoted for public accountability as well as business purposes in low and middle income countries by aid agencies and private enterprises.

One of the countries that merit particular attention is Chile. The Chilean education system has been subject to one of the more radical variants of GERM and there is an interest in calculating VAM scores for teachers (Taut et al. 2014).

Moreover, varieties of VAM focusing on expected student exam scores at school level have been put in place in some countries. For example, in the Nordic countries of Denmark and Sweden state authorities introduced a socio-economic reference in the publication of school results that allows for the comparison between actual school performance and expected school performance considering the contextual



factors of parents' educational background, gender, and student origin (Ministeriet for Børn, Undervisning og Ligestilling 2016; Skolverket 2016). In Denmark, this data material has been used for publishing school league tables on the basis of a so-called 'tuition effect' [undervisningseffekt in Danish] (CEPOS 2016). While such simpler varieties of VAM do not incorporate students' prior test scores, they are in an OECD report labelled as 'context value added' (Nusche et al. 2011, p.82).

Concerning organisations with international horizons of action, the Organisation for Economic Cooperation and Development (OECD) has since the mid-1990s been one of the major catalysts of the school effectiveness paradigm globally through its peer review activities and survey and assessment programmes such as PISA, PIAAC and TALIS.

Due to its nature as an intergovernmental organisation collecting information from its member countries, the OECD has addressed the use of VAM in the education sector in a number of reports (OECD 2005, 2008, 2009, 2013a, 2013b). In this light, the general caution shown towards the use of VAM in evaluation frameworks for schools and teachers is remarkable. Treading carefully, OECD tries to find a balance in describing the widespread endorsement of VAM in the US and England along with the scepticism elsewhere. OECD (2013b, pp.35-36) symptomatically, on the one hand, endorsed VAM as a valuable component of teacher evaluation frameworks and present model examples from the US, while on the other hand advocated caution by suggesting that VAM is more relevant for whole-school evaluations and that VAM should not be used as a sole measure of teacher performance.

The World Bank has in recent years launched several projects on teachers in development contexts. The World Bank is more strident than the OECD in its suggestions linking teacher performance with evaluation frameworks, pay and career advancement (World Bank, 2011; 2012, pp.34-36). Less caution is hence exercised in the suggestion that VAM could be incorporated in teacher evaluation frameworks despite the fact that it is recognised that the policy instrument is flawed (World Bank 2012, p.29).

Moreover, in the current context of the UN Sustainable Development Goals 2030 and the renewed global focus on measurement of learning, the World Bank embraces

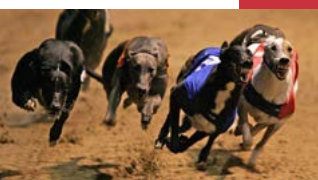
Results-Based Financing in Education. Various measures of teacher performance are heavily featured in RBF, and the World Bank asserts that there are many lessons to be learnt from the health sector (World Bank 2015).

Besides these well-known government-funded policy actors, the latest decade has seen the rise of private consultancies as influential knowledge brokers in the world of education. Companies like Pearson and McKinsey & Company have shown great interest in teachers and exercise considerable eclecticism in construing evidence meant to support sweeping claims (Coffield 2012). Pearson's Learning Curve Project (Pearson 2012, 2014) relies on input from some of the most prominent advocates of VAM, such as Eric Hanushek, Ludger Woessmann, and Raj Chetty. Considering Pearson's aspirations to be a global education company (Hogan et al. 2015; Junemann and Ball 2015), the company's interest in creating effective teachers commands vigilance particularly in low-income countries. The reports from McKinsey & Company follow a similar formula (Barber and Mourshed 2007, Mourshed et al. 2010; see Coffield 2012 for a critique).

Finally, we do not yet know much about current efforts of private school chains to assist school development in low and middle income countries by promoting VAM as part of 'school information systems' but the issue calls for research into the mapping of such initiatives and the conditions that enable this sort of VAM export (ARK 2015c; Elks et al. 2015).

The World Bank is more strident than the OECD in its suggestions linking teacher performance with evaluation frameworks, pay and career advancement (World Bank, 2011; 2012, pp.34-36)

Less caution is hence exercised in the suggestion that VAM could be incorporated in teacher evaluation frameworks despite the fact that it is recognised that the policy instrument is flawed (World Bank 2012, p.29)



Technical critique of the VAM bamboozle¹

¹ Section headline inspired by the blog VAMboozled.com edited by Audrey Amrein-Beardsley

VAM involves the ambition to create adequately complex statistical models that capture the essential and universal factors in what makes some schools and teachers more effective than others without sacrificing the complexity of education, teaching and learning.

However, the research literature points out a wide range of issues related to the validity and reliability of VAM models (AERA 2015; ASA 2014; Amrein-Beardsley 2008, 2014; Baker et al. 2010; Baker et al. 2013; Capital Hill Briefing 2011; Darling-Hammond 2015; Darling-Hammond et al. 2011; Haertel 2013; Hoyle and Robinson 2003; Goldring et al. 2015; Goldstein 2008; Gorard 2010; Gorard et al. 2013; Holloway-Libell 2015; Leckie and Goldstein 2009).

All of these issues cannot be discussed in-depth in this paper but this section provides a brief overview of the main areas of concern. Before we can do this, VAM needs to be introduced in more detail as a policy instrument based on statistical modelling.

there is a variety of VAM models. Their common characteristic is that they (...) seek to calculate a measure of change, the 'added value', to student learning over a period of time on the basis of standardised test scores

In fact, there is a variety of VAM models. Their common characteristic is that they for the purpose of raising the quality of education seek to calculate a measure of change, the 'added value', to student learning over a period of time on the basis of standardised test scores. VAM is thus a family of statistical models that typically use a form of regression model predicting student scores or growth on standardized tests from background variables (including prior test scores), and, possibly, student and school characteristics. The output is a set of estimates of value-added scores, one for each teacher or school, purporting to represent their relative effectiveness in improving student test scores. If a teacher's students have high achievement growth relative to other students with similar prior achievement, then the teacher will have a

high VAM score (Braun 2015).

The American Educational Research Association (AERA 2015) distinguishes between the following varieties of VAM models:

1. Growth models, also called gain score-based or mean gain models, which simply aggregate difference scores derived from subtracting previous scores from current scores on tests;
2. Transition-based models, or categorical models, which compute aggregate changes in performance categories over a period of 2 or more years;
3. Student growth percentiles-based models, which answer the question "What is the percentile rank of a student's current test score compared to students with similar previous test scores?" The individual teacher's VAM score is then calculated as the median or mean percentiles aggregated across her or his students;
4. Value-added measures-based models (VAM), which establish an expected current test score for students based on test scores from previous years, along with (possibly) other demographic characteristics of the student, classroom, and the school in attempting to account for the impact of factors beyond student achievement to isolate the teacher's impact.

VAM does not explain anything. VAM examines correlations and not causality. This means that positive or negative effects attributed to a teacher may have been caused by other factors not captured in the VAM model

Moreover, we should also note the basic qualities of VAM models. These are similar across the various models and imperative to acknowledge as much of the critique of VAM is related to these qualities.

First, since VAM is based on standardised test scores, VAM does not directly measure potential teacher contributions toward other student outcomes. Second, VAM does not



VAM cannot be relied upon when considering initial and propagated error. Whatever it is that VAM is measuring, if it is measuring anything at all, it is not a consistent characteristic of schools

explain anything. VAM examines correlations and not causality. This means that positive or negative effects attributed to a teacher may have been caused by other factors not captured in the VAM model. It also means that while VAM scores can be claimed to identify areas where improvement is needed, they do not provide information on how to go about it (ASA 2014).

The critique most commonly raised against VAM in both the US and England is that the VAM scores for schools and teachers are unstable

and unpredictable. This goes against the assumptions of VAM and undermines the claims of validity attached to the scores.

The fluctuating nature of VAM scores is associated with the distinct short-termism of evaluation procedures in many of the locations where the policy instrument has been adopted. As Bird et al. (2005) pointed out in their authoritative review of performance monitoring in public services there is a *"clear tension between the aim of performance monitoring to identify contemporary competence among practitioners or institutions and statistical potential to do so."*

In statistical terms, more extended periods of observation would provide a better basis for ranking or estimation. In particular, the common practice of comparing most recent values, such as this year's results with those from last year, might be very misleading (Bird et al. 2005). This critique applies more generally also to the publication of league tables, ranking individuals or organizational units, which tend to be produced on a yearly basis (Goldstein and Spiegelhalter 1996; Goldstein 2008).

In the English context, Stephen Gorard is one of the most vocal critics of VAM. He has in a series of papers systematically argued against the use of VAM in school and teacher evaluation frameworks (Gorard 2010, 2013; Gorard et al. 2013). Gorard and colleagues (2013) argues that it is remarkable how seriously VAM has been taken by researchers and governments since VAM scores are "useless or worse than useless" with current datasets due to misleading and detrimental effects when incorporated into evaluation frameworks. They flesh out the VAM procedure for schools in the following way:

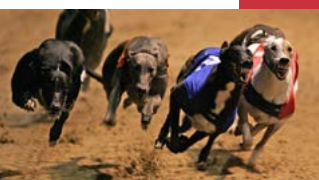
- Data on all pupils in the relevant school population is used to predict as

accurately as possible how well each pupil will score in a subsequent test of attainment.

- Any difference between the predicted and observed test result is then used as a residual.
- The averaged residuals for each school are termed the 'school effects'. These are intended to represent the amount by which students in that school progress more or less in comparison to equivalent pupils in other schools.
- A school with an average residual of zero is estimated to be 'performing' about as well as can be expected, given its intake.
- A school with an average above zero is doing better than expected.
- This judgement about progress is intended to be independent of the raw-score figures, making it fairer than assessment of raw scores.
- The 'school effect' is deemed a characteristic of the school, not its specific cohort of pupils. The 'school effect' should in principle be reasonably consistent over time where staff, structures, curriculum, leadership and resources of the school remain similar over time.

However, based on their analysis of the contextualised value-added scores (the VAM model used in England at the time) of all secondary schools in England 2006-2010, Gorard and colleagues (2013) show that the results of value-added approaches to estimate 'school effects' are not consistent, but very unstable over time, even over a 5 year period and with high quality data:

- The study asks how many schools with at least 99% of their pupils included in the VA calculations, and with data for all years, had VA measures that were clearly positive for five years. The answer is - none.
- VAM cannot be relied upon when considering initial and propagated error. Whatever it is that VAM is measuring, if it is measuring anything at all, it is not a consistent characteristic of schools.
- To find no schools with five successive years of positive VA means that parents could not use it as a way of judging how well their primary age children would do at age 16 in their future secondary school.
- What is true of the contextualised value-added model is almost certainly true of VAM approaches more generally, whether for schools, colleges,



departments or individual teachers, in England and everywhere else.

- Until their problems have been resolved by further development to handle missing and erroneous data, value-added models should not be used in practice. Commentators, policy-makers, educators and families need to be warned.
- If value-added scores are as meaningless as they appear to be, there is a serious ethical issue wherever they have been or continue to be used to reward and punish schools or make policy decisions.

Gorard et al. (2013) point out that the data quality of student records is imperative in VAM. Missing data creates an initial error component, that is, a source of inaccuracy and bias, and there is no way of adjusting for this statistically (Amrein-Beardsley 2008). Attempts to consider social complexity in the form of student background characteristics and contextual variables means that more data is needed and thereby the risk for missing data is further raised.

Gorard labels such calculations as “pseudo-quantification of the worst kind” because the initial figures are not accurate enough to sustain this kind of procedure

Moreover, creating valid, comparable and reliable attainment scores on the basis of standardised tests is in itself fraught with difficulty (Gorard et al. 2013). This is a fundamental issue because it means that VAM models generally have been found unable to fully account for the differences in student backgrounds and learning differences.

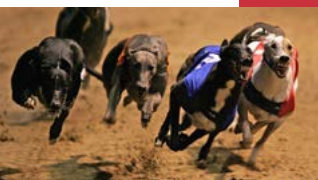
AERA (2015) thus points out that standardised tests in the US vary in the degree to which they fully capture the target constructs, as well as in their levels of precision across the range of reported scores. In addition, current state tests, by federal requirement, measure only grade-level standards without including items needed to measure growth for students who perform well below or well above grade level.

On the basis of Gorard's technical critique (Gorard 2010; Gorard et al. 2013), we might summarise the limitations of VAM as a statistical tool in these four points:

1. All initial errors ‘propagate’ through the VAM calculations. In other words, everything we do with our achieved measures we also do with their measurement errors, and these are therefore compounded to generate a far higher level of error in the residuals. The calculations thus leave the final VAM scores to be disproportionately made up of relative error terms.
2. The field of school effectiveness research works on the invalid assumption that errors in the data are random in nature and so can be estimated, and weighted for, by techniques based on random sampling theory. These techniques are fatally flawed, in their own terms, even when used ‘correctly’ with random samples.
3. VAM scores are heavily dependent on the raw-scores that have been rejected as a fair assessment of school effectiveness. This means that VAM may not be as fair as it claimed to be.
4. Since VAM scores are operationally defined simply as a measure of school or teacher effectiveness, we cannot compare them against anything except themselves. There is no criterion-related validity (see Amrein-Beardsley 2008), that is, any external referent or standard scale to judge them accurate or inaccurate with. We might say that there is nothing to calibrate VAM scores with. VAM scores thus emerge as magic figures from a long-winded and quasi-rational calculation.

The latter point is driven home by the fact that in England 2006-2010, VAM scores were based on calculations supposed to be accurate to at least four decimal places. Individual point scores represented to two decimal places (such as 29.56) are thus multiplied with coefficients with four decimal places (such as +0.3807). Gorard (2010) labels such calculations as “pseudo-quantification of the worst kind” because the initial figures are not accurate enough to sustain this kind of procedure.

The same impression comes from contemplating the calculations predicting attainment scores for any pupil in the English Key Stage(KS) 4 in 2007 (see table 1). Equivalent models of VAM apply to other stages of schooling.



- 162.1
- +0.3807 * (the squared school average KS2 score)
- 5.944 * school average KS2 score
- +1.396 * (KS2 English points - school average KS2 score)
- 0.109 * (KS2 maths points - school average KS2 score)
- 27.1 (if in care)
- 59.51 * IDACI score (index of deprivation)
- 34.37 (if School Action Special Educational Needs)
- 65.76 (if Action Plus or statement of Special Educational Needs)
- 73.55 (if joined after September of year 10)
- 23.43 (if joined not in July/August/September of years 7-9)
- +14.52 (if female)
- 12.94 * (age within year, where 31 August is 0 and 1 September is 1)
- + for English as an additional language pupil only (-8.328 -0.1428*(school average KS2 score)² + 4.93 * school average KS2 score)
- + ethnicity coefficient (from pre-defined list with 19 different ethnic groups)
- + for Free School Meals pupils only (-22.9 + FSM/ethnicity interaction, from a pre-defined table)
- + 1.962 * cohort average KS2 score
- 4.815 * standard deviation of cohort average KS2 score

Table 1. Calculation of predicted VAM score of pupil in Key Stage 4 education in England in 2007 (Adopted from Gorard 2010, pp.747-748)

At this stage, it appears relevant to quote Edward Haertel, Emeritus Professor at Stanford University (Haertel 2013, p.24):

“No statistical manipulation can assure fair comparisons of teachers working in very different schools, with very different students, under very different conditions.”

11

Most of the critical points above are included in the statements from the two major research associations of ASA (2014) and AERA (2015). Both of them call - in more subtle language - for the fundamental reconsideration of VAM and its uses as a policy instrument in the evaluation of schools

and teachers due to their scientific and technical limitations in actually capturing the complexity of what makes some schools and teachers ‘better’ than others, as measured by student attainment in tests.

Referring to the US context, the ASA and AERA statements urge that linking VAM scores with evaluation frameworks might potentially lead to unfair treatment of teachers and leaders in terms of advancement, compensation, and termination. Moreover, due to the risks for bias resources may be misdirected, and the educational system as a whole can be degraded. Ranking teachers by their VAM scores can have unintended consequences that reduce quality (ASA 2014; AERA 2015).

Moreover, the use of VAM presents additional substantial challenges in the evaluation of principals and nonteaching staff, and the limitations of using VAM are further compounded when used to compare the effectiveness of educator preparation programs based on the aggregation of graduates’ performance as teachers or leaders (AERA 2015).

In this respect, we should also note the critique of one recent major research effort into teacher effectiveness, the Gates Foundation’s Measuring Effective Teaching (MET) project. In the light of the points made above, it is particularly interesting given the focus and scope of the project that Jesse Rothstein in his thorough reviews of the main reports points out many of the same limitations. Rothstein notes that MET’s positive conclusions on the relevance of VAM and the weight given to student performance in teacher evaluation framework (between 33-50 percent) do not appear to be supported by the data but predetermined and a matter of judgement (Rothstein 2011; Rothstein and Mathis 2013).

However, given the overwhelming critique it is remarkable that the US professional associations of ASA and AERA do not call

Both of them call - in more subtle language - for the fundamental reconsideration of VAM and its uses as a policy instrument in the evaluation of schools and teachers due to their scientific and technical limitations in actually capturing the complexity of what makes some schools and teachers ‘better’ than others, as measured by student attainment in tests



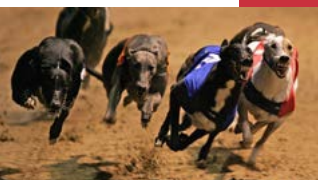
1. VAM scores must only be derived from students' scores on assessments that meet professional standards of reliability and validity for the purpose to be served.
2. VAM scores must be accompanied by separate lines of evidence of reliability and validity that support each claim and interpretative argument.
3. VAM scores must be based on multiple years of data from sufficient numbers of students.
4. VAM scores must only be calculated from scores on tests that are comparable over time.
5. VAM scores must not be calculated in grades or for subjects where there are not standardized assessments that are accompanied by evidence of their reliability and validity.
6. VAM scores must never be used alone or in isolation in educator or program evaluation systems.
7. Evaluation systems using VAM must include ongoing monitoring for technical quality and validity of use.
8. Evaluation reports and determinations based on VAM must include statistical estimates of error associated with student growth measures and any ratings or measures derived from them.

Table 2. AERA technical requirements for the use of VAM (AERA 2015)

While the body of research literature on methodological issues with VAM is growing, there is not much research on the experience of VAM by those most directly affected by the policy instrument, that is, teachers and school leaders

12

This baffling practice fails on two counts: the scores neither reflect the achievements of the teachers' own students nor contain any useful information for improvement



for the abandonment of VAM. In particular, AERA (2015) call for more research into VAM, with more rigid demands for data quality to meet a very high 'technical bar'. These are presented as a series of 'technical requirements' for the use of VAM to be 'scientifically rigorous and fair' (see table 2).

Thinking about the school effectiveness paradigm and VAM as an 'institutional regime' with a path-dependent trajectory sustained over time by groups of researchers and policy-makers, we might understand the call from AERA – a prominent research community with a direct stake in the policy instrument – to 'raise the bar' and further entrench VAM research as a reflection of the rigidity of the institutional regimes and the distinctive rationality it is based on.

While the body of research literature on methodological issues with VAM is growing, there is not much research on the experience of VAM by those most directly affected by the policy instrument, that is, teachers and school leaders.

However, the available research findings indicate that their experience corresponds with the weaknesses identified in the literature on methodological issues. Teachers and school leaders thus express that they find VAM inaccurate and biased and therefore of very limited use for evaluation, for making instructional decisions, or identifying potential areas for professional development (Collins 2012; Goldring et al. 2015; Jiang et al. 2015).

Furthermore, educators' trust in VAM scores are most likely negatively affected by the egregious practice that a majority of teachers' VAM scores are based on students or subjects that they do not teach. The vast majority of teachers are excluded from VAM because only teachers working with students in subjects that involve standardised testing are typically included in the models. This has led many states to attribute an aggregate, school-level value-added score to the non-tested grade level and content area teachers. This baffling practice fails on two counts: the scores neither reflect the achievements of the teachers' own students nor contain any useful information for improvement (Braun 2015; Collins & Amrein-Beardsley 2014; Harris 2011; Jiang et al. 2015).

Hitting the targets but missing the point:

the constitutive effects of VAM²

² The phrase 'hitting the targets but missing the point' adopted from Robert Wachter's opinion piece (2016)

This section focuses on the wider implications of VAM for teaching and learning and the students and parents engaging with schools. VAM is not merely a statistical tool adopted for the evaluation of schools and teachers. Like any policy instrument, VAM cannot help but have constitutive effects that go beyond the allegedly 'technical' evaluation of schools and teachers and the direct consequences in terms of feedback and appraisal, rewards and punishment. Ultimately, the use of VAM might have implications for how quality in education, teaching and learning is understood, as well as the role of education in society.

In terms of the curriculum, VAM encourages schools and teachers to prepare students for tests and focus on what is assessed and feed into the VAM scores, and to neglect those parts of the curriculum that are not

Again, there is no shortage of research and commentary on these issues. However, constitutive effects cannot be isolated as being discrete to VAM. As noted earlier, where VAM has been adopted as a policy instrument, it forms part of a cluster of policies that mutually sustain each other, usually for the purpose of entrenching market-based accountability systems in the education sector.

In her review for Education International, Figazzolo (2013) noted some of the main constitutive effects, including the promotion of a culture of competitiveness in school life, more teaching to standardised tests, and narrowing of the curriculum. Moreover, VAM might affect where teachers would like to work.

These effects are indeed among those most often identified in the research literature as well as commentary such as that provided by Diane Ravitch (2013, 2014), one of the most prominent and outspoken critics of VAM.

First of all, VAM have multiple implications for teachers and school leaders when incorporated into evaluation frameworks.

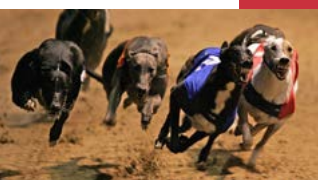
Effectively, this means that they are being rewarded or punished on the basis of unstable and unpredictable evidence. For schools, VAM scores tend to be used to determine funding allocations and potentially threaten them with closure. Moreover, the administrative burden of trying to understand and respond to VAM scores mean that there is less time for something more productive and meaningful, or that teachers and school leaders are under pressure to work more hours (Gorard 2010; Howard and Wood 2013).

In that respect, it is symptomatic that teachers in England are among those working most hours in a week and spending time on administrative tasks according to the OECD TALIS 2013 results (Micklewright et al. 2014, pp.47-50; OECD 2014, pp.387-388).

On this basis, it would appear obvious that the use of VAM for evaluation purposes has an impact on the behaviours and practices of teachers and school leaders. So far little research has been undertaken on such issues, but what we know is that the use of VAM has several detrimental effects, potentially furthering school segregation and inequality of educational opportunities and outcomes. In other words, VAM does not appear to serve educative objectives.

In terms of the curriculum, VAM encourages schools and teachers to prepare students for tests and focus on what is assessed and feed into the VAM scores, and to neglect those parts of the curriculum that are not. In this sense, VAM is associated with a narrow understanding of what education is for. Moreover, teachers are prompted to target instruction to students most likely to show growth, and an unwillingness to collaborate or share best practices with other teachers who were seen as competitors (Collins 2012; Darling-Hammond 2015; Jiang et al. 2015).

ASA hints at the same effect with their argument that overreliance on VAM scores may foster a competitive environment, discouraging collaboration and efforts to improve the educational system as a whole



ASA (2014) hints at the same effect with their argument that overreliance on VAM scores may foster a competitive environment, discouraging collaboration and efforts to improve the educational system as a whole.

On this basis, we might ask what such constitutive effects mean for teacher quality and the attractiveness of the teaching profession in the longer term?

(...) there is a real risk that those teachers and schools working with the most challenging students in the most challenging contexts, as well as those whose students are among the most capable, and those who eschew teaching to the test, are identified as the less capable ones

In this respect, Goldhaber (2015) outlines three broad categories of concern: i) how might VAM help in changing the supply of people who opt to pursue a teaching career and are selected into the labour market; ii) changing the effectiveness of those currently teaching and; iii) changing which teachers elect to, or are permitted to, stay in teaching.

While Goldhaber (2015) himself adopts the cautious position that the jury is still out on how the quality of the teacher workforce might be affected by VAM. This is to some degree a respectable statement given that there is not any available evidence on such large-scale effects due to the relatively short period that VAM has been in use, and moreover it would be nonsensical to speak of isolated VAM effects given that VAM policies are part of larger policy clusters.

However, taking the amount of critique on VAM and the constitutive effects into

account, it appears highly unlikely that VAM would make any positive contribution to the quality of teaching or the teacher workforce. Linda Darling-Hammond (2015) is adamant in this judgment too.

Concerning the attractiveness of the teaching profession, VAM might at first glance add some superficial allure of technical empiricism (Gorard 2010). But, since any claims to scientific value are wholly unwarranted, we might ask who would like to enter a profession where random noise from misguided statistical tools feeds into the evaluation of you and your workplace?

Moreover, due to the fact that VAM models are unable to fully account for the differences in student backgrounds and levels of knowledge and skills, there is a real risk that those teachers and schools working with the most challenging students in the most challenging contexts, as well as those whose students are among the most capable, and those who eschew teaching to the test, are identified as the less capable ones (Darling-Hammond 2015). A likely behavioural response – and indeed completely rational within the VAM regime – is that the use of VAM especially might further discourage teachers from working in high-need schools or with high-need students (Jiang et al. 2015; Johnson 2015).

In this respect, an overemphasis on VAM scores might take the focus away from the actual achievement gaps in the raw attainment scores between privileged and disadvantaged student groups, rich and poor, or between ethnic and language groups. Context-sensitive VAM models might thus end up disguising the notion of equity or render it invisible by factoring them into the VAM predictions (Gorard 2010).

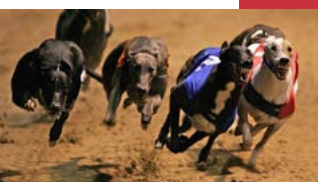
VAM, blame-avoidance and defensive education

In the discussion of constitutive effects, we might look to that other large public policy sector, the health sector, for signs of what VAM could contribute to in the longer term.

Like the education sector, health care has in some locations become excessively subjected to metrics and measurements. Wachter (2016) argues that while nobody is arguing that professionals should not be held accountable, the focus on numbers has gone too far. Trying to forge education and health care in the mould of business and

enterprise in the ways quality is measured block the very altruism that motivates people to enter the helping professions in the first place. More generally, VAM scores might show that targets are being met, yet the point of the larger endeavour is missed.

Hood (2011) identifies parallels in the ways professionals in the education and health sectors attempt to put themselves in a position where they cannot be blamed for failure. In particular, he asks whether public service organisations with the increased



Trying to forge education and health care in the mould of business and enterprise in the ways quality is measured block the very altruism that motivates people to enter the helping professions in the first place

emphasis on prescribed evidence-based routines are forced towards a hyper-defensiveness to avoid blame professionally, politically, and publicly?

Hood points out that the paradigm case of defensive medicine which has been around in the US for circa forty years can also be observed in the education sector in the forms of avoidance and assurance behaviour. The former involves excluding or expelling students and parents who would be believed to cause problems when following the standard best practice protocols. Assurance behaviour includes the rolling out of elaborate testing regimes so that evidence of educational progress is always at hand to counter charges of falling standards. In short, the world of defensive education is one in which rigid adherence to standard best practice protocols is far less risky than bespoke responses (Hood, 2011, pp.126-128).

Such observations appear particularly relevant in the case of VAM scores due to their instability and unpredictability, combined with their political and media appeal. As such, defensive education provides an exemplary yet unsettling case of the potential constitutive effects of VAM to be closely followed wherever the policy instrument is adopted.

VAM increases the importance of being able to 'game the system' and hence blame-avoidance. Hood's notion of 'defensive education' underpinned by assurance and avoidance behavior appears relevant as a pointer for future research on the wider, constitutive effects of VAM. Especially considering the fluctuating, unstable nature of VAM scores, it seems highly probable that teachers and school leaders would stick to conservative decisions and practices while crossing their fingers that this will get them safely through high-stakes assessment and evaluation procedures.

Moreover, Hood (2011) suggests that the increasingly used policy procedure of

protocolization furthers the defensiveness of public policy provision. Protocolization is meant to give due diligence protection to individuals and organisations in the event of blame or liability (Hood, 2011, pp.112-113; see also Bird et al. 2005 on performance monitoring protocols).

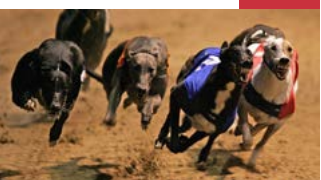
In short, the world of defensive education is one in which rigid adherence to standard best practice protocols is far less risky than bespoke responses

In this respect, we should note that there have been legal cases concerning questionable evaluation practices related to VAM. For example, a group of Florida teachers filed a lawsuit in April, 2013, on the grounds of being evaluated based on students whom they do not teach (Jordan, 2013). Other cases are likely to arise in the US due to the reliability and validity issues with VAM models currently used in state and district evaluation policies (Baker et al. 2013).

For example, a group of Florida teachers filed a lawsuit in April, 2013, on the grounds of being evaluated based on students whom they do not teach

This leads us back to Sahlberg's (2011) critique that contemporary trends in global education reform might undermine system innovation in the longer term. This hypothesis appears perfectly reasonable in the context of VAM. VAM models do not deliver valid and reliable results that can be used for innovation or development, individually or systemically. VAM is a misleading and wasteful exercise in terms of time and resources for everybody with a stake in education, students, parents, teachers, school leaders, and politicians (Braun 2015).

Successful education reform takes clear visions, dedication, and patience. With its short-term perspective and noise, VAM feeds anxiety and blame-avoidance and therefore hinders such processes to unfold.



Out of touch: VAM sidelines educators

Bangs and Frost (2012, p.27) ask whether teacher evaluation is seen as *“something done to teachers rather than a reflection of shared accountability where teachers’ professional learning is nurtured?”* By epitomizing a standards-based Fordist mode of production that de-professionalises teachers and de-politicises their work, VAM is clearly an example of something done to teachers. VAM is out of touch with educators, and educators are out of touch with VAM.

VAM is out of touch with educators, and educators are out of touch with VAM

The core issue underlying the debate on VAM concerns our concepts of quality in education, how it is to be measured, and who are involved in setting the measurement criteria.

If VAM adds anything to political and public debate, it does so on a misleading basis. Moreover, the complexity of VAM models excludes and disempowers the vast majority from that debate, including those, who pay taxes, work in or send their children to schools. Most academics are also excluded from understanding school effectiveness and VAM work (Normand, 2008).

Diane Ravitch decries the distance between VAM researchers and the lived experience of people engaged in education. Her comments can be taken further when considering the political implications of school and teacher evaluation frameworks based on highly sophisticated statistical models that only a selected few can operate, understand and explain

VAM involves a particularly detrimental outlook on the role of teachers and school leaders as agents in education. Diane Ravitch (2013, 2014) decries the distance between VAM researchers and the lived experience of people engaged in education. Her comments can be taken further when considering the political implications of school and teacher evaluation frameworks based on highly sophisticated statistical models that only a

selected few can operate, understand and explain.

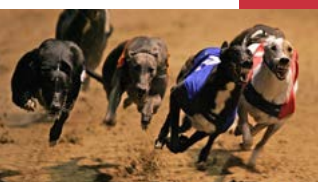
From the perspective of the teaching profession, this is a fallacy. It is unsettling for those committed to a scope of professional autonomy and expertise that ever more complex models comprehended by fewer and fewer people are being developed while politicians and researchers either do not understand them or ignore the serious issues associated with VAM (Gorard 2010). The use of VAM as a policy instrument effectively depoliticises the work of school and teachers, reducing teaching and learning to yet another subset of social engineering in a technocratic and utopian quest for perfectability (Crick 2013; Rittel and Webber, 1973, p.158; Stevenson and Wood 2013).

Despite the overwhelming concerns regarding VAM, the distinctly “heroic” assumptions underpinning the policy instrument continue to be propagated and too often accepted without challenge. The preponderance of sources propagating unfounded assertions are fostering a sort of VAM echo chamber that seems impenetrable by even the most rigorous and trustworthy empirical evidence (Holloway-Libell and Amrein-Beardsley, 2015).

VAM does not produce substantial information that can be used for system-level improvement or for identifying relevant areas of professional development and does not feed into professional development plans. Moreover, information on how VAM scores are calculated, error margins and limitations, is generally not available to administrators and teachers in accessible formats (Holloway-Libell and Collins 2014). So, VAM scores tend to remain mysterious and incomprehensible to teachers and school leaders (Goldring et al. 2015), captured by the remark from a teacher in Houston, Texas (Amrein-Beardsley and Collins 2012, p. 15):

“I do what I do every year. I teach the way I teach every year. [My] first year got me pats on the back. [My] second year got me kicked in the backside. And for year three my scores were off the charts. I got a huge bonus. ... What did I do differently? I have no clue.”

Unfortunately, Darling-Hammond (2015, p.134) is likely to be right in her assessment that educators’ current low levels of



The use of VAM as a policy instrument effectively depoliticises the work of school and teachers, reducing teaching and learning to yet another subset of social engineering in a technocratic and utopian quest for perfectability

confidence in VAM scores would be shaken even further if they knew how much error is associated with VAM.

Finally, the incorporation of VAM scores into evaluation frameworks might lead larger reform efforts to fail. When teachers in Chicago were introduced to a new evaluation framework, they were concerned about the inclusion of VAM scores, and the concerns turned to dissatisfaction with the overall framework over the course of implementation. Sixty-five percent of teachers thus reported that their evaluation relied too heavily on student growth, and half of them felt that the test data were not an accurate measure of their students' learning (Jiang et al. 2015).

There are calls for moderation in the use of VAM as a policy instrument, with the scores being one component in a more comprehensive use of educator or program evaluations. The argument goes that other measures of practice and student outcomes should always be integrated into judgments about overall teacher effectiveness (AERA 2015; ASA 2014). On this basis, the major statements from ASA (2014) and AERA (2015) on VAM call for raising the scientific rigour in the production of scores and moderation by limiting the importance of the scores in evaluation frameworks.

VAM market-making in low and middle income countries

The debate on VAM has so far centred heavily on the US and England. In these places VAM and school effectiveness has become an industry. There is money to be made and jobs to be created, in state and local bureaucracies, research institutions, schools, charities, foundations, consultancies, and software companies (Gorard 2010; Holloway-Libell and Amrein-Beardsley, 2015).

But, while other high-income countries remain cautious towards the use of VAM, there are troubling indications that VAM might be promoted in low and middle income countries.

However, on the basis of the body of research findings part of which has been referenced in this paper, one question needs to be asked:

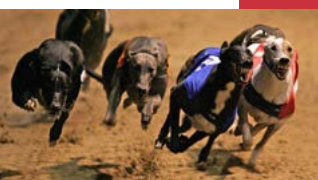
Do VAM scores have *any* place as a component in school and teacher evaluation frameworks?

It is clear that the monitoring of student test scores has its place as a component in evaluation frameworks of schools and teachers at school, district or system level. Where evaluation frameworks are based on professional standards, classroom observations, curriculum development, and a wide range of factors associated with teaching and teacher perspectives, such comprehensive and development-oriented methods are able to provide valuable information for school and system improvement. In this respect, the trust and commitment of educators and their unions are best gained when they are taking part in the conception of teacher-appraisal arrangements, policies and criteria (Figazzolo 2013).

Considering the technical critique and the constitutive effects of the policy instrument, it appears impossible to reconcile VAM with comprehensive, development-oriented evaluation frameworks.

One question needs to be asked:

Do VAM scores have any place as a component in school and teacher evaluation frameworks?



Yet, given the global prominence of the school effectiveness paradigm, and the current trends of increasing commercialisation and privatisation in and of education, the promotion and imposition of VAM in low and middle income countries is hardly surprising

for all with stakes in school effectiveness, and the global 'learning crisis' and United Nations Sustainable Development Goals with its enhanced focus on the measurement of learning might add momentum to these developments.

In this respect, we might note that VAM entrepreneurs could be large well-known companies such as Pearson with its global learning strategy to uncover new market

opportunities (see Junemann and Ball, 2015) as well as smaller operators.

One example of the latter is UK academy chain ARK which is currently opening schools in Uganda and Delhi, India, including the development of school information systems for data monitoring (ARK, 2015b, 2015c). ARK has also run VAM trials in Uganda. The model is taken from England, and thus one of those so heavily criticised by commentators like Stephen Gorard. Nonetheless, ARK finds their VAM pilot study so convincing that they recommend scaling the use of this model up to strengthen the school accountability system in Uganda (ARK 2015a; Elks et al. 2015).

Conclusion: VAM and the politics of distraction

The school effectiveness paradigm in education research and policy has over the decades become an institutional regime, developed and entrenched by research and policy communities with profound implications for policies and institutional arrangements in education across the globe. Held firmly in place by a wide range of policy-makers and researchers, the regime continues to influence our way of thinking about education, creating a sort of self-reinforcing bounded rationality based on a belief in progress through measurement, rational choice and management of incentives. In this sense, the regime of school effectiveness is rigid and path-dependent. It is clear that it cannot just be rolled back or dismantled.

VAM constitutes a frontier for the school effectiveness movement, a borderland where abstract principles of quantification correspond so little with the messy complexity of social reality that the use of such principles violate our sense of fairness and views of what education ought to be (see table 3). Perhaps, VAM could, due to its very hubris, spark some much needed reflection on the basic propositions of the school effectiveness paradigm, including the trade-off between abstract modelling and complexity that it relies on.

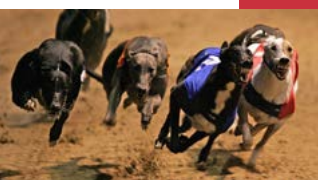
The use of VAM as a policy instrument in the education sector started out being based on a positive vision. But, when VAM is incorporated into market-based

accountability systems, there is a real risk that the quest for raising standards hollow out the meaning of education as an individual pursuit and collective good.

The critique of VAM is overwhelming. The policy instrument leads to predictable frustrations for all stakeholders, with recriminations that further undermine the prospects for real, sustainable improvements

According to the proponents of VAM, the policy instrument should make education systems more efficient. However, the research suggests that VAM due to ever instable and fluctuating VAM scores effectively paralyzes school systems in a sort of short-term thinking that is in stark contrast to the general consensus that wide-ranging educational reform takes time and dedication.

The critique of VAM is overwhelming. The policy instrument leads to predictable frustrations for all stakeholders, with recriminations that further undermine the prospects for real, sustainable improvements (Braun 2015).



VAM scorecard

Does VAM produce reliable and valid scores that can be used as an unbiased and fair basis for evaluation of schools and teachers and system level development and reform? No.

Does VAM produce distorting and distracting noise that further short-term thinking and anxiety among students, staff and parents? Yes.

Does VAM help to get the teacher profession on board with educational reform? No.

Does VAM further educational inequalities and segregation? Yes.

Does VAM seriously limit the capability of the system to innovate? Yes.

Table 3. A critical summary of VAM effects

So, why are policy-makers holding on to this biased and divisive policy instrument and accepting it is 'good enough' (Harris, 2011; Holloway-Libell and Collins 2014)?

As an attempt to synthesise the research evidence presented in this paper, it is hard to ignore the fact that VAM has been taken furthest in two political entities with some of the highest levels of inequality among high-income countries.

It was noted above that VAM might enhance blame-avoidance and defensive education. This critique can be extended by associating VAM with a politics of distraction. VAM is the perfect lever for both blame-avoidance and distraction since the policy instrument keeps the focus fixed on the 'education system' and maximising its performance by managing and adjusting incentives for teachers, school leaders, students and parents.

In terms of distraction, this narrow focus helps to distract from factors external to the system. The exposure of allegedly failing schools and teachers disguise that the most serious challenges to ensuring educational opportunities are related to poverty and disadvantage, issues beyond the control of schools and teachers (Berliner and Biddle 1995; Darling-Hammond 2015; Holloway-Libell and Collins 2014).

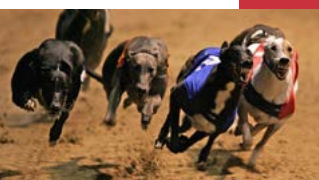
In terms of blame-avoidance, Bird et al. (2005) reminds us that performance monitoring in public services serves three purposes:

- assess the impact of Government policies on public services;
- identify well performing or underperforming institutions and public servants;
- public accountability of Ministers for their stewardship of the public services.

VAM is very much focused on the second purpose and hence constitutes a convenient solution for legislators and policy-makers as the system logic ensures that there are always new winners and losers of schools and teachers that can be used for setting policy agendas and gaining media and public attention, while keeping any notion of public accountability at a safe distance.

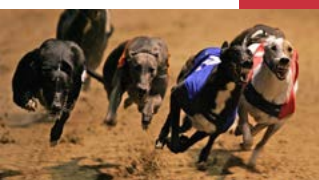
Finally, we might see it as the hallmark of the politics of distraction and blame-avoidance that VAM scores are wholly self-referential. They do not explain anything or tell how improvement is possible. In this way VAM is highly useful as a lever to keep the spin going while appeasing parents (the voters) and nurturing their customer behaviour on a misleading basis.

VAM constitutes a frontier for the school effectiveness movement, a borderland where abstract principles of quantification correspond so little with the messy complexity of social reality that the use of such principles violate our sense of fairness and views of what education ought to be

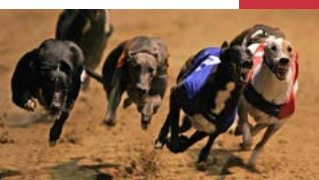


References

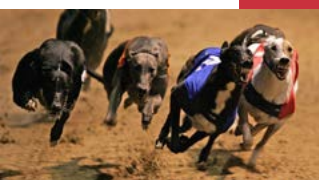
- AERA [American Educational Research Association] (2015). AERA Statement on Use of Value-Added Models (VAM) for the Evaluation of Educators and Educator Preparation Programs. *Educational Researcher*, 44(8), pp. 448–452, DOI: 10.3102/0013189X15618385
- Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system, *Educational Researcher*, 37(2), 65-75. <http://dx.doi.org/10.3102/0013189X08316420>
- Amrein-Beardsley, A. (2014). *Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability*. Routledge
- Amrein-Beardsley, A., & Collins, C. (2012). The SAS Education Value-Added Assessment System (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and unintended consequences. *Education Policy Analysis Archives*, 20(12). Available at <http://epaa.asu.edu/ojs/article/view/1096>
- ARK (2015a). *How we discovered the best school in Uganda (and what it means for the other schools)*. 17 December 2015. <http://arkonline.org/blog/how-we-discovered-best-school-uganda-and-what-it-means-other-schools>
- ARK (2015b). *International programmes*. <http://arkonline.org/our-programmes/international-programmes>
- ARK (2015c). *School Information Systems*. <http://arkonline.org/programmes/school-information-systems>
- ASA [American Statistical Association] (2014). *ASA statement on using valueadded models for educational assessment*. Retrieved from https://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf
- Baker, B.D. (2012). *Fire first, ask questions later? Comments on Recent Teacher Effectiveness Studies*. January 7, 2012. School Finance 101: Data and thoughts on public and private school funding in the U.S. <https://schoolfinance101.wordpress.com/2012/01/07/fire-first-ask-questions-later-comments-on-recent-teacher-effectiveness-studies/>
- Baker, E.L., Barton, P.E., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R.L., Ravitch, D., Rothstein, R., Shavelson, R.J., & Shepard, L.A. (2010). *Problems with the Use of Student Test Scores to Evaluate Teachers*. The Economic Policy Institute Briefing Paper 278. <http://www.epi.org/files/page/-/pdf/bp278.pdf>
- Baker, B. D., Oluwole, J., & Green III, P.C. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the race-to-the-top era. *Education Policy Analysis Archives*, 21(5).
- Ballou, D., & Springer, M.G. (2015). [Using Student Test Scores to Measure Teacher Performance: Some Problems in the Design and Implementation of Evaluation Systems](#). *Educational Researcher*, 44(2), 77-86.
- Bangs, J. & Frost D. (2012). *Teacher Self-efficacy, Voice and Leadership – Towards a Policy Framework for Education International*. Education International Research Institute and University of Cambridge Faculty of Education. Brussels: Education International.
- Barber, M., & M. Mourshed (2007). *How the world's best-performing school systems come out on top*. McKinsey & Company.
- Berliner, D. C., & Biddle, B. J. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools*. Reading, MA: Addison-Wesley Publishing Company, Inc.



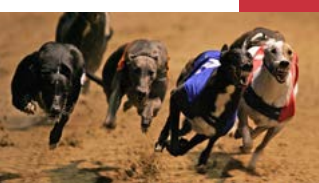
- Berliner, D.C. (2009). *Poverty and potential: out-of-school factors and school success*. Boulder, CO and Tempe, AZ: Education and the Public Interest Center, University of Colorado/Education Policy Research Unit, Arizona State University. Retrieved from <http://epicpolicy.org/publication/poverty-and-potential>
- Berliner, D. C. (2013). Effects of inequality and poverty vs. teachers and schooling on America's youth. *Teachers College Record*, 115(12). Retrieved from: <http://www.tcrecord.org/Content.asp?ContentID=16889>
- Bird, S., Cox, D., Farewell, V., Goldstein, H., Holt, T., & Smith, P. (2005). Performance indicators: Good, bad, ugly. *Journal of the Royal Statistical Society*, 168(1), 1–27.
- Braun, H. (2015). The Value in Value Added Depends on the Ecology. *Educational Researcher*, 44 (2), 127 –131.
- Campbell, D. T. (1976). *Assessing the impact of planned social change*. Hanover, NH: Public Affairs Center, Dartmouth College.
- Capitol Hill Briefing (2011). *Getting teacher evaluation right: A challenge for policy makers*. A briefing by E. Haertel, J. Rothstein, A. Amrein-Beardsley, and L. Darling-Hammond. Washington DC: Dirksen Senate Office Building (research in brief). Retrieved from <http://www.aera.net/Default.aspx?id=12856>
- Carter, B., Stevenson, H. & Passy, R. (2010). *Industrial Relations in Education: Transforming the School Workforce*. Abingdon, Oxon: Routledge.
- CEPOS (2015). *Undervisningseffekten*. Retrieved from <http://www.cepos.dk/publikationer-projekter/undervisningseffekten>
- Chetty, R., Friedman, J.N., & Rockoff, J.E. (2013a). *Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates*. NBER Working Paper No. 19423.
- Chetty, R., Friedman, J.N., & Rockoff, J.E. (2013b). *Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood*. NBER Working Paper No. 19424.
- Coffield, F. (2012). Why the McKinsey reports will not improve school systems, *Journal of Education Policy*, 27(1), 131-149. <http://dx.doi.org/10.1080/02680939.2011.623243>
- Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfield, F. & York, R.L. (1966). *Equality of educational opportunity*. National Center of Educational Statistics. Washington, US Government Printing Office.
- Collins, C. & Amrein-Beardsley, A. (2014). Putting growth and value-added models on the map: A national overview. *Teachers College Record* 116(1). Retrieved from: <http://www.tcrecord.org/Content.asp?ContentID=17291>
- Crick, B. (2013). *In Defence of Politics*. 5th edition. London: Bloomsbury.
- Darling-Hammond, L.(2015). Can Value Added Add Value to Teacher Evaluation? *Educational Researcher*, 44(2), 132-137.
- Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E.H., & Rothstein, J. (2011). *Getting Teacher Evaluation Right: A Background Paper for Policy Makers*. American Educational Research Association and National Academy of Education. Retrieved at <http://www.aera.net/Portals/38/docs/New%20Logo%20Research%20on%20Teacher%20Evaluation%20AERA-NAE%20Briefing.pdf>



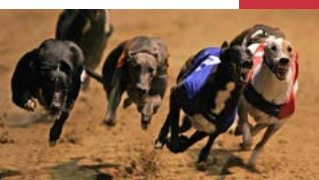
- Duncan, A. (2011). In Ferguson — and All of Our Communities — Education Can Be the Great Equalizer. Posted on *The Root*, 9 February 2015. Retrieved from <http://blog.ed.gov/2015/02/secretary-duncan-visits-ferguson/> Elks, P., Collins, J., & Morley, S. (2015). *Value-Added Measurement: using student progress rather than raw test scores to measure performance in secondary schools in Uganda*. Paper presented at UKFIET 2015, Oxford, 15-17 September 2015.
- Figazzolo, L. (2013). The use and misuse of teacher appraisal: An overview of cases in the developed world. Education International. Retrieved from <http://download.ei-ie.org/Docs/WebDepot/TeacherAppraisal.pdf>
- Figazzolo, L., & Harris, B. (eds. 2011). *Global corporate taxation and resources for quality public services*. Brussels: Education International Research Institute and Global Unions.
- Froese-Germain B. (2011), Weighing In On the Teacher Merit Pay Debate, March 2011, CTF Notes. <http://files.eric.ed.gov/fulltext/ED532555.pdf>
- Fourcade, M., Ollion, E., & Algan, Y. (2014). The superiority of economists. (Discussion Paper 14/3).
Retrieved from the Max Planck Science Po Center. website: http://www.maxpo.eu/pub/maxpo_dp/maxpodp14-3.pdf
- Goldhaber, D. (2015). Exploring the Potential of Value-Added Performance Measures to Affect the Quality of the Teacher Workforce. *Educational Researcher*, 44(2), 87-95.
- Goldring, E., Grissom, J.A., Rubin, M., Neumerski, C.M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make Room Value Added: Principals' Human Capital Decisions and the Emergence of Teacher Observation Data. *Educational Researcher*, 44(2), 96-104.
- Goldstein, H. (2008): Evidence and education policy – some reflections and allegations, *Cambridge Journal of Education*, 38(3), 393-400.
- Goldstein, H., & Spiegelhalter, D.J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion). *Journal of Royal Statistical Society A*, 159(3), 385–443.
- Gorard, S. (2000). Underachievement' is still an ugly word: Reconsidering the relative effectiveness of schools in England and Wales. *Journal of Education Policy*, 15(5), 559–573.
- Gorard, S. (2010). 'Serious doubts about school effectiveness', *British Educational Research Journal*, 36(5), 735-766. <http://dx.doi.org/10.1080/01411920903144251>
- Gorard, S. (2013). What difference do teachers make? A consideration of the wider outcomes of schooling. *Irish Educational Studies* 32(1): 69-82.
- Gorard, S., Hordosy, R. & Siddiqui, N. (2013). How stable are 'school effects' assessed by a value-added technique? *International Education Studies* 6(1): 1-9.
- International Summit on the Teaching Profession (2016). *Implementing highly effective teacher policy and practice*. Retrieved from <http://www.istp2015.org/>
- Haertel, E. (2013). *Reliability and validity of inferences about teachers based on student test scores* (Angoff lecture). Princeton, NJ: Educational Testing Service.
- Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review*, 61(2), 280-288.
- Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *The Journal of Human Resources*, 14(3), 351-388.

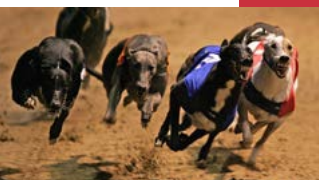


- Hanushek, E.A. (2011). The economic value of higher teacher quality. *Economics of Education Review* 30, 466–479.
- Hanushek, E. A., & S. G. Rivkin. (2010). Generalizations about Using Value-Added Measures Teacher Quality. *American Economic Review*, 100(2): 267–71.
- Harris, D.N. (2011). *Value-added measures in education: what every educator needs to know*. Cambridge, MA: Harvard Education.
- Hogan, A., Sellar, S., & Lingard, B. (2015) Commercialising comparison: Pearson puts the TLC in soft capitalism. *Journal of Education Policy*, doi:10.1080/02680939.2015.1112922
- Holloway-Libell, J. (2015). Evidence of Grade and Subject-Level Bias in Value-Added Measures. *Teachers College Record*.
- Holloway-Libell, J., & Amrein-Beardsley, A. (2015). “Truths” Devoid of Empirical Proof: Underlying Assumptions Surrounding Value-Added Models in Teacher Evaluation. *Teachers College Record*.
- Holloway-Libell, J., & Collins, C. (2014). VAM-based teacher evaluation policies: Ideological foundations, policy mechanisms, and implications. *InterActions: UCLA Journal of Education and Information Studies*, 10(1).
- Hood, C. (2011). *The Blame Game: Spin, Bureaucracy, and Self-Preservation in Government*. Princeton, New Jersey: Princeton University.
- Hoyle, R., & Robinson, J. (2003). League tables and school effectiveness: a mathematical model, *Proceedings of the Royal Society of London B*, 270, 113–199. <http://dx.doi.org/10.1098/rspb.2002.2223>
- ILO (2012). *Handbook of good human resource practices in the teaching profession*. Geneva: ILO. http://www.ilo.org/wcmsp5/groups/public/---ed_dialogue/---sector/documents/publication/wcms_187793.pdf
- Isore, M. (2009), *Teacher Evaluation: Current Practices in OECD Countries and a Literature Review*. OECD Working paper 23, Paris, OECD
- Jiang, J., Spote, S.E., & Luppescu, S. (2015). Teacher Perspectives on Evaluation Reform: Chicago’s REACH Students. *Educational Researcher*, 44(2), 105–116
- Johnson, S.M. (2015). Will VAMS Reinforce the Walls of the Egg-crate School? *Educational Researcher*, 44(2), 117–126.
- Junemann, C. & Ball, S. (2015). *Pearson and PALF: The Mutating Giant*. Retrieved from http://download.ei-ie.org/Docs/WebDepot/ei_palf_publication_web_pages_01072015.pdf
- Lamprianou, I. (2009). Comparability of examination standards between subjects: an international perspective, *Oxford Review of Education*, 35(2), 205–226. <http://dx.doi.org/10.1080/03054980802649360>
- Lazear, E.P. (2000). Economic Imperialism. *The Quarterly Journal of Economics*, 115(1), 99–146.
- Leckie, G., & Goldstein, H. (2009). *The limitations of using school league tables to inform school choice*, Working Paper 09/208, Bristol: Centre for Market and Public Organisation
- Lubienski, S., & Lubienski, C. (2006). School sector and academic achievement: a multi-level analysis of NAEP Mathematics data, *American Educational Research Journal*, 43(4), 651–698. <http://dx.doi.org/10.3102/00028312043004651>
- MacBeath, J. (2012), *Future of Teaching Profession*. Education International Research Institute and University of Cambridge Faculty of Education. Brussels: Education International.



- Mann, H. (1848). *Twelfth Annual Report to the Secretary of the Massachusetts State Board of Education*. Retrieved from: http://americanhistory.unomaha.edu/Source_Doc/Horace%20Mann%20Twelfth%20Report%20to%20MA%20Board%20of%20Education%201848.rtf
- McNeely, C. L. (1995). Prescribing national education policies: The role of international organizations. *Comparative education review*, 483-507.
- Measures of Effective Teaching (2013). *Feedback for Better Teaching: Nine Principles for Using Measures of Effective Teaching*. http://collegeready.gatesfoundation.org/wp-content/uploads/2015/05/MET_Feedback-for-Better-Teaching_Principles-Paper.pdf
- Meyer, R.H. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review*, 16(3), 283-301.
- Micklewright, J., Jerrim, J., Vignoles, A., Jenkins, A., Allen, R., Ilie, S., Bellarbre, E., Barrera, F., & Hein, C. (2014). *Teachers in England's Secondary Schools: Evidence from TALIS 2013*. Department for Education. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/322910/RR302_-_TALIS_report_NC.pdf
- Ministeriet for Børn, Undervisning og Ligestilling (2016). *Socioøkonomisk reference for grundskolekarakterer*. Retrieved from <http://www.uvm.dk/Service/Statistik/Statistik-om-folkeskolen-og-frie-skoler/Statistik-om-elever-i-folkeskolen-og-frie-skoler/Socioøkonomisk-reference-for-grundskolekarakterer>
- Mourshed, M., C. Chijioko, & M. Barbjder (2010). *How the world's most improved school systems keep getting better*. McKinsey & Company.
- Normand, R. (2008) School effectiveness or the horizon of the world as a laboratory, *British Journal of Sociology of Education*, 29(6), 665-676.
- Nolan, B., Salverda, W., Checchi, D., Marx, I., McKnight, A., Tóth, I.G., & Werfhorst, H.G. (2014). *Changing Inequalities and Societal Impacts in Rich Countries: Thirty Countries' Experiences*. Oxford University.
- Nusche, D., Halász, G., Looney, J., Santiago, P., & Shewbridge, C. (2011). *Sweden. OECD Reviews of Evaluation and Assessment in Education*. Retrieved from <http://www.oecd.org/sweden/47169533.pdf>
- OECD (2005). *Teachers Matter: Attracting, Developing and Retaining Effective Teachers, Education and Training Policy* [online]. Retrieved at <http://www.oecd.org/education/preschoolandschool/attractingdevelopingandretainingeffectiveteachers-finalreportteachersmatter.htm>
- OECD (2008). *Measuring improvements in learning outcomes: Best practices to assess the value-added of schools*. Paris: OECD.
- OECD (2009). *Evaluating and Rewarding the Quality of Teachers: International Practices*. <http://www.oecd.org/edu/school/evaluatingandrewardingthequalityofteachersinternationalpractices.htm>
- OECD (2011). *Establishing a Framework for Evaluation and Teacher Incentives: Considerations for Mexico*. <http://www.oecd.org/edu/school/establishingframeworkforevaluationandteacherincentivesconsiderationsformexico.htm>
- OECD (2013a). *Synergies for Better Learning: An International Perspective on Evaluation and Assessment*. OECD Reviews of Evaluation and Assessment in Education. Paris: OECD.
- OECD (2013b). *Teachers for the 21st Century: Using Evaluation to Improve Teaching*, OECD Publishing.
- OECD (2014). *TALIS 2013 Results: An international perspective on teaching and learning*. Retrieved from http://www.oecd-ilibrary.org/education/talis-2013-results_9789264196261-en
- OECD (2016). *Calidad Educativa/Quality Education*. <http://www.oecd.org/edu/school/calidadeducativaqualityeducation.htm>





- Pearson (2012). *The Learning Curve: Lessons in country performance in education*. London: Pearson.
- Pearson (2014). *The Learning Curve: Education and skills for life*. London: Pearson.
- Pierce, C. (2013). *Education in the Age of Biocapitalism: Optimizing educational life for a flat world*. New York: PalgraveMacmillan.
- Piketty, T. (2014). *Capital in the twenty-first century*. Harvard University.
- Robertson, S.L. (2012). Placing Teachers in Global Governance Agendas. *Comparative Education Review*, 56(4), 584-607.
- Ravitch, D. (2013). *Chetty-Friedman_Rockoff Non-Sense*. Posted on Diane Ravitch's blog: A site to discuss better education for all, April 17th, 2013. <http://dianeravitch.net/2013/04/17/chetty-friedman-rockoff-nonsense/>
- Ravitch, D. (2014). *Raj Chetty, VAM, Audrey Amrein-Beardsley, and Me*. Posted on Diane Ravitch's blog: A site to discuss better education for all, June 2, 2014. <http://dianeravitch.net/2014/06/02/raj-chetty-vam-audrey-amrein-beardsley-and-me/>
- Rittel, H.W.J., & Webber, M.M. (1973). Dilemmas in a General Theory of Planning. *Policy Sciences*, 4(2), 155-169.
- Robertson, S.L. (2000). *A Class Act: Changing Teachers' Work, The State, And Globalisation*. New York: RoutledgeFalmer.
- Rothstein, J. (2011). *Review of Learning about Teaching, by Thomas J. Kane and Steven Cantrell*. Boulder, Colorado: National Education Policy Center, University of Colorado. <http://nepc.colorado.edu/files/TTR-MET-Rothstein.pdf>
- Rothstein, J., & Mathis, W.J. (2013). *Review of two culminating reports from the MET Project*. Boulder, Colorado: National Education Policy Center, University of Colorado. <http://nepc.colorado.edu/files/ttr-final-met-rothstein.pdf>
- Rutkowski, D. J. (2007). Converging us softly: how intergovernmental organizations promote neoliberal educational policy. *Critical Studies in Education*, 48(2), 229-247.
- Rutter, M., Maughan, B., Mortimore, P., & Ouston, J. (1979). *Fifteen thousand hours: Secondary schools and their effects on children*. London: Open Books
- Sahlberg, P. (2011). Fourth Way of Finland. *Journal of Educational Change*, 12(2), 173-185.
- Schochet, P. Z., & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics*, 38(2), 142-171.
- Skolverket (2016), SALSA. Retrieved from <http://salsa.artisan.se/>
- Stevenson, H., & Wood, P. (2013). Markets, managerialism and teachers' work: the invisible hand of high stakes testing in England. *The International Education Journal: Comparative Perspectives*, 2013, 12(2), 42-61.
- Stevenson, H., & Carter, B. (2009). Teachers and the State: forming and reforming 'partnership'. *Journal of Educational Administration and History*, 41(4), 311-326.
- Tareen, S. (2012). Teacher evaluations at center of Chicago strike. *Huffington Post*, September 13, 2012. Retrieved from http://www.huffingtonpost.com/2012/09/13/teacher-evaluations-at-ce_0_n_1880264.html
- Taut, S., Valencia, E., Palacios, D., Santelices, M.V., Jiménez, D., & Manzi, J. (2014). Teacher performance and student learning: linking evidence from two national assessment programmes, *Assessment in Education: Principles, Policy & Practice*, DOI: 10.1080/0969594X.2014.961406



- Verger, A., & Curran, M. (2014). New public management as a global education policy: its adoption and re-contextualization in a Southern European setting. *Critical Studies in Education*, 55(3), 253-271.
- Wachter, R.M. (2016). How Measurement Fails Doctors and Teachers. *New York Times Sunday Review*, January 16, 2016. http://www.nytimes.com/2016/01/17/opinion/sunday/how-measurement-fails-doctors-and-teachers.html?_r=1
- Wilkinson, R., & Pickett, K. (2010). *The Spirit Level: Why Equality is Better for Everyone*. London: Penguin.
- Woessmann, L. (2011). Cross-country evidence on teacher performance pay. *Economics of Education Review*, 30(3), 404-418.
- World Bank (2011). *Learning for All: Investing in People's Knowledge and Skills to Promote Development. World Bank Group Education Strategy 2020*. Retrieved from http://siteresources.worldbank.org/EDUCATION/Resources/ESSU/Education_Strategy_4_12_2011.pdf
- World Bank (2012). *SABER: Systems Approach for Better Education Results. What Matters Most in Teacher Policies? A Framework for Building a More Effective Teaching Profession*. Retrieved from: http://siteresources.worldbank.org/EDUCATION/Resources/278200-1290520949227/SABER-Teachers-Framework-Updated_June14.2012.pdf
- World Bank (2015). *The Rise of Results-Based Financing in Education*. Retrieved from http://www.worldbank.org/content/dam/Worldbank/Brief/Education/RBF_ResultsBasedFinancing_v9_web.pdf

Comments and Suggestions

This is intended to be a discussion paper. If you have any questions about VAM, any comments or suggestions about how it works or might work in your context, or you just want to contribute to the debate please do not hesitate to get in touch.

Email us at: research@ei-ie.org



Education International Discussion Paper by Tore Bernt Sørensen

Value-added measurement or modelling (VAM)

#VAMboozled



Education International
Internationale de l'Éducation
Internacional de la Educación

Head Office

5, Bd du Roi Albert II
1210 Brussels, Belgium
Tel +32 2 224 06 11
Fax +32 2 224 06 06
headoffice@ei-ie.org
<http://www.ei-ie.org>

research@ei-ie.org

Tore Bernt Sørensen

Graduate School of Education,
University of Bristol
and EI research fellow

t.b.sorensen@bristol.ac.uk