Test-Based Teacher Evaluation

Over the past decade, few policies have generated more contentious debate than judging teachers’ performance using value added (VA), a type of complex statistical model designed to isolate teachers’ impact on their students’ testing progress. The American Educational Research Association calls the use of these models in teacher evaluations “one of education’s most controversial issues.”¹

As of 2015, all but a handful of states required that teachers be evaluated in part based on “student growth,” which, for most teachers in tested grades and subjects, refers to VA scores.² Just a decade earlier, virtually no states did so.

There is empirical support for trying out value-added and other types of growth models (henceforth simply “value-added”) as one component in multiple measure teacher evaluations. But details matter in policy. At the outset there was an almost complete lack of knowledge as to how VA estimates should be used in high stakes performance evaluations. This problem was compounded by the fact that VA had a credibility problem among educators.

In too many states, due in no small part to the “flexible constraints” of federal initiatives, and an overly narrow focus on the possibility of short term testing improvement via differentiation, VA policymaking was characterized by rigidity, imprudence, and haste.

The central argument of this chapter is that the manner in which VA was incorporated into evaluations made a difficult situation worse, and may have compromised the potentially useful role these estimates can play. The reasons why this occurred, moreover, offer important lessons for education reform going forward.
The Era of Teacher Quality

The body of research on value-added estimates, which now spans decades, has demonstrated that the effectiveness of teachers in raising test scores varies widely, with some teachers compelling very strong growth from their students while others’ students fall far behind.³

This core conclusion of the VA literature ran a direct collision course with descriptive research showing that teachers in several states and large districts were all receiving ratings of “satisfactory” or better, year after year.⁴ The juxtaposition of these two findings – that teachers’ measured effectiveness varied widely but evaluation ratings varied barely at all – provided the foundation upon which was built a sweeping effort to overhaul teacher evaluations for the vast majority of U.S. public school teachers.

Compelling this change was perhaps the signature education initiative of the Obama Administration, a competitive grant program called Race to the Top. States, most of which were suffering budget deficits due to a massive recession, could apply for federal RTTT funds by reforming several education policies, including teacher evaluation.

About two years after the announcement of RTTT came the second major cause of evaluation reform: USED’s program to grant states relief from the stringent requirements of the Elementary and Secondary Education Act (ESEA), the most recent iteration of which was No Child Left Behind (NCLB). In exchange for ESEA “flexibility,” or “ESEA waivers,” states, among other things, had to reform teacher evaluation systems though a great many had already done so as part of their RTTT efforts.
In the end, all but a handful of states enacted teacher evaluation reform, many via state legislation. The role of VA in this incredibly rapid, widespread policy shift is the topic of this chapter, but one quick caveat bears mention. That is, it is too early to draw anything beyond tentative conclusions about the impact of teacher evaluation reform, to say nothing of the unique impact of including VA estimates in those evaluations systems. The latter question will be particularly difficult to address empirically, even over the long term.

Moreover, insofar as any policy should in part be evaluated relative to its predecessor, it would be difficult to argue that the new evaluations are anything but an improvement over the old systems, which tended to be single measure systems that produced little or no variation in results. And, indeed, the evidence so far is encouraging.5

The purpose of this chapter, therefore, is to examine and derive lessons from the widespread incorporation of VA into evaluation systems, rather than proclaim success or failure for evaluation reform as a whole.

The Case for Value-Added

The body of research establishing the importance of teachers for short term student testing performance has rightly motivated increased attention to teacher quality in education policy. There is, however, a big difference between showing that teachers matter and actually improving the distribution of teacher quality. VA may have played a huge role in illustrating the potential benefits of improving teacher accountability, but it does not necessarily follow that it should play a role in accountability systems.
Critics of VA have made a strong case against using these estimates, one that has focused largely on their technical properties. It is worth noting that this case is based almost entirely on the value-added literature, which has provided arguably more knowledge about the properties of VA estimates than is available for any other common evaluation component.

A comprehensive review of the properties, strengths, and limitations of value-added models is beyond the scope of this chapter. The key point for our purposes here is that the limitations of VA are important and policy-relevant, but virtually all are shared by critics’ preferred alternatives, most notably classroom observations.

For example, it is well-established that VA scores are imprecise and unstable over time. Teachers can receive two rather different scores in two consecutive years, and this volatility is mostly a result of random error rather than “true” changes in performance. This is a real problem, but it is hardly unique to VA. In fact the available evidence suggests that classroom observation scores are similarly imprecise and unstable over time.

The same basic point applies to the argument that VA scores from some (but not all) models tend to be lower for teachers who teach more disadvantaged students. Once again, this is an important, policy-relevant property (though some of that difference may in fact be “real”). It is, however, shared by classroom observations.

Any evaluation component must be assessed vis-à-vis feasible alternatives. The “technical” case against VA certainly has implications for policy design, but, as a standalone argument for excluding these estimates entirely, it does not hold up particularly well under scrutiny.

So the available evidence does support, or at least does not preclude, trying out VA estimates as one component in multi-measure teacher accountability systems. From this
perspective, the key question is less about whether to use these estimates, and more about how to use them productively.

Value-Added’s Long and Unbeaten Path

From day one, two formidable challenges stood in the way of using VA in teacher evaluations. The first was inevitable, and requires little elaboration: there was very little evidence as to how these measures should be used. This includes crucial design issues such as model specification and how much VA scores should count vis-à-vis other measures.

The second and perhaps more serious obstacle was that many teachers didn’t trust VA estimates, and thus did not want to be judged by them. To be clear, teachers are far from a monolith, and their attitudes toward testing and accountability are nuanced. Yet the available evidence points to widespread resistance against using state test-based measures in personnel policies.

The issue of weighting gives a great example of the design complexities at play. All of these new teacher evaluation systems used VA estimates in concert with other measures of teacher performance. That might mean classroom observations, student surveys, or a host of other metrics. To evaluate a teacher, all of those indicators have to be combined into a score.

There is no “correct” weight for VA. There is a tradeoff. On the one hand, weighting the estimates more heavily will tend to generate greater differentiation in final ratings and ensure that a larger proportion of those ratings are based directly on student performance (at least to the degree tests are valid for this purpose). On the other hand, higher weights may crowd out the signal of other meaningful measures, and may also foster opposition to the new systems given the unpopularity of VA.
The Measures of Effective Teaching (MET) project, an extremely thorough and influential study of teacher evaluations released in 2012, recommended a weight for VA of 33-50 percent along with roughly equivalent additional weight allotted to classroom observations and student surveys. The empirical basis for this recommendation was analyses of how well different weighting schemes predicted the “overlap” between these three measures the next year. But the underlying implication here is that weighting is a judgment call: if you want to predict test-based effectiveness, weight VA more heavily; if you want to predict classroom practice, weight observations more heavily; and so on.

An informal review of the documentation for the states imposing a VA requirement reveals a dizzying and changing array of rules and recommendations for how districts could or should use these measures. For their teachers in tested grades and subjects, most states set a concrete weight specifically for VA or other growth models using state testing data, usually with an additional proportion reserved for alternative growth measures (the sum of these two percentages represented the “student growth” component).

There was variation in the specific weights set in states’ first iterations of evaluation reform. Of the 36 states that had, by 2015, imposed a requirement of using some sort of student growth component, at least 20 specified the minimum proportion that had to be based on VA per se (usually just for teachers in tested grades and subjects).

The percentages ranged from 20-25 percent in states such as North Carolina and Connecticut to 50 percent in Florida, Tennessee, and Washington, D.C. Virtually none set VA minimums below 20 percent.

In the remaining states, districts were required to use student growth measures as at least a “significant component,” sometimes specified as a percentage, but there was no explicit weight
specifically for VA. In a few cases, such as Wisconsin, this was because VA scores were not yet available statewide. In others, districts were granted varying degrees of leeway in how much they weighted VA versus alternative growth measures, such as student learning objectives (SLOs), which are usually locally-designed measures using local assessments. For instance, Indiana’s plan required that growth measures “significantly inform” teachers’ overall evaluation ratings but VA, when available, had to be the “primary” measure used for that purpose.

Virtually no states had VA estimates available but did not require districts to use them. One exception is Colorado, which allowed districts to use the state’s growth model for as little as one percent of teachers’ ratings, although the weight assigned to the student growth component was very high at 50 percent.

These technical issues played out against a backdrop of deep teacher skepticism of VA-based teacher evaluation. There are, somewhat surprisingly, relatively few representative national surveys that query teachers’ attitudes on the role of test-based measures in evaluations. One exception is a 2014 Gallup Poll which found that nine out of ten teachers had a negative view of linking their evaluations to Common Core testing results (note, however, the inclusion of Common Core in this question likely influenced responses).

The data are somewhat more abundant when it comes to attitudes toward using state testing outcomes in other personnel policies. For example, a 2009 survey by Public Agenda showed that 75 percent of teachers opposed basing teachers’ salaries, in part on students’ progress on state tests, while 60 percent opposed using state tests in tenure decisions. Similarly, 64 percent of teachers opposed the idea of financial incentives for teachers “whose kids routinely score higher than similar students on standardized tests,” according to a 2011 survey by Education Sector. This was an eight percentage point increase from 2007.
Underlying these attitudes toward test-based accountability is widespread skepticism toward state tests in general. A number of surveys show teachers’ unfavorable views of state tests as a measure of student learning\textsuperscript{17}, and opposition to the state testing requirement of NCLB.\textsuperscript{18}

Indeed, the very mention of standardized tests affects teachers’ opinions of policies. The Education Sector survey found that roughly half of teachers opposed the idea of performance-based pay based on students’ “academic progress,” but almost two thirds expressed disapproval of that policy when the question specified that progress would be measured using standardized tests.

Complicating the situation further was the fact that VA models have little if any formative value, and are far more complex than the simple proficiency rates to which educators had become accustomed under NCLB. Many teachers were understandably hesitant to have their performance judged, even in part, by an econometric model that very few understood and which provided no real information about how to improve. And the publication of teachers’ names and scores in the \textit{Los Angeles Times} and several major New York newspapers embodied teachers’ worst fears about the potential for the improper use of VA estimates.

Teachers’ attitudes about VA are important because the potential benefits of evaluation will stem not just from direct actions such as dismissals, but, perhaps more importantly, from \textit{voluntary} behavioral changes.\textsuperscript{19} This entails, for example, teachers changing their instructional practices based on evaluation results and feedback, identifying successful practices among their colleagues, and, of course, staying in the profession if they are successful and leaving if they are not.\textsuperscript{20} On the flip side, accountability systems, to the degree possible, should avoid incentivizing counterproductive behavior which includes so-called “teaching to the test,” switching to untested
grades or subjects or lower poverty schools (or being switched to tested at the expense of non-tested grades/subjects) or in rare cases, cheating.\textsuperscript{21}

Productive behavioral changes are less likely to occur if teachers do not “buy in” to the system in question, which will not happen if it employs measures that teachers do not consider valid or reliable. And any benefits of involuntary changes, such as dismissals, can be offset if teachers feel that they are unjustified.\textsuperscript{22} Teachers, like all workers, evaluate the risks and rewards of their jobs.

Right at the outset of the period of rapid evaluation reform, these two major issues – the scarcity of evidence regarding the proper use of VA and its credibility problem among teachers – made for a situation in which policymakers had to traverse a hostile environment without so much as a map. And the outcome of this process, insofar as the debate over VA tended to dominate the debate about evaluations in general, would play an outsized role in the entire evaluation reform enterprise.

Flexible Constraints and the Institutionalization of Value-Added

Tension between the federal government and states is hardly unusual in U.S. policymaking, but in the education arena, long the purview of states and localities, the waters are particularly rough. The Obama Administration, through RTTT and the ESEA waivers, attempted to navigate these waters by making evaluation reform voluntary, though one might argue that this voluntarism was somewhat nominal, given states’ budget crises at the time and their desire to avoid the consequences of failing to meet NCLB’s impossible goals.

In the end it was the districts that would have to do the hard work of evaluation reform. Yet it was states that designed evaluation plans, or at least set the requirements for districts’
designs, under RTTT and ESEA waivers, and only under the latter could districts choose not to participate.

To be sure, as discussed below, state policymakers, themselves operating within the framework of federal guidelines, varied in terms of how much flexibility they permitted for districts. A few states required a single system for all districts but most either let districts design systems that met certain specifications, or recommended a state model from which districts could opt out if they designed a comparable alternative.

In the context of VA, though, there was at least one detail of RTTT and ESEA waivers that made a huge difference in shaping the degree of flexibility passed down from states to districts, and that was the establishment of strong incentives to include a “significant” role for “student growth” measures in new evaluations systems. Given the paucity of alternative growth measures, the “significant role” guideline came fairly close to a *de facto* VA requirement.

As a result, all but 10 of the 46 states that engaged in evaluation reform, and the thousands of districts under their supervision, mandated using VA or some other type of state test-based growth model for teachers in tested grades and subjects (and in many cases, for non-tested teachers as well).²³

So, from a district perspective, when it came to using VA in teacher evaluations, what most were given was at best “flexible constraints.” They had to do something, but had options regarding how to do it.

The “significant role” recommendation was in no small part a direct response to the aforementioned failure of previous evaluations to differentiate meaningfully between teachers’ effectiveness. The systems in place at the time were a formality. VA models, on the other hand,
are guaranteed to produce a spread of different scores, since teachers are judged relative to each other.

This strong push for differentiation was in turn motivated by a strong belief among many policymakers and advocates that the ability to sort teachers by performance, and attaching stakes to those ratings, would by itself produce substantial testing gains. VA scores are designed, however imperfectly, to isolate teachers’ effects on testing growth, a proxy for student learning. Selecting and deselecting teachers based in part on this measure therefore offered substantial appeal, particularly given that both RTTT and ESEA waivers placed enormous emphasis on boosting short term testing results. Critics, in contrast, saw it as misguided testing obsession.

In any case, the federal push was indisputably successful in compelling evaluation reform, and, again, it is difficult to argue that the new systems are inferior to the old. In retrospect, however, the drive to ensure widespread inclusion of VA in these new systems, though defensible in theory, was not especially well executed in practice. Given how little was known about using VA in evaluations, as well as its credibility problem among educators, the design and implementation of VA’s role in the new systems should have emphasized flexibility, caution, and gradualism. There was a correct way to do this. It didn’t quite happen.

The Three Leverage Points

Policy choices typically represent tradeoffs, rather than clear right or wrong decisions. Policymakers at the state and local levels faced with the task of using VA, an unpopular, untested measure, had to adjudicate where to come down on these tradeoffs based on their views of what was best for their teachers and students.
In the context of incorporating VA into teacher evaluations, three such decisions were arguably most consequential: How much to count VA toward teachers’ final ratings; how to address the issue of imprecision; and implementation timelines. More than any others, these three decisions were the focal points determining the early impact of using VA in evaluations, including the crucial issue of how teachers would react to it.

The Weighting Game

In the final analysis, when it came to the role of VA, there was certainly a degree of flexibility for districts, at least in some places. On the whole, however, states’ policies tend either to require VA or have plans to do so in a future year, and usually at a relatively high weight of at least 20-25 percent.24

In some cases, the rigidity bordered on absurdity. In Maryland, for instance, several districts’ proposed designs were rejected by the state because they did not weight VA heavily enough. Included among these districts was Prince George’s County, which proposed using VA as 15 percent of final ratings. The state wanted 20 percent.25

The point here is not to say that lower weights were necessarily better, or higher weights necessarily worse. The point, rather, is that districts in most states should have had more freedom to make those decisions for themselves. Context matters. Such autonomy would have allowed districts to exercise their judgment based on local needs and situations, which can not only improve educator buy-in, but student outcomes as well.26

The focus instead seemed to be on identifying and dismissing ineffective teachers, which, policymakers were told, based mostly on misinterpretations of test-based simulations27, would generate rapid improvement. It was, in a sense, differentiation for differentiation’s sake.
Such a mindset was evident, for example, in New York. When the first round of evaluation results showed very few teachers receiving the lowest ratings, the state’s governor reacted by blaming principals for giving high observation scores, and putting forth a proposal to increase the weight of VA from 30 to 50 percent.28 This was a particularly clumsy, ill-considered reaction, but it reflects a more widespread view of VA as a means to sort teachers, rather than as a tool to compel productive changes in educators’ practice.

To be sure, the argument that evaluations would have been better had districts been given greater flexibility to choose their own weights is speculative. What is clear, however, is that the decision was not made by the people most in touch with how best to educate the particular children they serve. It was, in short, an arranged marriage.

The primary role in exercising this judgment should have rested at the local level, and there was a lot of room for innovation. Districts, for example, could have phased in VA slowly, starting at a very low weight (say, around 5-10 percent), with the stipulation that this would increase in future years, perhaps based on clear criteria, such as teacher and principal feedback or how the results turned out.

There were also innovative alternatives to a system in which each component is granted a specific weight, and final ratings calculated based on those weights. For instance, VA could have been used as a kind of noisy “screen,” identifying teachers for additional observation and intervention, while cross-checking VA against other measures, and vice-versa, in order to ensure observer quality and guard against misclassification due to imprecise VA estimates.29

Finally, of course, there might have been the option for districts not to use VA at all, perhaps with the stipulation that they demonstrate the components they did choose were
associated with meaningful outcomes and served to differentiate teachers. The argument that evaluations need VA to work is not consistent with the available research.\textsuperscript{30}

Such innovation, or even districts’ ability to consider it, was the exception rather than the rule. And the primary reason was the “significant role” guideline embedded in RTTT and ESEA waivers. This language discouraged any approach other than assigning concrete percentages, and was easily interpretable as precluding flexibility to set low weights, even with the plan to increase them.

\textit{The Unfought War on Error}

Among the primary reasons why VA was so controversial in evaluation reform was that opponents, somewhat ironically, seized on key findings from the VA literature, which also compelled the focus on teacher quality, in order to argue that VA scores are too error-prone to be useful.

These concerns were justified. As discussed above, due largely to small samples, VA estimates are imprecisely estimated in any given year and teachers’ scores therefore can vary quite a bit between years. This is perhaps the most significant “on paper” problem when using individual teachers’ VA scores in accountability systems, as it might penalize or reward teachers based largely on random error. One analysis of New York City teachers’ VA scores, which received a great deal of attention but is typical of many other studies’ results, found that almost one in three of the English teachers who were in the bottom VA quintile in 2007 ended up in one of the top two quintiles in 2008.\textsuperscript{31}

And this issue is not solely about misclassification or other “technical” concerns. Although, as mentioned above, the imprecision of VA is shared by classroom observations (or
any other performance measure worth its salt), VA’s mere reputation as inaccurate represents an obstacle to realizing its potential benefits in new evaluations. If teachers regard this component of their ratings as unreliable or mostly luck, the credibility of their entire evaluation is compromised.

Yet there are straightforward ways to at least partially address this serious issue. For example, states and districts can increase the sample size upon which teachers’ scores are based by using multiple years of data, or at least using them when available. In addition, unlike classroom observations, VA scores come with a margin of error, or a confidence interval, which systems can employ when converting scores into “evaluation points.” Both steps would mitigate the problem of teachers being classified as high or low performing due to random error, and could also go a long way toward easing teachers’ concerns.

One simulation, for example, estimated the proportion of teachers who would be either erroneously identified as high or low performing, or erroneously identified as average, at around 35 percent using one year of data, but this error rates was reduced almost 30 percent using three years of data.32

The decisions about whether and how to address error directly were made largely at the state level, and some states took a proactive approach. In Tennessee, for instance, confidence intervals inform the incorporation of VA scores into evaluation scores, while Colorado provides the intervals and encourages districts to use them. A few states such as North Carolina, Colorado, and Oklahoma, use or strongly encourage multi-year rolling averages of VA scores to boost precision. And, finally, the models used in a couple of states, such as D.C., employ statistical techniques for adjusting (“shrinking”) VA estimates by sample size, which mitigates the problem of random error.
The benefits of these types of approaches, to be clear, entail costs. For instance, requiring multiple years of data means that many educators, such as first year teachers, cannot receive scores in any given year, while employing confidence intervals would reduce the ability of VA estimates to differentiate between teachers since so many of their scores would be interpreted as average even though, statistically, their “true performance” is more likely to be closer to the estimate than to the bounds of the confidence interval.

Given the severity of the imprecision problem, however, as well as teachers’ grave concerns about it, there was a strong case here for accepting these costs, and districts should have been encouraged to do so. Despite recommendations from numerous scholars and organizations, such attempts were made too infrequently, or at least ineffectively. For the vast majority of teachers whose evaluation ratings are based in part on VA, scores are based on a single year of data. Sample size requirements are set at a bare minimum, if at all. Error margins are not factored into ratings, and in some cases they are not even reported.

It has not gone unnoticed. One New York teacher filed a (successful) lawsuit against the state’s evaluation system, because her VA score had fluctuated wildly between two years.\(^{33}\) Pushback specifically against the imprecision of VA has been a major factor in opposition to evaluations in states such as Massachusetts\(^{34}\) and Georgia\(^{35}\), as well as the 2012 teacher strike in Chicago.\(^{36}\)

So, most new systems include VA, a measure that has caused a great deal of controversy due to its imprecision, and yet these systems are ignoring readily available information that can be used to address that issue. They prioritize differentiation and selection over caution and proper inference.
It is, of course, implausible to assert that using this information would have prevented the opposition. At the very least, though, given the pre-existing hostility toward VA and the lack of evidence about how to use it, taking steps to address error would have been good policy. And more districts should have had the option or encouragement to make this choice.

*The Great Value-Added Stampede*

The burden of teacher evaluation reform inevitably fell on districts which had to implement (and in some cases design) these new systems with new measures, while also going about about their usual business of educating children. Given the complexity of this set of tasks, this was no time for rushing.

States’ timelines, as usual in evaluation reform, varied a great deal, but speed was the most common approach. Most of the 36 states that employed VA or other growth models implemented evaluations without VA, and installed it 1-3 years later (usually due to logistical requirements, rather than a deliberate effort to proceed gradually), and several states had to extend their original timelines.

Even counting these delays, around 20 of the 36 states allowed five or fewer years from the time evaluations were legislated or otherwise required to the time they were implemented with VA included (in almost all cases, with stakes attached).  

All but 10 of these states allowed three or fewer years for design and piloting. Seven did not even pilot their new systems, 11 piloted for only one year, and only a handful of states’ pilots lasted longer than two years. A few states, such as Michigan and Alaska, did choose to “phase in” VA by increasing its weight over time, but it was typically from 25-30 to 40-50 percent rather than gradual phasing starting at a low weight.
Over the span of just a few years, value-added went from an obscure measure familiar only to researchers, to a significant component of performance evaluations for hundreds of thousands of U.S. public school teachers. And, just a few years later, it was the rule rather than the exception.

Making the situation worse, the adoption of the Common Core State Standards was occurring at roughly the same time as most states were in some stage of reforming evaluations. This was not only a major strain for district capacity, but it presented a unique challenge for VA, since states were also phasing in new Core-aligned assessments (which would be used to calculate VA scores). Many states, to their credit, chose to suspend the use of VA while new assessments were administered, but at that point, there was no good choice. The delicate rollout of VA in evaluations either had to push ahead recklessly, despite new tests, or stop after having started, causing yet more hardship and confusion. This unfortunate situation was perfectly foreseeable, and could have been avoided with longer timelines.

Such haste, once again, was motivated in part by federal guidelines. RTTT’s process rewarded states that already had evaluation reform in progress prior to their submitting applications. In other words, states that could promise new evaluations in place more quickly stood a better chance of winning. For example, the first two winners of RTTT grants, Tennessee and Delaware, promised quick implementation and immediate results. Tennessee named its program “First to the Top.”

Other states got the message: get systems up and running as quickly as possible, in order to boost student achievement. ESEA waivers pushed this urgency by requiring states granted “flexibility” to have new evaluation systems in place (including stakes) by 2015-16.
To be sure, some districts had the capacity to meet these deadlines, but there is ample evidence that the timelines established by states were unrealistic for many others. In addition to copious reports about individual districts struggling to get their new systems up and running, a 2013 GAO report found that most states awarded RTTT grants failed to meet their timelines for implementing evaluation reform\textsuperscript{38}, while at least 12 states granted ESEA waivers requested additional time.\textsuperscript{39}

This rush also had consequences for the design of new systems. For example, districts in many states had to decide how to weight each component before receiving a single year of data showing how these measures turned out. Others lacking the time or capacity to develop alternative growth measures in time, chose to use schoolwide VA, thus holding some teachers, particularly those in non-tested grades and subjects, accountable for the performance of students they never taught. A series of lawsuits in states such as Florida, Tennessee, and Texas have challenged this requirement.

In short, federal incentives/guidelines and the push for rapid achievement gains compelled many states to rush ahead with evaluation reform. It was too much too soon. This was a risky course of action regardless, but particularly so given that the new systems had to include untested, unpopular new measures. The potential of VA depended on using it correctly, and building its credibility among teachers. In this context, patience was a virtue, as was acknowledgment of variation in districts’ capacity to accomplish this type of reform.

The Early Results

As stated above, there is no real empirical evidence regarding the impact of using VA specifically in evaluations on teacher or student outcomes. Indeed, the evidence on the impact of
evaluations as a whole is just beginning to emerge. Moreover, given the constraints imposed on districts in how they were to incorporate VA, there may not be enough within-state variation to assess the impact.

When it comes to VA, however, the signs thus far are worrisome. The survey evidence cited above indicates that teachers remain overwhelmingly opposed to using VA in evaluations and other personnel policies, and that their opposition to standardized testing regimes in general may even have increased somewhat. Furthermore, a 2016 study by the Institute for Education Sciences found that teachers whose evaluations included test-based measures were less satisfied with their systems than their colleagues who were evaluated based on non-test measures.40

Lawsuits challenging evaluations have been filed in New York, Tennessee, Texas, and New Mexico, all focusing on the VA component. And the pushback to VA in evaluations no doubt played some role in the current “opt out” movement, which encourages parents to refuse to allow their children to take state tests.

Several states including Tennessee and D.C., whose evaluation systems received nationwide attention as exemplars, have rolled back the weights they assign to VA. Instead of slowly increasing, the weights are decreasing.

Finally, and most recently, the reauthorization of ESEA in the form of the Every Student Succeeds Act (ESSA) severed the link between the federal government and states on teacher evaluations. A few states, such as Alaska and Georgia, have already removed the student growth requirement from their evaluation laws. It remains to be seen how many others will follow suit.

And, in a final irony, it seems that even the priority on differentiation failed to work out as expected. A recent analysis of the results in 24 states with new evaluation systems found that
the proportion of teachers receiving the lowest rating in virtually all of these states was less than one percent. As a result of the push to install VA, there may have been insufficient attention to other measures, most notably classroom observations.

The environment for incorporating VA into teacher evaluations was hostile from the start and so the flexibility, caution, and patience recommended above certainly would not have prevented setbacks. But states, fueled by federal constraints and an overemphasis on differentiation and selection, likely made a difficult situation worse. The role of VA is under siege in many places despite its potential to be one useful component in teacher evaluation systems.

Lessons

The story of VA in teacher evaluations, at least thus far, offers at least three important lessons for policymakers and advocates.

*Even seemingly minor attempts to limit district autonomy in education reform, or to embed flexibility in constraints, carry great risks.* The architects of RTTT and ESEA waivers focused a great deal of the importance of flexibility for states and districts designing new evaluation systems. These promises were not all double talk. States indeed had some leeway in the design and implementation of new teacher evaluation systems, and they passed much of that down to districts.

When it came to the role of VA in evaluations, however, the flexibility was in important respects just nominal, particularly in ESEA waivers, which unlike RTTT, was not optional for districts. The language recommending or requiring growth measures as a “significant component” was but a miniscule part of these federal initiatives, but it ended up carrying
enormous consequences for evaluation reform. It established, in essence, a *de facto* chain of command from the federal government down to states and then to districts, and put a rather hefty finger on the scale when it came to crucial decisions about policy design and implementation. This may have compromised the strong potential value of VA in teacher evaluations.

The lesson here, put simply, is that federal efforts to compel policy change at the state and district levels should think very carefully about a “one size fits all” approach *even for seemingly small details*. Had the federal government taken a more hands off approach toward VA, the decisions discussed above would still have been necessary, but they would have been in the hands of states and, preferably, their districts.

Moreover, the “flexible constraints” approach, in which the big priorities are largely pre-determined but the specifics of how to accomplish them are less prescribed, is not always an effective compromise. The details can threaten the larger goals. NCLB, for example, was quite prescriptive in terms of the testing goals states had to set, but gave them some freedom on details such as subgroup selection, sample size, and the treatment of mobile students. Yet it was these seemingly arcane details, rather than actual student achievement, that drove much of the variation in outcomes.\(^{42}\)

Similarly, requiring states and districts to use VA (or other growth measures) as a “significant component” but letting them decide how to define “significant” forced a great many districts to use measures they didn’t want to use, and may have compelled others to use them in an ill-considered manner.

*Don’t let measurement and high stakes decisions overwhelm attention to the voluntary behavior of stakeholders.* Until recently, teacher personnel policy focused largely on inputs, such as teacher education and experience. This has changed dramatically over the past 15-20 years, in
no small part due to the wide availability of annual testing data, and the enormous research opportunities this has fostered, including that represented by the VA literature.

The idea of transitioning VA from a policy analysis tool to a direct accountability measure for individual teachers was based in large part on this body of research, and VA was (rightly) viewed as one potentially useful means of differentiating teachers based on student performance, an outcome about which we all care (or at least an imperfect proxy for such an outcome). But the haste and rigidity of evaluation reform in many places betrayed a mindset in which sorting and firing teachers was the priority and key to improvement, often to the detriment of how the new systems would influence the behavior of stakeholders.

Finding this balance is a recurring theme in school accountability systems, where NCLB and state-designed systems have for years focused on assigning performance designations to schools, usually based on testing outcomes. But the ratings themselves accomplish very little. What matters instead is how teachers and administrators respond to them, and these responses can be desirable (e.g., increasing instructional time) and undesirable (e.g., gaming). So, while these schoolwide systems have been shown to have a modest positive impact on short term testing outcomes, there is mixed evidence as to why this improvement has occurred.43 From this perspective, credibility of and buy-in to systems is no less important than differentiation and high stakes decisions.

Some advocates of VA scoff at the idea of downplaying the role of these measures simply because teachers don’t trust them. This is understandable. No policy will enjoy universal support, and it is not unusual for highly controversial policies to end up working. Moreover, if there are no rewards or consequences for teacher evaluations, teachers may simply ignore them completely.
That said, such positive effects are less likely – perhaps unlikely – if policies are rushed or forced on an unwilling workforce. The recent expansion of the role of measurement in education policy is a good thing, but the assumption that simply getting policies installed will shape the teacher workforce risks under-acknowledging the fact that it is actually teachers that shape policies’ impact.

Informal test-based accountability can work against formal test-based accountability. Co-existing with test-based accountability systems for schools and teachers is a vast and complex system of informal accountability, in which individuals and education institutions are held accountable for testing results not by state and federal governments, but rather by the public.

For instance, the education policies of governors and mayors, as well as the reputations of superintendents and principals, depend largely, often entirely, on short term testing outcomes. If test results are positive, virtually any other outcome or issue takes a backseat. Indeed, terms such as “effective” and “high performing” almost always refer to test-based impact. Policies “work” if they boost scores.

This context of informal accountability has a meaningful influence on formal accountability policy, and vice-versa. There is, for instance, a strong, almost irresistible incentive for policymakers to do anything to boost test scores, and to do it quickly. This was, as we’ve seen, a big part of the rationale for the growth measure requirements under RTTT and ESEA waivers, as well as, perhaps, the haste with which evaluation reform proceeded. Select and deselect people and institutions based on test scores, and test scores will increase.

Similarly, in their applications for ESEA waivers, states were allowed to abandon the ridiculous 100 percent proficiency requirements of NCLB, but virtually all replaced those with
“ambitious” goals that remained unattainably high. Anything less would have risked informal accountability sanctions, such as public accusations of “low expectations.”

To be clear, using test-based productivity measures in formal accountability is not some kind of misguided conspiracy to reap political benefits. Done correctly, it can be good policy. But it can also go too far.

And one reason it goes too far is the incentive structure of informal accountability, in which rapid short term testing gains are the coin of the realm. This is incompatible with the fact that good policy takes time to design and implement, and also to have the desired impact, which may not always be apparent in short term testing results.

For instance, experiments with performance-based compensation in the U.S. indicate that providing bonuses based on their students’ testing results did not have an appreciable impact on teachers’ test-based effectiveness. On the one hand, this illustrates that simply identifying and rewarding “effective” teachers did not serve to compel greater effort, which is important. Yet these experiments led to a widespread belief, based solely on short term test effects, that performance pay does not work. This ignored the possibility that changes in compensation systems could have longer term impacts, such as improving recruitment and differential retention.

So long as the reputations of people and institutions depend so heavily on short term testing results, test-based accountability policies, as well as other policies, are at risk of incomplete, misguided design, rushed implementation, and short sighted evaluation. Put simply, test-based accountability can be among its own worst enemies.

11 “Gathering Feedback for Teaching.”


Steinberg and Donaldsen, “The New Educational Accountability.”

Ibid.


Taylor and Tyler, “The Effect of Evaluation on Teacher Performance.”


