Making Teachers Accountable

Making schools and teachers more accountable for results, especially student learning outcomes, has become a central challenge for education policy makers in both developed and developing countries. The quantity and variety of policy innovations in this area has increased significantly over the past decade, with an especially striking increase in the developing world. This chapter reviews both the theory and the evidence base around two key types of reform focused on teacher accountability: contract tenure reforms and pay-for-performance reforms.

The first section summarizes the theoretical and empirical rationales for teacher accountability reforms. The second section reviews recent global experience with these reforms. The third and fourth sections put forth a typology of contract tenure and pay-for-performance approaches being adopted in developing countries and review the evaluation evidence around each type of reform. The fifth section compares the empirical evidence with the theoretical literature on performance incentives to identify key design issues. The final section draws cautious conclusions from existing evidence on how to design effective incentives for better teaching.

Throughout the chapter, the developing country experience and research literature are surveyed broadly, but emphasis is placed on recent evidence from well-evaluated reforms in the developing world.

Teacher Accountability Reforms: Why?

Growing interest in teacher accountability reforms stems from a confluence of factors. First, an increasing number of countries are convinced that
student learning outcomes are the core barometer of education system performance. In developing countries, the use of nationally standardized tests has increased substantially over the past 10 years, as has the number of developing countries participating in internationally benchmarked assessments such as the Organisation for Economic Co-operation and Development’s (OECD’s) Programme for International Student Assessment (PISA). The latest wave of education research on long-term correlations between (internationally benchmarked) student learning levels and gross domestic product (GDP) growth supports this focus. Studies conclude that education investments contribute to faster GDP growth only if schooling is effective in raising student learning—and the higher the learning levels, the faster the growth (Hanushek and Woessmann 2007). In short, an increase in the quantity and quality of student learning data in developing countries has created a capacity that did not exist earlier to monitor and potentially reward school-level learning improvements.

Second, there is increasing evidence that teachers’ ability to generate student learning is highly variable. Recent careful studies of the “value added” of individual teachers working in the same grade in the same school have begun to document that while students with a weak teacher may master 50 percent or less of the curriculum for that grade over a single school year, students with a good teacher can get an average gain of one year, and students with great teachers may advance 1.5 grade levels or more (Hanushek and Rivkin 2010; Farr 2010). A series of great or bad teachers over several years compounds these effects and can lead to unbridgeable gaps in student learning levels. The most rigorous research to date in this area is from the United States, and it may not proxy the variance in teacher effectiveness in other settings. But new research in Brazil, discussed later in this chapter, suggests that teachers’ classroom performance and effectiveness spans a large spectrum in at least some developing country settings as well.

Third, education policy makers wishing to recruit or groom “great teachers” to raise overall learning results confront the empirical reality of recruitment and compensation systems with weak links, if any, between rewards and performance. The vast majority of education systems are characterized by fixed salary schedules, lifetime job tenure, and flat labor hierarchies, which create rigid labor environments where extra effort, innovation, and good results are not rewarded. Nor is it possible to sanction poor performance; the percentage of tenured teachers ever dismissed for poor performance is exceedingly small (Weisberg and others 2009).

Almost universally, teacher recruitment and promotion are based on the number of years of preservice training, formal certificates, and years in service. Yet an extensive body of research has documented the lack of correlation between these “observable” factors and teachers’ actual effectiveness
in the classroom, measured by their ability to produce learning improvement in their students (Hanushek and others 2005). The record on in-service professional development leading to measurable improvements in teacher performance is also strikingly poor (Borko 2004; Garet and others 2001; Cohen and Hill 2001). The clear implication of available research is that most school systems are recruiting and rewarding teachers for the wrong things, failing to encourage the capacities and behaviors that contribute most directly to student learning results, and unable to sanction ineffective performance.

A disconnect between the incentives teachers face and the results school systems seek is manifest in many developing countries by egregious performance failures. A study of teacher absenteeism across six different developing countries in 2004 found that, on average, 20 percent of the teaching force was absent on any given day, suggesting low accountability for attendance and performance (Chaudhury and others 2005). Classroom observations in representative samples of schools in Latin America in 2009 found more than 30 percent of instructional time lost because teachers arrived late, left early, or otherwise failed to engage in teaching (Bruns, Evans, and Luque 2010). PISA 2006 data show that teacher applicants in most developing countries are drawn from the weakest students in secondary and higher education, even as the cognitive content of basic education becomes more demanding. Above all, average learning outcomes in developing countries are low and, in most countries, have failed to improve.

Developing countries today spend an average of 5 percent of GDP on education, and many countries are on track to increase this. The impact of this investment on their subsequent economic growth hangs largely on how effectively they use the 4 percent of GDP (80 percent of total education spending) that goes to pay teachers. In a growing number of countries, the drive to improve student learning outcomes is translating into creative and sometimes radical policy reforms aimed at changing the incentives for teachers. While the share of reforms being rigorously evaluated remains small, the evidence base to guide the design of new efforts is becoming more robust. This chapter aims to distill that evidence and its implications for policy.

**Recent Global Experience with Teacher Accountability Reforms**

As Vegas (2005) and others have pointed out, individuals are attracted into the teaching profession and gain satisfaction from their work for a wide range of reasons. All of these factors, shown in figure 4.1, constitute part of the *incentives* for teaching. Correspondingly, school systems have numerous
monetary and nonmonetary levers to use in attracting and rewarding effective teachers.

Beyond the policies that touch teachers directly, broader education policies that shape the school environment also affect whether high-capacity teachers are attracted into a school system and motivated to perform. It is relatively common for private schools and charter schools to recruit teachers with qualifications equivalent to those of public school teachers while paying lower salaries or offering less job security. Teacher surveys report that attractive features of the school environment—including the quality of infrastructure, class size, availability of teaching materials, quality of principals, engagement of parents, time for collective work with other teachers, and opportunities for professional growth—influence employment decisions and can offset lower wage and benefit compensation.

This chapter zeroes in on two of the incentives pictured in figure 4.1: job stability (that is, teacher contract tenure) and bonus pay or pay-for-performance programs.

We differentiate bonus pay from other policies affecting salary differentials in that bonus programs generally leave base salaries and the salary scale intact but create incentives at the margin, with the offer of an annual (or monthly) bonus based on some measure of teacher performance. The performance measure may be an input measure, such as teacher attendance; an outcome measure, such as school or student results;
or a combination of the two. The distinguishing feature of pay for performance is that it rewards teachers for what they do or achieve during a specified period (typically the prior school year). In this way, pay for performance is conceptually distinct from the factors that generally determine other teacher salary differentials, such as teachers’ qualifications, geographic posting, type of service (that is, hardship pay for teaching in certain schools or disciplines), or even tests of skills and capacity—all of which reward teachers for what they are capable of doing.

In contrast to most of the other incentives for teachers, both contracts without guaranteed tenure and pay-for-performance programs establish direct links between teachers’ performance and their rewards or sanctions. Thus, contract tenure and pay-for-performance reforms are potentially two of the strongest instruments at the disposal of education policy makers to increase teachers’ accountability for results. These reforms are also of special policy interest for several other reasons:

- **They are important.** While there is evidence that intrinsic motivation plays a stronger role in the teaching profession than in many other occupations, there is also evidence that compensation and contract status are key determinants of who goes into teaching, how long they remain, and how they perform (Chapman, Snyder, and Burchfield 1993; Guarino, Santibañez, and Daley 2006; Rivkin, Hanushek, and Kain 2005; Murnane and others 1991).

- **They are expensive.** Teacher contracting and pay policies are important drivers of the overall teacher wage bill, which is by far the largest component of education spending. As such, these policies have ripple effects on the resources available for all other education investments. In most developing countries, the teacher wage bill is also a large enough share of public spending to have implications for overall fiscal policy.

- **They present major challenges for policy makers in most countries.** Historically rigid policies tying teacher recruitment, tenure, and compensation to formal certification processes and seniority typically leave education officials with limited room to maneuver to either “de-select” teachers who are not effective or reward high performers. While there may be a long-term need in many countries to adjust the base wage and salary scale for teachers relative to other categories of public or private sector employment, policy makers struggle with the recognition that the impact of across-the-board increases is fiscally prohibitive and may still fail to create stronger incentives for performance.

These factors are inspiring increased experimentation in developing countries with reforms that use the levers of contract tenure and bonus pay to try to make teachers more accountable for performance. Of the two, pay
for performance has the largest number of experiences under way that are being subjected to rigorous impact evaluation, in both developing and developed countries. However, both strategies are increasingly being used in both low- and middle-income developing countries. The next two sections review the leading developing-country experiences and emerging evidence base for each type of reform.

**Contract Tenure Reforms**

Alternative contracting, in its broadest form, means recruiting teachers on contracts that do not grant the civil service status and tenure protection offered to regular teachers. Contract tenure reforms thus can overcome the rigidity of existing teacher policies by establishing a parallel teacher corps or career stream with different rules of the game alongside the existing teacher stream.

Although fairly uncommon in most European countries, alternative contracting is used in Australia, the United Kingdom, the United States, some other OECD countries, and most developing countries. Alternative contracts are typically for one year and renewable based on performance. Alternative contracting in developing-country settings is often—but not always—associated with entry standards and pay levels that are lower than for regular civil service teachers. In OECD contexts, such contracting is often used to bring highly qualified mid-career professionals from other fields into teaching at comparable salaries.

The rationale for alternative contracting in the developing world is usually either a shortage of teachers who meet the civil service entry standards or a shortage of public sector funding for the more expensive civil service contracts. Tightening teacher accountability for performance is rarely the explicit rationale. However, being on a short-term, renewable contract can clearly generate stronger incentives for an individual teacher to meet the performance goals of the contracting unit. The share of civil service teachers dismissed on performance grounds is extremely low in most systems, and these dismissals are almost always for egregious abuses, not simply for poor effectiveness in producing student learning outcomes. Good data on turnover of contract teachers unfortunately do not exist, but in the most carefully studied environments turnover has been substantially higher—in the range of 20–40 percent annually. However, there are no data on the share of contract teacher exits strictly related to performance, and many other factors are also typically at play.

The core policy innovation is the creation of a parallel teacher cadre hired on short-term contracts, typically covering a single school year and renewable based on performance. In a large share of cases, the hiring of
contract teachers is also devolved to a lower level of the education system—the school or village level—than the hiring of civil service teachers, who are typically hired at the national level. This creates potential for more effective monitoring of teacher performance by the contracting unit.

Despite this potentially tight link between contract renewal and teacher performance, goals or targets in such contracts are rarely precise or explicit. There are few developing-country cases of explicit links between contract teachers’ renewal and student learning improvements. If specific metrics are mentioned at all, they are more commonly teacher attendance or community satisfaction measurements, defined fairly broadly and somewhat subjectively. Student learning progress, however, may be an implicit factor in school managers’ or community satisfaction with a contract teacher’s performance. Table 4.1 summarizes the evidence base from rigorously evaluated cases in the developing world.

Confounding analysis of the pure accountability effects of less-secure contract terms and possibly closer performance monitoring is the fact that both the background characteristics (level of education and socioeconomic status) and pay levels of contract teachers can be different and sometimes lower than those of civil service teachers. While economic theory predicts that contracts renewable annually based on performance and monitored locally will increase teachers’ motivation to perform in line with performance measures established by the hiring unit, both teacher effort and results may be decreased by the lower compensation embodied in many of these contracts or by lower teacher capacity (at least in terms of formal preparation). How these competing tendencies play out in practice, and how they affect learning outcomes for students, has been a subject of much debate. Only recently has enough robust evidence from diverse settings begun to offer some answers.

Impact Evaluation Evidence

India: Balsakhi contract teacher program

The Balsakhi (“child’s friend”) program, implemented in two cities in India (Mumbai and Vadodara), paid locally recruited young women with a tenth-grade (secondary school) education roughly 10 percent of a regular civil service teacher’s salary to teach basic literacy and numeracy skills to third- and fourth-grade children who needed remedial support. The targeted children left the classroom in the afternoon and received tutoring for two hours per day. The program was highly effective in boosting the learning of these children, to the extent that average math and reading test scores in the treatment schools increased by a 0.14 standard deviation in the first year and a 0.28 standard deviation after two years over the averages in the comparison schools (Banerjee and others 2007). Most of the gains were attributable
### Table 4.1 Summary of Evaluated Contract Tenure Reforms

<table>
<thead>
<tr>
<th>Country (evaluation date)</th>
<th>Design, coverage</th>
<th>Performance measure</th>
<th>Contract teacher relative wage</th>
</tr>
</thead>
<tbody>
<tr>
<td>India: Balsakhi teacher program (Banerjee and others 2007)</td>
<td>pilot program, two cities</td>
<td>student test scores; teacher absence</td>
<td>10% of civil service wage; lower educational level</td>
</tr>
<tr>
<td>Kenya: contract teachers (Duflø, Dupas, and Kremer 2009)</td>
<td>pilot program, 140 schools in rural Kenya</td>
<td>student test scores; teacher attendance and time spent actively teaching; student attendance</td>
<td>30% of civil service wage; same educational level</td>
</tr>
<tr>
<td>India: Andhra Pradesh (Muralidharan and Sundararaman 2010a)</td>
<td>pilot program, 100 schools in rural Andhra Pradesh</td>
<td>student test scores; teacher attendance and time spent actively teaching</td>
<td>20% of civil service wage; lower educational level</td>
</tr>
<tr>
<td>Mali, Niger, and Togo (Bourdon, Frölich, and Michaelowa, 2007)</td>
<td>national programs</td>
<td>student test scores</td>
<td>—</td>
</tr>
<tr>
<td>Peru (Alcazar and others 2006)</td>
<td>provincial program</td>
<td>teacher absence</td>
<td>—</td>
</tr>
<tr>
<td>Locus of contracting</td>
<td>Contract term</td>
<td>Period observed</td>
<td>Evaluation method</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------------</td>
<td>----------------</td>
<td>------------------</td>
</tr>
<tr>
<td>local NGO</td>
<td>annual, renewable</td>
<td>3 yrs.</td>
<td>RCT</td>
</tr>
<tr>
<td>school councils</td>
<td>annual, renewable</td>
<td>3 yrs. (2 of intervention, 1 of follow up)</td>
<td>RCT</td>
</tr>
<tr>
<td>school committees</td>
<td>annual, renewable</td>
<td>2 yrs.</td>
<td>RCT and matching</td>
</tr>
<tr>
<td>school councils</td>
<td>annual, renewable</td>
<td>—</td>
<td>matching</td>
</tr>
<tr>
<td>—</td>
<td>—</td>
<td>—</td>
<td>matching</td>
</tr>
</tbody>
</table>

(continued next page)
Table 4.1 Summary of Evaluated Contract Tenure Reforms *Continued*

<table>
<thead>
<tr>
<th>Country (evaluation date)</th>
<th>Design, coverage</th>
<th>Performance measure</th>
<th>Contract teacher relative wage</th>
</tr>
</thead>
<tbody>
<tr>
<td>India: Madhya Pradesh (Goyal and Pandey 2009a)</td>
<td>statewide program, 200 schools</td>
<td>student test scores; teacher attendance and activity</td>
<td>20–25% of civil service wage; higher educational level</td>
</tr>
<tr>
<td>India: Uttar Pradesh (Goyal and Pandey 2009a)</td>
<td>statewide program, 200 schools</td>
<td>teacher attendance and activity</td>
<td>20–25% of civil service wage; higher educational level</td>
</tr>
</tbody>
</table>

*Source:* Authors’ compilation.

*Note:* RCT = Randomized control trial. NGO = nongovernmental organization. — = not available. SD = standard deviation(s).
<table>
<thead>
<tr>
<th>Locus of contracting</th>
<th>Contract term</th>
<th>Period observed</th>
<th>Evaluation method</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>district office</td>
<td>3 yrs., renewable</td>
<td>2 yrs.</td>
<td>matching</td>
<td>(1) Contract teachers absent 27% of time vs. 37% for regular teachers. (2) Contract teachers found teaching 37% of time vs. 25% for regular teachers during spot visits. (3) Student test scores positively associated with teacher effort. (4) Contract teachers’ absence and activity rates worsen in second contract year but still better than regular teachers.</td>
</tr>
<tr>
<td>village education committee</td>
<td>10 mos., renewable</td>
<td>2 yrs.</td>
<td>matching</td>
<td>(1) Contract teachers absent 26% of time vs. 39% for regular teachers. (2) Contract teachers found teaching 36% of time vs. 19% of time for regular teachers during spot visits. (3) Student test scores positively associated with teacher effort. (4) Contract teachers’ absence and activity rates worsen in second contract year to similar levels as regular teachers.</td>
</tr>
</tbody>
</table>
to increases in learning by the children who had received remedial tutoring. Even one year after the program ended, the average student in the Balsakhi schools had test scores that were 0.1 of a standard deviation higher than the average student in comparison schools.

Although the program was clearly effective as a targeted, complementary instruction strategy using low-cost community-based tutors, it was not designed as a “head-to-head” trial of the cost-effectiveness of contract teachers as opposed to civil service teachers. In addition to their different contract status and pay levels, the Balsakhi teachers had different curriculum goals, teaching hours, and class sizes from those of the regular schools. Due to the design of the study, intermediate variables of teacher effort, such as absence rates and use of instructional time, were not monitored. However, the large learning gains produced by teachers contracted at 10 percent of the prevailing civil service teacher salary provides evidence that, at least in some contexts, contract teachers’ lower average qualifications, lower pay, and lack of long-term contract security do not impede effective performance.

Kenya: Extra Teacher Program

The Extra Teacher Program (ETP) in western Kenya provided funding to a randomly selected set of schools to allow their school committees to hire a local contract teacher. Given an excess supply of graduates from Kenyan teacher training schools, contract teachers in Kenya, unlike India, have the same academic qualifications as civil service teachers (Duflo, Dupas, and Kremer 2009). However, the contract teachers were paid less than one-third of the civil service pay and could be fired by the school committee after each annual performance review.

The evaluation found that contract teachers were 15 percentage points more likely to be in class and teaching during unannounced visits than were civil service teachers in the comparison schools. The attendance record of contract teachers was even more impressive—30 percentage points higher—when compared with that of civil service teachers in their own schools, whose absence rates increased when the contract teachers began working. However, training programs for the school committees in how to monitor teacher attendance and performance mitigated the “shirking” behavior of civil service teachers over time.

The overall ETP program involved a number of different school-level interventions, including cutting class size in half (through the hiring of contract teachers) and assigning students to smaller classes, either randomly or tracked by ability. Overall, the program produced significantly higher student test scores. Among the program’s results was clear evidence that students randomly assigned to contract teachers scored a 0.21 standard
deviation higher on reading and math tests than their schoolmates assigned to civil service teachers. Students assigned to contract teachers also attended school more regularly (88 percent of the time versus 86 percent, corresponding to an 11 percent drop in absence rates). The stronger learning outcomes produced by contract teachers provides evidence of higher—or at least more effective—teaching effort than that of civil service teachers, despite the contract teachers’ lower salaries. However, the fact that contract teachers were assigned to a single class of students and stayed with those students for two successive years (unlike the typical pattern of a new teacher each year) may have played a role as well.

The evaluation evidence was less clear on why students of Kenyan contract teachers performed better. The contract teachers were absent less and their students’ attendance was slightly better, but relatively few school committees exercised the option not to renew a teacher’s contract on performance grounds. Nonetheless, provision of school-based management (SBM) training to school committees (on how to manage the hiring, performance monitoring, and renewal decisions for contract teachers) was correlated with positive long-term impacts of the program. At the three-year follow-up, only students of contract teachers in schools whose committees had been trained performed significantly better than students in control schools. The SBM schools were also more likely to use funding from the community to retain the contract teacher once program funding ended; 43 percent of them did so, compared with 34 percent of non-SBM schools.

As in other country cases, the authors noted questions around the long-term sustainability of the contract teacher advantage. The superior performance of these teachers could have been the result of better choice of teachers by the local school committees or the stronger incentives those teachers faced. But the contract teachers’ performance could also have reflected their motivation to perform well as a stepping stone toward higher-paid civil service teaching positions. While the ETP program did not provide a formal pathway to civil service status, it did allow teachers to acquire valuable experience; by the end of the program, 32 percent of the contract teachers did, in fact, acquire civil service teaching positions. Duflo, Dupas, and Kremer (2009) cautioned against assuming that average teacher performance, in Kenya or elsewhere, would match the performance of contract teachers in their study if all teachers were placed on alternative tenure contracts. In every setting where they were used, contract teachers worked alongside civil service teachers with higher pay and job security; the possibility of parlaying contract teaching experience into eventual civil service positions was a part of the incentive structure that contract teachers faced.
Andhra Pradesh, India: Contract teacher program

As part of a broader study of teacher incentives, contract teachers were hired in 2005 for a randomly selected set of schools across the state of Andhra Pradesh, India. As in the Balsakhi program, the contract teachers had lower average education than the civil service teachers in these schools and also tended to be younger, female, and more likely to live in the local villages. Whereas 84 percent of civil service teachers had a college degree and 99 percent had a formal teaching certificate, only 47 percent of contract teachers had completed college and only 12 percent had a teaching certificate. Eighty-one percent of the contract teachers lived in the local village, compared with 9 percent of the civil service teachers. Seventy-two percent of the contract teachers were female, with an average age of 24, while only 34 percent of the civil service teachers were female, with an average age of 39. The contract teachers’ pay was also dramatically lower: less than one-fifth of the civil service wage.

Over the two years of the program, the contract teachers were absent significantly less—16 percent of the time, compared with 27 percent for civil service teachers in the same school, with the performance differential higher in the second year than in the first (Muralidharan and Sundararaman 2010a). In both years, the contract teachers were also more likely to be found teaching during spot visits by observers. Among contract teachers, those with lower absence rates and higher rates of observed teaching activity in the first year of their contract had higher rates of contract renewal for the second year.

Higher effort from the contract teachers (measured by absence rates and observed teaching activity) plus the reduction in average class size that their recruitment permitted had positive effects on student learning. After two years, students in schools assigned a contract teacher scored a 0.15 standard deviation higher in math and a 0.13 standard deviation higher in language than their counterparts in the same grade in schools without contract teachers (Muralidharan and Sundararaman 2010a). In short, less-educated teachers who were paid a small fraction of the civil service wage in Andhra Pradesh appeared to be more accountable for performance than their civil service counterparts—in terms of both attendance and teaching activity—and helped boost overall school outcomes.

Although the study was a randomized trial, the recruitment of a contract teacher also had a major effect on class size in these small schools, so it is difficult to disentangle how much student learning outcomes were driven by the smaller classes rather than the differential effort or effectiveness of contract as opposed to regular teachers. Muralidharan and Sundararaman conducted a number of persuasive nonexperimental tests and concluded that the contract teachers were as effective as the regular teachers in improving their students’ learning. Given the contract teachers’ dramatically lower
wages, they were five times more cost-effective in producing educational results.

**Other contract teacher studies**

No other randomized evaluations have directly compared the performance of contract teachers and civil service teachers, but several other developing country studies have used matching methods to explore these issues. Contract teachers are widely used in West Africa, and Bourdon, Frölich, and Michaelowa (2007) compared the experience in Mali, Niger, and Togo, updating an earlier study of Togo by Vegas and De Laat (2003). They found that the presence of a contract teacher was positively correlated with the learning performance of low-ability students in the early grades but negatively correlated with the results of high-ability students in the upper grades.

A study in a rural province in Peru was the only study to date to document higher absence rates for contract teachers than for civil service teachers, by 12–13 percentage points (Alcazar and others 2006). The authors speculated that the lower salaries of contract teachers, in a context of weak local supervision, were a major reason for those teachers’ apparently lower effort. Unfortunately, the study did not include any data on the student learning performance of contract and civil service teachers.

In two additional states in India that made extensive use of contract teachers—Madhya Pradesh and Uttar Pradesh—Goyal and Pandey (2009a) carried out a nonexperimental analysis of the relative performance of contract and civil service teachers. In these states, unlike Andhra Pradesh, contract teachers typically were more educated than regular teachers, although they were much younger, had much less teaching experience, and were more likely to come from the local community. Consistent with the results of experimental studies in India, the contract teachers working in Madhya Pradesh and Uttar Pradesh public schools consistently demonstrated higher effort than regular teachers, whether measured as daily attendance or as the likelihood of being actively engaged in teaching during an unannounced visit. This higher effort was also correlated with significantly better learning outcomes for their students on language and math tests after controlling for other school, teacher, and student characteristics.

Goyal and Pandey followed the performance of the contract teachers over two years and noted that effort levels (attendance rates and likelihood of being found teaching) declined for teachers in their second contract period in both states and, in the case of Uttar Pradesh, became indistinguishable from regular teachers. They speculated that weak de facto oversight by school-level committees significantly reduced contract teachers’ incentives to perform. In both states, less than 6 percent of school
committee members were even aware that selection and oversight of contract teachers was one of their core responsibilities. In practice, most contract teachers were rehired after their first year, and a number of contract teachers in Madhya Pradesh were able to move into the regular teacher stream (although at a reduced salary).

Weak de facto performance oversight, high likelihood of contract renewal, and opportunities for transfer into the regular civil service stream (or of successful union mobilization to equalize contract and regular teachers’ employment terms) all reduced the differential in performance incentives between contract teachers and regular teachers and thus weakened the potential power of alternative contracting to strengthen teacher accountability for results. While there is uncertainty in the Indian context, as elsewhere, about the long-term sustainability of contract teacher policies, the evidence on their cost-effectiveness as a strategy for improving education accountability and outcomes is strong.

Contract Teacher Programs: The Balance of Evidence

The most rigorous of the available studies found contract teachers to be more cost-effective than regular civil service teachers—in India, dramatically so. In both Kenya and India, randomized trials have found learning outcomes for students of contract teachers to be better than those of civil service teachers, despite contract teachers’ much lower salaries. Non-experimental studies in India have found similar results. Earlier evidence on community-hired teachers in Central America (not included here but summarized thoroughly in Vegas 2005) was less robust, but that evidence also suggested that contract teachers achieved similar or better student grade progression and learning outcomes (controlling for student background) at lower cost.

Although contract teachers almost always work for lower salaries than their civil service counterparts, the cost-effectiveness of a contract teacher policy is likely to depend on country characteristics and the level of education involved. For example, all of the recent cases cited involved contract teachers at the primary level (where the supply of potential teachers with adequate capacity is not as likely to be constrained as at the secondary level or for specialty subjects such as sciences and math). It cannot be assumed that in all contexts it will be possible to recruit adequately qualified teachers at lower salaries.

In addition, there are major questions about the sustainability of this policy over time. Most of the evaluated cases suggested that contract teachers may have accepted the lower salaries and insecure tenure because they were queuing for civil service positions. Teachers’ unions have also aided contract teachers in some African countries and in Central America
to press successfully not only for tenure but also for equivalent wages—undermining the rationale for alternative contracting. In many of the cases evaluated, a large share of contract teachers do end up entering the civil service.

Nonetheless, the new wave of evidence on the short-term impacts of contract teacher reforms is fairly consistent: the use of contract teachers can strengthen the scope for local monitoring of teacher performance by parents and school councils, which results in higher teacher effort, which results in better student learning outcomes. In contexts where the supply of adequately trained teachers is not constrained, these positive outcomes can also be achieved at lower costs per student.

The evidence supports a theory of action in which the positive impacts of contract teacher reforms hinge on the de facto effectiveness of local monitoring. In evaluated cases where decisions about the hiring and retention of contract teachers were made higher up the administrative chain and not at the school level, or where local school committees were not equipped or empowered to put “teeth” into contract renewal decisions, impacts have been lower or have broken down relatively quickly. Alternative contracting can stimulate higher teacher effort and resulting improvements in student learning, but only if the hiring authority actually exercises the scope for holding teachers more accountable for performance.

Pay-for-Performance Reforms

Salary scales for teachers, unlike salary scales in most other sectors of the economy, are typically highly compressed, and movement across salary bands is rarely linked to individual results. These facts are all the more striking given the research evidence that individual teachers’ ability to produce educational results varies widely. There is less consistent evidence across countries (and over time) on whether the average level of teacher salaries is sufficiently competitive with those of comparable occupations to attract high-capacity individuals to teaching. But there is remarkable concurrence in the literature that the widespread pattern of relatively flat salary progression over teachers’ careers plus promotion policies rigidly linked to seniority combine to create weak incentives for teachers to perform to the best of their abilities and for high performers to remain in teaching. Indeed, in analyzing the large-scale migration of high-ability women out of the teaching profession in the United States between 1960 and 2010, Hoxby and Leigh (2004) argued that the “push” of compressed pay scales played a stronger role than the “pull” of more lucrative jobs in other sectors (Ballou and Podgursky 2002; Delannoy and Sedlacek 2001; Odden and Kelley 1997, cited in Umansky 2005).
To address these issues, school systems—and, indeed, other sectors and employers trying to stimulate higher worker productivity and efficiency—have resorted to bonus or “merit pay” schemes that establish a direct financial reward for the desired performance. In the United States alone, there have been hundreds of merit pay programs in education over the past century (Murnane and Cohen 1986). These programs typically share the objective of introducing some degree of variable annual compensation for teachers based on a measure of relative performance. The rewards can be based on input measures of performance (such as teacher attendance) or outcome or results measures (such as student learning progress). They can be targeted to either individual teachers or groups of teachers, typically at the school level.

Most of the cases reviewed in this chapter award bonus pay based on outcomes. The increased availability of student assessment data appears to be leading more school systems to try to link teacher pay directly to the performance measure they value most: student learning progress. Bonus pay is an attractive format because it does not increase the base salary bill and maintains the annual carrot of an incentive. Group bonuses, which reward all staff in a school for the school’s average results, are more common than individual teacher bonuses. One reason for this is sheer practicality; individual (classroom-level) teacher bonuses require the ability to measure learning gains for every subject and grade, which is costly administratively.

An interesting new model being used in Brazil is a combined approach. The core bonus is a group reward (for all school personnel, including administrative staff), calculated based on average school learning results, but downward adjustments are made in the amounts paid to individual teachers and staff based on their individual absence rates.

Impact Evaluation Evidence

There is tremendous policy innovation and interest in education pay for performance currently, in the OECD as well as the developing world, and a complete review is beyond the scope of this chapter. We focus on (1) developing-country experiences that have been rigorously evaluated and (2) important or innovative developing-country cases that are currently being rigorously evaluated, even if final results are not yet available.

To organize the diverse experiences, we use a typology based on two dimensions: what is rewarded, and whether the rewards are individual- or group-based. Table 4.2 presents a summary of data from the following reviewed cases:

- **Bonuses based on student learning outcomes**
  - Individual bonus (Andhra Pradesh, India 2009; Israel 2009)
  - Group bonus (Andhra Pradesh, India 2009; Kenya 2010)
• **Bonuses based on student learning plus other student outcomes**
  — Group bonus (Pernambuco, Rio de Janeiro, Sao Paulo, and Minas Gerais, Brazil forthcoming; Israel 2002)

• **Bonuses based on student outcomes and teacher input measures**
  — Group bonus (Chile 2009)

• **Bonuses based on teacher input measures only**
  — Individual bonus (Rajasthan, India 2010; Kenya 2001)

### Bonuses Based on Student Learning Outcomes

**India: Andhra Pradesh**

An ongoing randomized study in the Indian state of Andhra Pradesh offers the most persuasive evidence to date of the potential for individual and group bonuses for teachers to motivate more effective teacher performance in a developing-country setting. In a statewide representative sample of 500 schools, Muralidharan and Sundararaman (2009) carefully measured the impact of four alternative treatments applied in 100 schools each:

- An individual teacher bonus
- A group teacher bonus
- Provision of one extra contract teacher (input strategy)
- A block grant to the school (input strategy)

One hundred schools, not eligible to receive either incentives or inputs, served as the comparison group. To minimize Hawthorne effects, all schools (including the comparison schools) received the same amount of monitoring and measurement, differing only in the treatment received. Beginning and end-of-year tests were administered to all students and used to estimate value-added (gain) scores.

All interventions were designed to cost the same: Rs 10,000 per school (around $200), roughly equivalent to a teacher’s monthly salary (including benefits).\(^1\) Even though the four interventions were calibrated to cost the same, in practice the group incentive treatment ended up costing less—about Rs 6,000 ($125) per school.\(^2\)

The teacher incentive bonus was structured as a fixed performance standard, meaning that awards were distributed to any teacher or school that raised test scores by 5 percentage points or more over their baseline test scores. Below this threshold, the bonus was zero. Above this threshold, the bonus was calculated as the percentage additional gain in average test scores, multiplied by a slope of Rs 500 ($10). The average bonus was calibrated to be around 35 percent of a typical teacher’s monthly salary. Individual bonus payments were based on the average improvement in test scores for that teacher’s particular class. In group-incentive schools, teachers received the same bonus, based on the average school-level
## Table 4.2 Summary of Evaluated Pay-for-Performance (Bonus Pay) Reforms

<table>
<thead>
<tr>
<th>Country (evaluation date)</th>
<th>Bonus type</th>
<th>Design, coverage</th>
<th>Performance measure</th>
<th>Award process</th>
<th>Predictability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bonus based on student learning outcomes—Individual</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>India: Andhra Pradesh (Muralidharan and Sundararaman 2009)</td>
<td>indiv. pilot, 100 schools</td>
<td>gain in student TS (end-of-yr. vs. start-of-yr. TS for each classroom)</td>
<td>piecewise formula: bonus a function of % gain in TS, above a threshold of 5% gain</td>
<td>60% of teachers in first yr. got some bonus</td>
<td></td>
</tr>
<tr>
<td>Israel (Lavy 2009)</td>
<td>indiv. pilot, 629 teachers in 49 high schools</td>
<td>avg. student scores on matriculation exams and avg. pass rate relative to predicted scores (adjusted for student SES)</td>
<td>rank order of teachers</td>
<td>302 of 629 (48%) teachers got some award</td>
<td></td>
</tr>
<tr>
<td><strong>Bonus based on student learning outcomes—Group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>India: Andhra Pradesh (Muralidharan and Sundararaman 2009)</td>
<td>group pilot, 100 schools</td>
<td>avg. gain in student test scores at school level relative to baseline scores</td>
<td>piecewise formula (see above)</td>
<td>(most teachers got some award)</td>
<td></td>
</tr>
<tr>
<td>Kenya (Glewwe, Ilias, and Kremer 2010)</td>
<td>group pilot, 50 schools</td>
<td>avg. gain in student test scores at school level relative to baseline scores</td>
<td>rank order of “top-scoring” and “most-improved” (schools can win in only one category)</td>
<td>24 of 50 schools got award</td>
<td></td>
</tr>
<tr>
<td><strong>Bonus based on student learning plus other student outcomes—Group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brazil: Pernambuco (Ferraz and Bruns, forthcoming)</td>
<td>group statewide, 950 schools</td>
<td>school-level targets for improvement in IDEPE (state TS and student grade progression); individuals’ bonuses discounted based on absence rates</td>
<td>piecewise formula: above threshold of 50% of target attained, up to limit of 100% of target</td>
<td>479 of 929 schools (52%) in 2009; 758 of 954 schools (79%) in 2010 received bonus</td>
<td></td>
</tr>
<tr>
<td>Brazil: Sao Paulo (Ferraz and Bruns, forthcoming)</td>
<td>group statewide, 5,500 schools</td>
<td>school-level targets for improvement in IDESP (state TS and student grade progression); individuals’ bonuses discounted based on absence rates</td>
<td>piecewise formula: from 1% of target attained, up to limit of 120% of school target</td>
<td>2009: 100% of schools, 87% of personnel got some bonus (33% of personnel received none for excess absence); 2010: 73% of schools, 77% of personnel got some bonus</td>
<td></td>
</tr>
<tr>
<td>Brazil: Minas Gerais (Ferraz and Bruns, forthcoming)</td>
<td>group statewide, 3,972 schools</td>
<td>school-level targets for improvement in IDEMG (state TS and student grade progression); individuals’ bonuses discounted based on absence rates; school targets negotiated annually with regional admin.</td>
<td>piecewise formula: above threshold of 60% of target attained on first phase of institutional evaluation, capped at 100% of target</td>
<td>33% of schools in 2009</td>
<td></td>
</tr>
<tr>
<td>Monitoring and support</td>
<td>Bonus size, distribution</td>
<td>Bonus frequency</td>
<td>Cost</td>
<td>Evaluation method</td>
<td>Results</td>
</tr>
<tr>
<td>--------------------------------------------------------------------------------------</td>
<td>--------------------------</td>
<td>-----------------</td>
<td>-------------------</td>
<td>-------------------</td>
<td>-------------------------------------------------------------------------</td>
</tr>
<tr>
<td>substantial: feedback on test results, interviews, classroom monitoring</td>
<td>35% of MW</td>
<td>annual during 2-yr. experiment</td>
<td>Rs 1.9 million ($42,200)</td>
<td>RCT</td>
<td>0.27 SD improvement in learning outcomes compared with control schools by second yr.</td>
</tr>
<tr>
<td>limited: some interviews conducted</td>
<td>$1,750–$7,500 per teacher per subject (70–300% of MW)</td>
<td>annual, 1-yr. experiment</td>
<td>— (est. $1.5 million)</td>
<td>quasi-randomized trial, RDD</td>
<td>(1) 14% higher math pass rate and 10% higher scores (2) 5% higher English pass rate and 4% higher scores</td>
</tr>
<tr>
<td>substantial (see above)</td>
<td>$125/school distributed equally across teachers (about 35% of MW)</td>
<td>annual during 2-yr. experiment</td>
<td>Rs 1.8 million ($40,000)</td>
<td>RCT</td>
<td>0.16 SD higher test scores compared with control schools by second yr.</td>
</tr>
<tr>
<td>limited: one round of interviews conducted</td>
<td>21-43% of MW</td>
<td>one-time</td>
<td>KSh 1.8 million ($30,700)</td>
<td>RCT</td>
<td>(1) 0.14 SD higher test scores compared with control schools in first yr. (2) Gains not sustained one yr. after (3) No decrease in teacher absence (4) Increase in exam prep sessions</td>
</tr>
<tr>
<td>school director surveys and classroom observations in sample of 220–300 schools</td>
<td>all school personnel eligible, teaching or nonteaching avg. bonuses: 180% of MW (2009), 140% of MW (2010)</td>
<td>annual</td>
<td>2009: Rs$28.8 million ($15.3 million) 2010: Rs$40 million ($21.3 million)</td>
<td>RDD</td>
<td>Schools with more ambitious targets improved 0.15–0.31 SD more than comparable schools with lower targets</td>
</tr>
<tr>
<td>school director and teacher surveys</td>
<td>all school personnel eligible, teaching or nonteaching avg. bonuses: 150% of MW (2009), 150% of MW (2010)</td>
<td>annual</td>
<td>2009: Rs$600 million ($319 million) 2010: Rs$655 million ($350 million)</td>
<td>DD</td>
<td>—</td>
</tr>
<tr>
<td>school director and teacher surveys; classroom observations in sample of 600 schools</td>
<td>all school personnel eligible, teaching or nonteaching</td>
<td>annual</td>
<td>2009: Rs$311 million ($173 million) 2010: Rs$371 million ($206 million)</td>
<td>DD</td>
<td>—</td>
</tr>
</tbody>
</table>

(continued next page)
### Table 4.2 Summary of Evaluated Pay-for-Performance (Bonus Pay) Reforms Continued

<table>
<thead>
<tr>
<th>Country (evaluation date)</th>
<th>Bonus type</th>
<th>Design, coverage</th>
<th>Performance measure</th>
<th>Award process</th>
<th>Predictability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil: Rio de Janeiro City (Ferraz and Bruns, forthcoming)</td>
<td>group municipality-wide, 922 schools</td>
<td>school-level targets for improvement in IDERio (municipal TS and student grade progression); individuals’ bonuses discounted based on absence rates</td>
<td>threshold of 100% of target for early grades, at least 50% of target for upper grades, and ceiling of 110% of targets; no bonus for personnel with more than 5 days’ total absences (excused or unexcused)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Israel (Lavy 2002)</td>
<td>group pilot, 62 high schools</td>
<td>avg. credits taken, percentage of students receiving matriculation certificates, school dropout rates, school scores relative to predicted scores adjusted for student SES</td>
<td>rank order of schools</td>
<td>top 33% earned award</td>
<td>—</td>
</tr>
<tr>
<td>Chile: SNED (Rau and Contreras 2009)</td>
<td>group national (all public and publicly subsidized basic education schools)</td>
<td>avg. student TS on national exam (SIMCE) (37%), SIMCE gains (28%), other school factors (35%)</td>
<td>rank order tournament for schools stratified by region, urbanicity, grade level, and SES</td>
<td>top 25–35% of schools get awards</td>
<td>—</td>
</tr>
<tr>
<td>India: Rajasthan (Duflo, Hanna, and Ryan 2010)</td>
<td>indiv. pilot, 113 rural NGO schools</td>
<td>teacher daily attendance monitored with a date-stamped camera</td>
<td>piecewise formula: bonus of 50 rupees per day for additional days worked over 10-day-per-month threshold</td>
<td>bonus attainment automatic upon attendance</td>
<td>—</td>
</tr>
<tr>
<td>Kenya: preschools (Kremer and others 2001)</td>
<td>indiv. pilot, 50 preschools</td>
<td>school headmasters given resources to award teachers bonuses for good attendance</td>
<td>piecewise formula: deduction from maximum potential bonus based on days of absence; remaining funds went to school general account</td>
<td>headmasters granted full bonuses to all teachers, regardless of their actual attendance</td>
<td>—</td>
</tr>
</tbody>
</table>

Source: Authors’ compilation.

Note: RCT = randomized control trial. RDD = regression discontinuity design. DD = difference-in-differences. PSM = propensity score matching. SD = standard deviation(s). SES = socioeconomic status. TS = test scores. MW = monthly wage. — = not available. IDEPE = Index of Basic Education Development, Pernambuco (Brazil). IDESP = Index of Basic Education Development, Sao Paulo (Brazil). IDEMG = Index of Basic Education Development, Minas Gerais (Brazil). IDERio = Index of Basic Education Development, Rio de Janeiro (Brazil). SNED = National System for Performance Evaluation of Subsidized Educational Establishments (Sistema Nacional de Evaluación del Desempeño de los Establecimientos Educativos Subvencionados). SIMCE = National System for Measuring the Quality of Education (Sistema Nacional de Medición de la Calidad de la Educación). NGO = nongovernmental organization.
<table>
<thead>
<tr>
<th>Monitoring and support</th>
<th>Bonus size, distribution</th>
<th>Bonus frequency</th>
<th>Cost</th>
<th>Evaluation method</th>
<th>Results</th>
</tr>
</thead>
</table>
| school director and teacher surveys and classroom observations in sample of 200 schools | all school personnel eligible, teaching or nonteaching individual bonuses heavily discounted for absence | annual | 2010: R$14.6 million ($8 million) | RDD, DD | — |}

| limited | $13,250–$105,000 per school; teachers got $250–$1,000 (10–40% of MW) | one-time | $1.44 million | RDD | (1) 0.13 SD improvement in learning outcomes (2) Modest increases in credits earned and % of students taking matriculation exam |

| Chile has numerous teacher evaluation, observation, and school support programs not directly related to SNED | Initially 40% of MW, currently 70–80% of MW; 90% of award to teachers, 10% to school | annual for 2 yrs., then new tournament starts | — | DD with PSM, RDD | (1) 0.07–0.12 SD increase in average learning outcomes associated with introduction of SNED (2) No evidence that winning bonus stimulated subsequent improvement |

| teacher surveys and spot visits, classroom observation | up to 25% of MW each month | monthly | — | RCT | (1) Teacher absence in “camera” schools fell from 42% to 23% (2) Student TS increased by 0.17 SD (3) Grade completion increased |

| spot school monitoring visits; teacher interviews | up to 300% of MW | per term (3 terms in each school yr.) | K Sh 4.6 million ($61,700) | RCT | (1) No decline in teacher absence rate (29%) (2) No change in pupil attendance or TS |
improvement in test scores. Teachers in all incentive schools were told that bonuses would be paid at the beginning of the next school year, conditional on average improvements during the current school year (first year of the experiment). Inputs (a block grant or an extra teacher) were provided unconditionally to selected schools at the beginning of the school year.

At the end of two years of the program, both the individual and group teacher incentives were effective in improving student test scores. Students in incentive treatment schools had considerably higher test scores than those attending comparison schools, by 0.28 and 0.16 of a standard deviation in math and language, respectively. The qualitative analysis suggested that the main mechanism for the incentive effects was not increased teacher attendance but greater, and more effective, teaching effort conditional on being present. In particular, teachers in incentive schools were significantly more likely to have assigned homework and class work, conducted extra classes beyond regular school hours, given practice tests, and paid special attention to weaker children. These results were obtained from self-reports—that is, answers to unprompted interview questions with treatment-school and control-school teachers. These behaviors, however, were not independently verified or observed in the classroom.

At the end of the two years of the study, there were significant (at the 10 percent level) differences between individual and group incentives, and between incentive schools and input schools. Individual incentives produced an average increase in student test scores of a 0.27 standard deviation, compared with a 0.16 standard deviation in group incentive schools. The input strategies also yielded positive effects when compared to control schools, but their magnitude (a 0.08 standard deviation) was substantially lower than incentive schools. In terms of cost-effectiveness, both the group and individual bonus programs were more cost-effective than the input programs, and they were roughly equal to each other in cost-effectiveness. Although the group bonus had a weaker impact on student learning results, this was offset by its lower costs.

To analyze the possibility that bonus payments induced negative impacts on student learning in other subjects, the researchers also studied learning outcomes in science and social studies, which did not enter into the calculation of the bonus. Test scores in these subjects in the incentive schools were also higher (0.11 and 0.18 of a standard deviation in science and social studies, respectively) at the end of two years of the program. Contrary to concerns that bonus pay can lead teachers to “teach to the test” at the expense of nontested subjects, at least in some contexts there may actually have been positive externalities from bonus programs. In contexts of low average literacy and numeracy skills, the researchers hypothesized that teacher efforts to increase test scores in math and language can have
positive spillover effects on students’ mastery of other subjects (Muralidharan and Sundararaman 2009).

In a related experiment in the same setting, the researchers analyzed whether providing teachers with more detailed feedback on their students’ performance could increase the power of bonus incentives to improve test scores. The NGO running the project provided individual teachers with written diagnostic feedback on the performance of their class of students, in both absolute and relative terms, at the beginning of the school year. The reports also included specific suggestions on how to improve learning levels in specific areas of weakness. The program was implemented in 100 rural primary schools that were randomly selected from the previous treatment schools across the state of Andhra Pradesh.

The authors found that, by itself, the feedback treatment did not appear to have any significant effects. However, feedback combined with teacher incentives had a significant effect on student test scores. Based on indicators of teacher activity, the researchers concluded that while teachers in all of the feedback schools could have used the reports effectively if they had wanted to, only teachers in the incentive schools seemed to have done so. This suggests positive interactions between incentives and inputs and the possibility for incentives to also raise the effectiveness of other school inputs such as teacher capacity building (Muralidharan and Sundararaman 2010b).

No other research to date provides a direct comparison of the impact of individual and group-based teacher bonuses in the same context. But there is additional experimental evidence on individual teacher bonuses based on student learning results from Israel and a group-based bonus program in Kenya.

Israel: Individual teacher incentive experiment
Lavy (2009) evaluated a tournament-type bonus program in Israel that ranked teachers on the basis of value-added contributions to their students’ test scores on high school matriculation exams, above and beyond the predicted scores for those students based on their socioeconomic characteristics, their level of study in the relevant subject, grade level, and a fixed school-level effect. Thus, teachers competed against each other in a fair way to produce learning gains in their classes.

The program covered 629 teachers, of whom 207 competed in English and 237 in math. Relative to other incentive programs, the bonuses for this program were large and could amount to $7,500 per class (one teacher won two first-place awards totaling $15,000—equal to six months of salary). Due to a measurement error in the way schools were assigned into the program, it was possible to approximate a randomized trial of the incentive offer. Lavy found that the program had significant positive effects on student achievement by increasing the test-taking rate among high school students.
seniors as well as the average pass rates and average test scores in both math and English.

Postprogram interviews with participating teachers generated interesting insights into how they responded to the incentives. Compared with teachers in schools not eligible for the program, teachers offered the incentive modified their teaching methods in several ways. They were significantly more likely to add after-school classes, to track students by ability within the classroom, and to tailor instruction to the needs of individual students, with special focus on students of weaker ability. These actions resulted in a higher share of students taking the matriculation exams than otherwise would have and higher pass rates and average scores across all test takers.

Lavy also found that teachers’ effectiveness (measured by their success in achieving the bonus) was not highly correlated with “observable” teacher characteristics such as age, education level, teaching certification, or years of experience. However, teacher performance was correlated with the caliber of university attended; teachers who had graduated from top-ranked Israeli universities were significantly more effective than those who had attended less-prestigious universities or teachers’ colleges. Unfortunately, although the program was initially designed to run for three years, a change of government caused its cancellation after a single school year.

Kenya: ICS teacher incentive program
A program in Kenya provided a group-based incentive to teachers, based on improvements in average student performance (Glewwe, Ilias, and Kremer 2010). Fifty rural schools were randomly selected (out of 100) for the incentive program, which was implemented over two years with one additional year of follow-up to observe long-term effects. Each year the program provided in-kind prizes, such as bicycles, valued at up to 43 percent of a typical fourth- to eighth-grade teacher’s monthly salary. The prizes were awarded based on the performance of the school as a whole on the Kenyan government’s districtwide exams. Performance was measured relative to baseline test scores at the beginning of the school year. Awards were offered to “top-scoring” or “most-improved” schools (schools could win in only one of the two categories). In each category, prizes were awarded for first place (three prizes), second place (three prizes), third place (three prizes), and fourth place (three prizes). Overall, 24 of the 50 selected schools received a prize of some kind, and teachers in most schools felt they had a reasonable chance of winning a prize (Glewwe, Ilias, and Kremer 2010).

The program was careful to track the students initially enrolled in each school to ensure that new (potentially talented) students could not be recruited to take the exam and that poor-performing students were not held back from taking the exam. During the two years the program
operated, a higher share of students in treatment schools than in comparison schools took the exam, which did generate gains in test scores. By the second year, students in the bonus schools scored, on average, a 0.14 standard deviation higher on the exams than did students in comparison schools, with the strongest improvements seen in geography, history, and religion (around a 0.34 standard deviation during the first and second years versus the baseline [“year 0”] on the district exam, and a 0.20 standard deviation versus the comparison schools in the second year on the district exam). The next-largest effects were in science and math (0.20 and 0.15, respectively, of a standard deviation on the district exam versus the baseline year), with no significant effects in other subjects.

However, these gains proved short-lived. One year after the program ended, there were no significant differences in test performance across the schools. Glewwe, Ilias, and Kremer (2010) speculated that teachers’ strategies for achieving the bonus focused on short-run efforts to boost performance on the government tests, such as after-school tutoring in test-taking techniques, rather than changes in their core pedagogy or effort levels that might have had a higher chance of promoting long-term learning. For example, teacher absence rates did not decline from the baseline level of 20 percent of school days missed. Classroom observations did not detect any changes in homework assigned or use of learning materials. But by the second year of the program, bonus-eligible schools were 7.4 percentage points more likely than comparison schools to conduct exam preparation sessions. The benefits of this strategy were narrow, however. When researchers applied tests using a format different from those of the government exams, they found no difference in student performance between treatment and comparison schools.

Bonuses Based on Learning Improvement Plus Other Outcomes (Group-Based)

We know of no evaluated pay-for-performance programs based on multiple student outcomes that reward individual teachers, but the evidence on group-based bonuses of this type is increasing, thanks to a wave of innovation in this area in Brazil. Part of the reform stimulus is a conviction among Brazilian policy makers that low teacher quality is the binding constraint to education improvement and that restructuring the incentives for teachers must be part of the solution.

Another factor appears to be the federal government’s establishment of the IDEB (Index of Basic Education Development) in 2007 (discussed further in box 4.1), which has provided states and municipalities with a convenient, transparent, and universally accepted metric for setting and monitoring annual school-level targets for improvement. Between 2008
Targets that Avoid Perverse Incentives: Brazil’s Index of Basic Education Development

The Brazilian Ministry of Education in 2007 introduced an innovative tool for systematic annual monitoring of basic education progress in every school, municipality, state, and region of the country: the Índice de Desenvolvimento da Educação Básica (Index of Basic Education Development, or IDEB). The innovation lies in IDEB’s combined measure of student learning results and student flows (grade progression, repetition, and graduation rates). Because the index is the product of both test scores and pass rates, it discourages automatic promotion of children who are not learning. However, it also discourages schools from holding children back to boost learning scores. Avoiding incentives for grade retention is important in Brazil, where the average repetition rate in primary school is approximately 20 percent, the highest in Latin America.

IDEB builds on the progress Brazil has made in scaling up its national student assessment system to a technically well-regarded learning assessment of math and language—called the Prova Brasil (Brazil Test)—that is applied every two years to every fifth-, ninth-, and twelfth-grade student. The IDEB measure combines Prova Brasil test results with administrative data on school enrollments, repetition, and grade promotion. The raw scale of the exams varies from 0 to 500, and the standardized scale ranges from 0 to 10. Pass rates are calculated based on the information reported by each school to the National School Census, applied yearly by the Ministry of Education.

The IDEB for each grade-subject is calculated as the product of the standardized Prova Brasil score and the average pass rate for the cycle evaluated (π):

\[
IDEB_{asj} = \text{ProvaBrasil}_{asj} \times \pi_{asj}
\]

(4.1)

where \(a\) is the subject evaluated (Portuguese or mathematics),
\(s\) is the cycle evaluated,
\(j\) is the school.

The average pass rate in the cycle varies between 0 and 1 (1 if the pass rate equals 100 percent). The standardized IDEB measure thus varies between 0 and 10.

The index has become rapidly accepted in Brazil as the leading metric for gauging the relative performance of individual schools as well as municipal, state, and private school systems. Biannual IDEB results are (continued next page)
widely reported in the media, and the federal government has established targets for IDEB improvement for all of Brazil’s 26 state and 5,564 municipal school systems.

IDEB has also facilitated the implementation of teacher bonus programs at both the state and municipal levels over the past three years. Although the different state and municipal programs in operation have a number of alternative design features, all are based on the IDEB metrics (states typically apply state-level tests that are equated with the Prova Brasil in the off years to generate a corresponding annual measure of performance). From the standpoint of federal education policy, IDEB has created a powerful platform for comparative analysis of state and municipal innovations in basic education.


and 2010, six states and one large municipality in Brazil adopted annual teacher bonus programs based on IDEB results.

Because each of the Brazilian programs is being implemented system-wide, it is not possible to evaluate rigorously the fundamental question of whether introducing a bonus program *causes* the education system to improve. In these systems, there are no schools operating outside of the bonus regime and thus no perfectly valid comparison group. However, researchers plan to track the evolution of results in the “bonus” and neighboring nonbonus states and municipalities over time, using difference-in-differences analysis. And rigorous evidence on how different design features of the bonus programs affect school performance is being generated, exploiting discontinuities in the ways that school-level targets have been set. The “rules of the game” for the bonus programs in Pernambuco state and Rio de Janeiro municipality are most conducive to this approach. Although both evaluations are in an early stage, some first-round results from Pernambuco are discussed here.

**Pernambuco, Brazil: Group-based teacher bonus program**

In 2008, Pernambuco became the third state in Brazil to introduce a pay-for-performance system that rewards school personnel for the attainment of annual school improvement targets. All schools that achieve at least 50 percent of their targets receive a proportional bonus, up to a cap of 100 percent. Because the state budgets one month’s education payroll for the program annually, the average bonus will exceed one month’s
Making Schools Work

salary if less than 100 percent of schools achieve it. In the first year of the program, 52 percent of schools achieved over 50 percent of their targets, and the awards averaged 1.8 months of salary for most recipients. In the second year, 79 percent of schools received the bonus, and the average award was 1.4 months of salary.

This is a relatively large incentive compared with other programs internationally. Pernambuco’s rule that schools achieving less than 50 percent of their targets receive nothing also creates a strong incentive. In some other Brazilian states, the rules of the game allow every school to receive some degree of bonus proportional to results, no matter how slight. Although the bonus is group-based (in that it rewards the whole school for its results), school directors have no discretion in how the funds are distributed. Each member of the school staff (teaching and nonteaching) receives the equivalent percentage bonus applied to the staff member’s monthly salary.

The “strength” of the incentives embedded in the Pernambuco program’s design make it an important case to analyze. The rules used to set the initial targets in Pernambuco—based on whether schools fell in the bottom 25 percent, 26–50 percent, 51–75 percent, or 76–100 percent of the performance distribution in 2007—created discontinuities in the targets that permit rigorous evaluation of their effects. Similarly performing schools had more or less ambitious targets depending on which side of these cutoffs they happened to fall. These discontinuities permit research on a number of important questions: How, in the short run, do schools respond to targets for improvement that are more or less ambitious? Do “stretch” targets motivate higher effort, or do they cause schools to give up?

Pernambuco’s rule restricting bonus payments to schools that achieve at least 50 percent of their targets allows analysis of how achieving or not achieving a bonus payment in a given year affects a school’s effort and strategies in subsequent years—including adverse impacts such as encouraging more time on tested subjects, teaching narrowly to the test, or inducing teachers to migrate to schools that are successful in achieving the bonus. Finally, as the evolution of Pernambuco’s student outcomes can be compared over time to those of other states without bonus programs (using difference-in-differences analysis), the comparison will shed some light on the ultimate question: how does introducing a system of pay for performance affect student learning, pass rates, and teaching practices?

Results reported here are preliminary (Ferraz and Bruns, forthcoming). The first bonus was awarded in June 2009, based on schools’ performance relative to their targets for the end-2008 school year (measured on standardized tests and administrative data collected by the federal government in December 2008). Fifty-two percent of Pernambuco’s 929 schools achieved the bonus. A new set of targets was established for each school for the 2009 school year, and in mid-2010 the second round of bonus
payments was made. This time, 79 percent of schools (and 81 percent of Pernambuco’s 41,337 school-level personnel) received the bonus. The key findings thus far from the Pernambuco study are summarized below.

Acceptance of the program was relatively high. Sixty-four percent of school directors surveyed believed that the policy is an appropriate one, and 66 percent believed the program was having a positive impact on their school—whether or not they received the first-year bonus.

Schools with more ambitious targets achieved more progress (other things being equal). In almost every performance category (fourth-, eighth-, or twelfth-grade math or Portuguese), schools falling on the “higher target” side of the performance cutoffs made larger test score gains than the comparison schools just below the cutoffs. The differential learning gains were sharpest for schools just above the 25th percentile of performance. For the eighth grade in 2008, for example, schools on the “higher target” side of the cutoff improved average Portuguese scores by a 0.31 standard deviation more than the schools just below the cutoff (with less-ambitious targets). In math, the differential gain was a 0.15 standard deviation. At the second cutoff, just above and below the 50th percentile in the performance distribution, improvements were also higher for the schools with more ambitious targets, but they were also of smaller magnitude. For the other tested grades (fourth and eleventh), impacts were in similar ranges but varied across subjects and, in a few cases, by cutoff point. Overall, however, the evidence was consistent—at least over the very short term—that higher targets in the presence of an attractive incentive in Pernambuco induced higher learning results by schools.8

Learning levels across the state improved significantly. By the end of 2009, the second year of the bonus program, Pernambuco’s state schools as a whole registered significant average improvements in learning, especially in Portuguese. Average Portuguese scores in the eighth and eleventh grades increased by 0.44 and 0.57 of a standard deviation, respectively. Math scores in the eighth and eleventh grades rose by 0.27 and 0.31 of a standard deviation, respectively. These learning gains are large relative to observed results from other teacher incentive programs. However, since this was a universally applied program within the state, these are raw gains, not gains relative to a comparison group. Difference-in-differences analysis will be needed to bound these gains. As Pernambuco ranked last among Brazil’s 26 states on the 2007 IDEB rankings for primary schools (through eighth grade) and 19th for secondary schools, a significant part of these gains likely reflects reversion to the mean.

Schools that just missed receiving the bonus in 2008 improved more in the following year than schools that barely achieved the bonus. A key research question is whether schools that received the bonus in 2008 would be more motivated in 2009, or would they exert less effort and coast. To examine this,
the performance of schools that fell just short of 50 percent of their 2008 targets (and did not receive the bonus) was compared with the performance of schools that achieved just over 50 percent of their targets and did get the bonus. Controlling for schools’ 2008 test results and other school characteristics, schools that barely missed the bonus in 2008 improved more than schools that barely achieved it. It appears that—at least for schools that came fairly close in 2008—not getting the bonus had a positive effect on schools’ motivation and performance.

_Schools whose teachers spent more time on instruction were much more likely to achieve the bonus._ In contrast to studies that have found no clear evidence of changes in teacher classroom practice to explain student learning improvements caused by the bonus, researchers in Brazil found significant correlations. A novel feature of the research is use of the Stallings “classroom snapshot” instrument (Abadzi 2009; Bruns, Evans, and Luque 2010) to generate detailed data on the pathways through which teacher incentive programs such as Pernambuco’s bonus affect teacher practice in the classroom. In theory, if an incentive _causes_ an improvement in student outcomes, it should operate through changes in teacher behavior that are induced by the incentive, such as increased or more-effective teaching effort.

Periodic observations are tracking comprehensive indicators of classroom dynamics (teachers’ use of instructional time, materials, interaction with students, and student engagement) in a large sample of Pernambuco’s 950 schools (1,800 classrooms in close to 300 schools, with oversampling of schools just above and below the original target discontinuities). Both tested and nontested subjects are being observed to try to capture adverse effects, such as diversion of school time from nontested subjects.

The first of several unannounced observations was carried out one month before the end-of-year tests that entered into the second-year bonus calculation. The data, shown in table 4.3, uncovered two main findings.

First, average teacher practice diverges widely from norms observed in the United States, where the Stallings instrument has been most extensively used. Against a U.S. good-practice benchmark of 85 percent of total class time effectively applied to instruction, the average observed in Pernambuco schools was 61 percent. Against a good-practice benchmark of 15 percent of time spent on routine classroom processes (such as taking attendance, passing out papers, or cleaning the blackboard), schools in Pernambuco averaged 28 percent.

A second finding, however, was that significant disparities existed across schools, and these were highly correlated with schools’ likelihood of achieving the 2009 bonus (paid in 2010 based on year-end performance in 2009). While teachers were off-task a very high 12 percent of the time across the whole sample—either out of the classroom due to late arrival or
early departure or engaged in social interaction with students or colleagues—such time loss was much more significant in the schools that did not subsequently achieve the bonus (17 percent of total class time) than in those that did (10 percent of time lost).

Teachers in both successful (bonus-achiever) and less-successful schools spent a high share of total class time on routine management processes by U.S. standards. However, teachers in the bonus schools registered significantly less time off-task and were able to devote this time to instruction: learning activities absorbed 62 percent of total class time in schools that did go on to achieve their targets, compared with only 53 percent of time in schools that did not. There is also evidence of more intensive use of learning materials and higher rates of student engagement. The research as yet cannot determine whether these teacher behaviors caused students to improve more or whether the behaviors simply reflected the fact that better students are easier to manage and teach. But the sample of bonus achievers included schools from all parts of the performance distribution, including a large number of low-income and low-performing schools, because the targets measured improvement from a school’s own baseline. The second year of observations will measure whether and how teachers’ practice and classroom dynamics evolve after schools process the “information shocks” and incentives of either achieving or not achieving the bonus.

The evaluation of Pernambuco’s pay-for-performance program is expected to continue for several more years, permitting deeper analysis of

<table>
<thead>
<tr>
<th>Teacher use of classroom time</th>
<th>U.S. good-practice benchmarks (percent)</th>
<th>Overall PE sample (percent)</th>
<th>Achieved 2009 bonus (percent)</th>
<th>Didn’t Achieve 2009 bonus (percent)</th>
<th>Difference, bonus Vs. nonbonus schools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning activities</td>
<td>85</td>
<td>61</td>
<td>62</td>
<td>53</td>
<td>0.09 (0.04)**</td>
</tr>
<tr>
<td>Classroom management</td>
<td>15</td>
<td>28</td>
<td>27</td>
<td>30</td>
<td>–0.03 (0.03)</td>
</tr>
<tr>
<td>Teacher off-task</td>
<td>0</td>
<td>12</td>
<td>10</td>
<td>17</td>
<td>–0.04 (0.02)***</td>
</tr>
<tr>
<td>Teacher out of classroom</td>
<td>0</td>
<td>8</td>
<td>8</td>
<td>12</td>
<td>–0.04 (0.02)***</td>
</tr>
</tbody>
</table>

Source: Bruns, Evans, and Luque 2010.

Note: PE = Pernambuco.

Standard errors in parentheses: ** significant at the 5% level; ***significant at the 1% level.
how annual targets and past rewards affect schools’ improvement strategies, teacher behaviors, and overall system progress. Pernambuco’s results will also be directly comparable with the new program in the municipality of Rio de Janeiro, which also sets targets for improvements in IDEB outcomes as the basis for the bonus and which also has established performance bands that generate discontinuities in school targets around several different thresholds. As in Pernambuco, a large sample of schools is being followed in a panel study that includes systematic classroom observation, using the Stallings instrument. The Rio program design is additionally interesting because it embodies strong sanctions against teacher absence: only employees with five or fewer absences for the entire school year (whether excused or not) receive the bonus. In 2010, 290 of Rio’s 1,044 municipal schools qualified for the bonus based on their schools’ 2009 IDEB improvements, but more than one-fourth of these schools’ 11,000 employees did not meet the bar for individual attendance. Rio’s program is currently the strongest effort to attack absenteeism; it will be important to measure how the bonus incentives affect these rates over time.

These and the other new Brazilian experiences offer a promising opportunity to generate comparative evidence on some key issues in the design of teacher pay-for-performance programs. Before the new wave of programs in Brazil, however, the main evidence on group-based bonus incentives for improvements in student outcomes was a small-scale program implemented in the 1990s in Israel.

**Israel: Ministry of education school performance program**

Lavy (2002) examined a tournament-type program implemented in 1995 that provided a group incentive to teachers in 62 nonrandomly selected secondary schools. The objective of the program was to reduce dropout rates and improve academic achievement. The program used three performance measures: average number of credit units per student, proportion of students receiving a matriculation certificate, and the school dropout rate. Participating schools competed for a total of $1.44 million in awards.

School performance was measured in two stages. First, school average outcomes (the three performance measures used in the program) were normalized relative to an expected score, estimated using regressions that controlled for socioeconomic background of the students. Second, schools were ranked according to improvement in average outcomes. The top third of schools won awards. In 1996, the highest-scoring of these schools won $105,000, and the lowest-scoring won $13,250. Seventy-five percent of the award was distributed among teachers in the school, and the remaining 25 percent could be used for schoolwide improvements (for example, teacher common rooms). Teachers’ individual share of the school-level bonuses ranged from $250 to $1,000 (10–40 percent of the average monthly
wage) depending on the size of the school’s award.\textsuperscript{11} All teachers in a school received the same award, regardless of individual contribution to the school’s average results.

Lavy used a regression discontinuity approach to compare student outcomes in the 62 treatment secondary schools with 119 similar schools that just missed the program’s eligibility rules. Lavy’s results suggested that the monetary incentives had some effect in the first year of implementation (mainly in religious schools) and by the second year caused significant gains in student outcomes in all schools. The program led to improvements in the number of credit units taken by students and average test scores in matriculation exams. Average exam scores across treated schools in the second year of the program increased 2.4 points more than in the comparison group. From Lavy’s descriptive statistics, this can be estimated as approximately .12 of a standard deviation. In addition, Lavy reported that more students in treatment schools gained matriculation certificates, and these schools also reduced dropout rates in the transition from middle school (grades 7–9) to high school (grades 10–12).

An interesting feature of the Lavy (2002) paper is that it compared the incentive intervention with a “resource” intervention. The resource intervention was a separate program that rewarded schools showing improvement with in-kind resources, such as teacher training, of identical monetary value to the bonus program. Twenty-two schools were selected for the resource-intervention program. A comparison group of schools not admitted into the resource program served as the basis for identification of program effects. The results of the resource program (evaluated only for secular schools) suggested that it also led to statistically significant improvements in student performance, but of much smaller magnitude. Lavy concluded that the school bonus program was more cost-effective.

**Bonuses Based on Student Learning Outcomes Plus Teacher Input Measures (Individual)**

Several new teacher bonus programs being implemented in Chile and at the state level in Mexico use a very different—and quite interesting—model. They provide individual rewards to teachers based on a combination of student outcomes and evaluation of teachers’ individual merit (including tests of teachers’ content mastery, direct observation of their classroom practice, and written assessments by principals and master teachers.)

The 2004 Chilean *Asignación Variable por Desempeño Individual* (Variable Allocation for Individual Performance, or AVDI) program is the most sophisticated example, given its use of comprehensive, high-stakes performance reviews (including videotapes of teachers’ classroom practice) and a significant bonus in the form of a four-year increase in base pay for teachers.
who are rated highly. New programs in the state of Mexico (PROEBB and Ser Maestro) and the Premio al Merito Docente (Teaching Award of Merit) in the state of Nuevo Leon also combine evaluation of teachers’ individual performance with student outcome measures. Unfortunately, none of these promising experiences is being evaluated rigorously.

Bonuses Based on Student Learning Outcomes Plus Teacher Input Measures (Group-Based)

Chile: National System for Performance Evaluation of Subsidized Educational Establishments (SNED)

One of the earliest experiences with performance pay in Latin America was Chile’s Sistema Nacional de Evaluación del Desempeño de los Establecimientos Educativos Subvencionados (National System for Performance Evaluation of Subsidized Educational Establishments) or SNED. The program was introduced in 1996 and offers group bonuses to schools every two years based on an index that combines student learning results with other indicators of school and teacher performance.

Student learning counts for 65 percent of the total score through a combined measure of a school’s results on the current year’s national assessment (37 percent) plus a value-added measure of the difference in its average scores over the past two cycles (28 percent). The other indicators include schools’ initiative (6 percent, based on school surveys); labor conditions (2 percent, based on presence of a complete teaching staff, replacement of absent teachers, and other indicators); equality of opportunities (22 percent, based on the school’s retention and passing rates and lack of discriminatory practices, also measured through survey data); and integration of parents and guardians (5 percent, also based on survey data).

SNED is conducted as a tournament. Once the SNED index is estimated, schools are ranked within homogenous groups. Schools in the top 25 percent of the ranking within each group receive the SNED award of excellence. Ninety percent of the bonus is distributed to teachers, while 10 percent is distributed by the principal. The bonus for teachers typically represents an annual increase equal to approximately 40 percent of a teacher’s monthly salary. About 20–30 percent of schools win the award.

The most recent evaluation of SNED estimated the impact of the incentive program by comparing its effects on privately managed, government-subsidized schools (about 30 percent of total enrollment) that are eligible for the incentive with results for privately unsubsidized schools, which are not eligible (Rau and Contreras 2009). Controlling for all other differences across these schools, the researchers found that the introduction of SNED in 1996 stimulated a statistically significant 0.12 standard deviation
improvement in average student learning outcomes in the eligible schools relative to schools not covered by the program. Rau and Contreras (2009) also explored the effects of winning the SNED award on subsequent school performance. Over six rounds of SNED awards, however, they found no statistically significant evidence that schools winning the award performed better in the next period. They did, however, find that although SNED’s stratified “homogenous group” performance bands are designed to help equalize schools’ chances of winning the bonus, almost 40 percent of schools have not yet done so.

**Bonuses Based on Teacher Input**

Two rigorously evaluated programs, in India and Kenya, have explored the impact of bonus pay for teachers based on attendance.

*Kenya: Pre-school teacher bonus program*

Kremer and others (2001) evaluated a program that allowed school headmasters in rural Kenya to award individual teachers bonus pay for regular attendance. The size of the bonus was large—up to three months’ salary for no absences. They found that the program had no impact on actual teacher attendance (measured by unannounced random visits). Absence rates remained at 29 percent. There was also no evidence of change in teachers’ pedagogy, pupil attendance, or pupil test scores, although it could be argued that both pedagogy and test performance at the preschool level may be noisy measures.

Researchers also found that headmasters simply distributed the full bonus to all teachers regardless of attendance. Even though there was a financial incentive for headmasters to hold back part of the funding (any funds not allocated to teacher bonuses reverted to the schools’ general fund), they chose not to do so. School headmasters clearly found it difficult to play a strict monitoring role at the school level.

*Rajasthan, India: “Camera” program*

An innovative program in rural India produced very different results (Duflo, Hanna, and Ryan 2010). In a randomly selected set of rural, NGO-run schools in Rajasthan, a schedule of monthly teacher bonuses and fines based on attendance was monitored in a creative way—with daily date- and time-stamped photographs. A student was asked to photograph the teacher and the children in the class at the beginning and end of each school day. Teachers’ salaries were a function of the number of “valid school days,” in which the school was open for at least five hours and at least eight students appeared in each picture. Unannounced visits to the “camera” and comparison schools measured actual absence rates and observed teacher activity.
The maximum bonus for a teacher with no days absent was approximately 25 percent of a month’s salary. The program over three years had a dramatic effect on teacher absenteeism, which fell from 42 percent to 23 percent in the treatment schools. While there were no observed changes in teachers’ classroom behavior and pedagogy (other than greater presence), student test scores rose by a 0.17 standard deviation, and graduation rates to the next level of education also rose significantly. While students’ attendance rates (conditional on the school being open) did not increase, there was a significant increase in the total amount of time children in treatment schools spent in classes: on average, 2.7 more days of schooling per month.

Duflo, Hanna, and Ryan (2010) noted that the intervention was quite cost-effective. The base salary for teachers in the treatment and comparison schools (Rs 1,000 per month) was the same. All of the other costs of the program (the bonuses, the cameras, and monitoring) totaled roughly $6 per child per year. In terms of raising test scores, the per-child cost of a 0.10 standard deviation increase in test scores was only $3.58.

These contrasting experiences suggest that it is possible to stimulate higher teacher attendance with bonus pay, but the credibility of the system for monitoring performance is important. At least in the rural India setting—a context of very high teacher absence—teacher attendance also appears to be an important correlate of the desired education outcome: student learning. In other high-absenteeism settings, or settings where standardized annual student learning data is not available, bonus pay linked to teacher attendance is a reasonable approach, although it would be good to have additional evaluation evidence confirming its impact on learning.

Pay-for-Performance Programs: The Balance of Evidence

The most recent and robust developing-country evidence on pay-for-performance programs suggests that bonus pay incentives can improve learning outcomes, at least in the contexts studied most carefully to date. This evidence is in contrast to the more mixed, but less rigorous, developing-country evidence that existed just five years ago. It is also in sharp contrast to the most recent evaluation evidence from U.S. programs. In carefully conducted randomized trials of relatively generous bonuses aimed at both individual teachers (Nashville public schools) and schools (group-based bonuses in New York City public schools), researchers have failed to find any impact on student learning outcomes.

Under the aegis of the National Center on Performance Incentives, an impressive number of new U.S. education pay-for-performance programs are being evaluated experimentally. A three-year randomized trial of an individual bonus for 297 math teachers in Nashville public schools that offered large bonuses (up to $15,000 per year or 400 percent of a teacher’s monthly wage) found no difference in average learning outcomes among
students of the teachers who were and were not eligible for the bonuses. The first-year results from a randomized trial of a school (group-based) bonus program in 323 schools in New York City, which also offered a relatively large award (up to $3,000 per staff member), have also shown no student test-score differences between treatment and comparison schools. However, in the New York case, the tests were administered only three months after the program was announced, and the researchers note that it is logical to expect that impacts from incentives may take more time to develop.

In no study to date have long-term effects of performance-based pay been analyzed. Both theory and experience with performance-based rewards in other sectors indicate that the scope for perverse behaviors, such as gaming, cheating, or teaching to the test, can rise with time as system actors become more familiar with the rules of the game. As performance-based pay becomes increasingly—and logically—linked to student test results in many countries, the validity of those tests and the legitimacy of their application become centrally important challenges for education systems.

In a context of persistently low education outcomes and widespread evidence of “accountability failures” on the part of teachers and other education system actors, the evidence that pay-for-performance programs and the use of contract teachers can raise student outcomes in developing-country contexts is important. But the contrasting U.S. evidence suggests that it is important to note that these developing-country contexts are characterized by

- **Weak systems for performance monitoring and accountability**—evidenced by relatively high teacher absence rates, low teacher dismissal rates, and low student learning performance
- **Relatively weak teacher professionalism**—evidenced, in most cases, by low standards for entry
- **Relatively large bonus size**—for example, an annual bonus equaling 30–300 percent of a month’s salary
- **Focused performance metrics**—emphasis on a small number of key, measurable results, notably student learning improvements or relatively easily measured teacher “inputs” such as monthly attendance, rather than more complex, subjective, and comprehensive performance evaluations
- **“Fair” performance metrics**—rewards to schools on a value-added basis (for progress relative to their starting point) or compared with schools with similar geographic and student socioeconomic conditions, not for absolute levels of performance
- **Rewards clearly linked to prior-period results**—annual bonuses directly linked to results for the previous school year, such as school-level learning improvement, or monthly bonuses for input measures monitored over the previous month, such as attendance.
These pay-for-performance programs in developing countries have “worked” in the sense that student learning outcomes improved in the presence of the bonus. In the most careful studies, the size of the effect—a 0.19–0.27 standard deviation increase in average student learning—is impressively large compared with the effects typically measured for other types of education programs.

The sole developing-country evaluation to date designed as a “head-to-head” comparison of individual bonus pay (rewarding each teacher for his or her own classroom’s average learning progress over the course of a school year) with “group” bonus pay (rewarding schools for their average learning improvement) showed similar results in the first year but a stronger effect on learning outcomes from the individual bonus by the second year. The impact measured in that program in rural Andhra Pradesh, India—a 0.27 standard deviation increase in language scores—remains the largest reported causal impact from an education pay-for-performance program. However, the group bonus alternative proved to be more cost-effective because the average amounts awarded were smaller. In general, school systems will likely find group bonus pay more technically feasible than individual bonus pay, which requires the ability to test students in every grade, subject, and classroom and thus presents significant technical challenges and costs.

Our understanding of the mechanisms through which bonus pay improves student outcomes is still weak. In several randomized trials, at least over the short term, the bonus program did not induce any reduction in teacher absence rates, which is one of the most obvious ways teachers can increase their efforts in response to an incentive. These teachers did nonetheless produce statistically significant improvements in their students’ learning outcomes relative to comparison groups of teachers who were not offered a bonus. Most did so by increasing the intensity of their work during school hours (assigning more homework and class work) and conducting test preparation sessions outside of school hours.

In Brazil, where all programs to date have been implemented system-wide, it is not possible to estimate how the introduction of a bonus per se affects schools’ performance. At the end of the day, this is the critical question for policy makers: how cost-effective is the bonus program as a whole relative to alternative uses of education funds? But the Brazil studies can help elucidate both the pathways through which bonus incentives can change teacher behavior and the kinds of changes that are most effective in improving learning. There is intriguing evidence from Pernambuco that more-ambitious targets stimulated larger increases in student learning than less-ambitious targets did in comparable schools. This evidence suggests that, in the presence of an attractive performance award, schools focus on and are motivated to try to achieve specific targets. The Brazil research has
also generated evidence of more efficient use of instructional time by teachers in the schools that subsequently earned the bonus. Teachers in bonus-earning schools lost less time due to late arrival or early departure, spent more time on instruction, made more intensive use of classroom resources such as learning materials, and kept a higher share of students engaged in learning.

**Designing Teacher Accountability Reforms**

What can recent experience tell us about the design of effective teacher incentives? How does the latest developing-country evidence square with issues that have long been associated with pay-for-performance programs, both in theory and in practice? In the United States alone, there is a long history of merit pay efforts in the education sector, and few have survived (Murnane and Cohen 1986). In this section, we review the theory on performance incentives and try to unpack the evidence from recent program experience to generate practical guidance for policy makers.

**Principal-Agent Theory**

Many of the issues in designing effective performance contracts transcend the education sector. As discussed in chapter 1, “principal-agent” (or employer-employee) relationships are a central topic of economics and industrial relations research because while employers (principals) need employees (agents) to help achieve organizational objectives, the two parties have divergent interests. Employees want to maximize compensation and minimize effort, and employers want the opposite. The two parties also have asymmetric information: the employer cannot perfectly monitor the effort and activities of the employee.

Under these circumstances, the role of the contract between the principal and the agent is to align their objectives by specifying the activities and results wanted and the compensation offered for these. An effective contract will motivate the agent to focus his or her efforts on efficiently achieving the principal’s objectives. On the other hand, if the contract is structured so that agents are paid a flat rate irrespective of their level of effort or ability, or if job tenure is guaranteed irrespective of performance, it is unlikely that employees will exert additional effort or focus on doing the things that matter to the principal. It is striking how closely this textbook definition of an ineffective performance contract parallels the typical teacher contract (Prendergast 1999, cited in Courty and Marschke 2003).

Some issues in designing effective performance-based contracts in education are shared more broadly by public or nonprofit sectors. In contrast to
firms that operate in a competitive environment, public sector agencies often face difficulties in specifying performance goals. Even when they have clear objectives, they can have difficulty establishing a clear hierarchy among competing objectives—something that is straightforward for firms, which seek to maximize profits or shareholder value. Baker (2002) argues that “the difficulty in defining ‘good’ performance measures in nonprofit organizations is one reason for the weak incentives that so often characterize organizations of this type, and for the dysfunctional consequences that often arise when these types of organizations try to use strong incentives.” In addition, agents may not “know the technology” for achieving complex organizational goals. Vegas (2005), among others, has questioned whether teachers in many developing-country settings have the core content mastery and teaching skills required to produce desired learning improvements, no matter how strong their incentives to do so.

On the other hand, the assumption that agents work only for extrinsic (financial) rewards has come under scrutiny in both the psychology and economics literature in recent years. In their review of the economics and industrial relations literature on contracting, Fehr and Falk (2002) observed that “while it is certainly true that [agents’] desires to avoid risk and to achieve income through effort are important, it is equally true that there are powerful non-pecuniary motives that shape human behavior”—such as the desire to reciprocate, the desire for social approval, and the desire to work in interesting tasks.

Pink (2006) has argued that once a certain threshold level of financial remuneration is achieved in knowledge professions, the most powerful incentives are individual workers’ own desires for autonomy, mastery, and a sense of contribution. It is interesting to note that all three of these incentives figure prominently in the core “industrial model” of the school: Teachers enjoy substantial autonomy within their own classrooms. The process of intellectual mastery is at the core of the work, for both teachers and their students. And abundant survey and other literature documents the attraction of the education profession for individuals seeking to make a social contribution.

These different strands of the academic literature suggest two hypotheses about performance contracts in education.

Hypothesis 1: Other things equal, it may be easier to attract individuals into teaching than into equivalently remunerated professions

The experience of Teach for America and its recent Teach for All offshoots in European and developing countries provides some support for this hypothesis. These programs have found it possible to attract high-capacity individuals to teach in disadvantaged schools (at prevailing wages) by stressing the scope for social contribution and offering strong organizational
support for professional growth and mastery. However, the numbers of teachers recruited through these or parallel programs remain small relative to the system as a whole; after 20 years of experience, Teach for America placements represent only 7,000 of 3.2 million teachers in the United States.

The broader lesson for public sector school systems is the power of these levers. The “intrinsic” rewards of teaching—even if they are explicitly maximized by a well-managed school system—cannot substitute indefinitely for financial remuneration. But public education systems in developing countries likely have more scope than they are currently exploiting to incorporate appeals to the intrinsic motivation of prospective teachers into their human resources policies.

**Hypothesis 2: Agents’ (teachers’) performance in education may be enhanced by clearer expression of performance goals and feedback**

Research shows that public and nonprofit agencies in general have difficulty in specifying or prioritizing performance goals. In this context, improving the quality of performance feedback to teachers and the clarity of targets and goals, in theory, should enhance the efficiency of agents’ performance. This belief is an important driver of the standards and accountability-based reforms in education seen in many OECD countries (such as Australia, New Zealand, the United Kingdom, and the United States) and some of the developing-country cases discussed in this chapter.

**Measuring Performance in Education**

The growing availability of nationally and internationally standardized data on student learning results over the past 20 years has reshaped the landscape on performance measurement in education systems. Despite deep controversies over the quality, credibility, and implementation of specific tests or testing in general, there is broad acknowledgment that learning outcomes are appropriate metrics for school system results. Education systems today in most countries are more advanced than other parts of the public sector in being able to track a meaningful indicator of system progress on an annual basis and, often, to be able to disaggregate results for specific subnational regions and target groups.

This advantage has clear implications for the design of effective incentive programs in education. Much of the theoretical literature on principal-agent contracts centers on the challenge of identifying an appropriate performance measure that will align the employer’s goals with the employee’s efforts. Not only must the measure be appropriate and meaningful, it must also be something that can be adequately measured—meaning regularly, inexpensively, and credibly. The scope for gaming (efforts by the employee
to set artificially low targets), outright cheating, and other perverse behaviors must be managed. Psychologists argue that even valid and informative performance measures have a tendency to “degrade” or become dysfunctional by the mere act of being used for incentive purposes (Darley 1991, cited in Baker 2002).

None of these challenges is easy. But it is increasingly evident that, in the context of education, a core performance measure in most incentive programs will be student learning outcomes. Thus, the sheer identification of a meaningful and appropriate performance measure—which presents major problems for many public sector and nonprofit actors—is less of a challenge in education. This does not mean that learning outcomes are, or should be, the only performance measure used by school systems. There are clearly other important goals (building children’s self-esteem, appreciation for culture and the arts, physical development, citizenship, and so on), and concerns are often expressed that these goals are shortchanged by excessive focus on learning outcomes. School systems must manage this tension. But what seems increasingly accepted is that a central task for every 21st-century school system is ensuring that well-designed and legitimately managed testing systems to track student learning progress are in place.

While high validity, well-administered testing programs are important for education systems irrespective of incentives, there is evidence that the stakes around test legitimacy and the scope for abuses become higher when incentives are introduced (Jacob and Levitt 2002). Rivkin (2009) recently articulated this point well:

Over the long run, the success of [accountability and pay-for-performance] reforms hinges on a number of factors, but the validity of the teacher effectiveness (a term I use interchangeably with teacher quality) measures is one of the most important. Unless the quality estimates are accepted as informative and fair, opposition will remain strong. Moreover, inaccurate or inadequate quality measures and poorly designed pay-for-performance programs will introduce adverse incentives to teach narrowly to test content, concentrate on only a fraction of the students, forgo important non-tested outcomes, or elicit frustration and distrust in response to reward structures with systemic flaws.

**Design Elements of Teacher Incentives: A Preliminary Typology**

Our review of 15 different teacher incentive programs implemented in developing countries and Israel over the past decade offers an opportunity to analyze some of their design features. In doing so, we try to move beyond the basic taxonomy most commonly used in the education literature—for example, group or individual; input-based or outcome-based, and so on. In focusing on additional aspects of incentive design that theory predicts are
important for effective performance-based contracts, we also hope to create a framework that can be tested as the evidence from these experiences accumulates. Because the current set of cases remains too small for any definitive conclusions, this is an exploratory exercise only.

Based on the evaluation evidence reviewed, three key dimensions may predict an incentive program’s effectiveness: controllability, predictability, and bonus size. We assess these features of different programs and examine the correlations with available empirical evidence of program impact.

Transcending these features is something that might be termed the legitimacy or credibility of the incentive program as a whole. Achieving this legitimacy involves, in Rivkin’s (2009) words, ensuring that the specific performance measures used are perceived as “informative” and “fair.” But legitimacy also speaks to the perception that all aspects of the program are managed honestly and transparently. Further, legitimacy implies an acceptance that the rules of the game will be stable over time or that adaptations will be undertaken with due process and consultation, rather than through ad hoc adjustments. Murnane and Cohen’s (1986) observation that the longest-standing pay-for-performance programs in the United States typically have been redesigned several times, with teacher input, suggests that sustaining credibility over time requires effort and processes that are perceived as legitimate.

We have argued that improved measurement of student learning outcomes in many countries has created an expanded political space for pay-for-performance programs because the technical platform—a relevant and meaningful source of data on schooling outcomes—has been established. A well-designed and cleanly administered national or subnational student assessment program could provide a ready anchor for pay-for-performance schemes. Brazil’s IDEB goes one step further by establishing a high-quality national data source on both student learning outcomes and grade progression. The combined measure mitigates at least one undesired teacher behavior that may be induced by a high-stakes focus on test scores: the risk that teachers trying to boost average student learning outcomes will hold back slower students. By combining the two measures of learning outcomes and grade progression, IDEB builds in protection against this form of distortion. The degree of legitimacy that Brazil’s high-quality national testing system and IDEB have achieved over the past five years appears to be a factor in the increasing adoption of pay-for-performance programs there. It is interesting to note that several earlier pay-for-performance programs in Brazil, based on individual states’ testing programs, had difficulty with credibility issues and were not sustained long.

Perceived legitimacy may be the single most important factor in the impact and sustainability of teacher incentive programs, but it is difficult to assess objectively and especially difficult to assess ex ante. Our typology,
therefore, focuses on more specific and tractable design features. However, it bears noting that the experimental design programs reviewed here—which constitute a strong part of the current evidence base on teacher incentives in developing countries—have special characteristics that positively affect their legitimacy. Given the context for teacher incentive reforms in most developing countries—a status quo of poor education results—the potential for undesired consequences is not per se an argument against teacher incentives so much as a call for eyes-open efforts to manage known risks. Some of these risks relate to specific choices in the design of incentives in the three areas we identified: controllability, predictability, and bonus size.

Table 4.4 shows the hypotheses regarding program effects that can be derived from the principal-agent model and our typology.

**Controllability in Incentive Program Design**

Controllability refers to how much of the measured outcome depends on actions taken by an individual teacher (or employee). Performance measures related to a teacher’s own behavior, such as attendance or training course participation, are directly dependent on the teacher’s actions. Awards based on such measures create incentives with a high degree of controllability from the standpoint of the employee or, conversely, low risks. Theory predicts that riskier contracts, based on less controllable factors, either demand a larger incentive (other things being equal) or will not be as effective.

However, while school systems could reward many behaviors that are highly controllable from a teacher’s perspective, these may not be the most important behaviors for producing the results school systems desire most. The problem for incentives design is that teacher behaviors that contribute to student learning improvement are more difficult to specify, observe, and measure directly across an entire school system. On the other hand, it is easy to see why basing incentive programs on system results, especially student learning outcomes, is often resisted by teachers; these outcomes are less controllable—that is, riskier for the agent.

The sources of risk can be unpacked into at least two dimensions that policy makers need to consider and manage: noise and locality.

*The noise factor*

“Noise” refers to the degree to which the specific performance measure is affected by factors outside of the agent’s control. Student learning results for a particular classroom, for example, cannot be completely attributed to a teacher; they are also influenced by noise from other variables such as students’ prior grade preparation, socioeconomic level, or distractions.
on the day of the exam. A specific test instrument can be noisy if it is not well mapped onto curriculum goals or fails to discriminate between students who have and have not mastered the content. The noisier a performance measure, the less responsive it is to the agent’s behavior and effort. Even on the same test instrument, cross-sectional student scores are noisier than longitudinal (value-added) measures that can identify

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controllability</td>
<td>Noise The degree to which factors other than the teacher’s behavior (for example, student characteristics or the actions of other teachers) can affect the performance measure, therefore making it a less credible and adequate measure of a teacher’s individual contribution</td>
<td>The less noise in the measure, the more controllable it is and the more effective the incentive will be (at least in the short term).</td>
</tr>
<tr>
<td></td>
<td>Locality The degree to which an individual teacher’s performance affects the result relative to other actors being incentivized. Measures based on individual performance (group size = 1) are maximally local. If group performance is being measured, as group size grows larger, the measure becomes less local.</td>
<td>The more local the measure, the more controllable it is and the more effective the incentive will be.</td>
</tr>
<tr>
<td>Predictability</td>
<td>Predictability An individual teacher’s likelihood of achieving the bonus</td>
<td>If probability of earning the bonus is close to either zero or 100%, the incentive will not be effective. Piecewise formulations (which allow workers to earn some bonus in line with their production) will create stronger incentives than tournament-style awards (which set an ex ante cap on the share of schools or teachers that can win).</td>
</tr>
<tr>
<td>Bonus size</td>
<td>Bonus payment as a proportion of monthly wages</td>
<td>The larger the bonus, the more effective the incentive will be, all things being equal. But larger bonus size can also stimulate cheating and other undesirable behaviors and can make bonus programs fiscally unsustainable or less cost-effective.</td>
</tr>
</tbody>
</table>

Source: Authors’ compilation.
learning gains for individual teachers’ classrooms. Finally, test scores for small classes are inherently much noisier than for larger classes (Kane and Staiger 2001).16

School systems can reduce the noise involved in test scores by grouping schools into comparative strata on socioeconomic grounds (as in Chile’s SNED) or using school-level targets (as in Pernambuco and other Brazilian states) that measure each school’s progress relative to its own baseline. A noteworthy feature of many of the new bonus programs is sensitivity to this issue, with built-in adjustments for schools’ baseline performance levels or performance comparisons for schools grouped by socioeconomic and geographic strata.

\textit{The locality factor}

“Locality” refers to the scope for an individual teacher to influence the results being rewarded. If the result is produced by a group of teachers or schools, it will be less controllable from the standpoint of an individual teacher because the teacher’s actions will contribute only a fraction to the result. As the group gets larger, the individual’s contribution gets smaller, and he or she is less able to control the final outcome. Controllability is higher (all things being equal) in highly local measures, such as classroom-level learning outcomes or individual contract teachers’ attendance records.

An important issue with group rewards (typically schoolwide rewards, in the case of education) is not only that they are less controllable from the perspective of an individual teacher but also that they can motivate free-rider behavior. Free riders are members of a group who refuse to exert effort while counting on others to produce the group result (Holmstrom 1979). In large schools, where an entire staff’s achievement of a school-level bonus may depend on the results of a small handful of tested subjects and grades, free-riding can become an important issue. If teachers perceive that others are free-riding on their efforts—that is, the locality and thus controllability of the incentive is lower—the effectiveness of the incentive will be weaker.

Does this mean that education incentives should be focused on outcomes at the teacher level? More sophisticated testing data in developing countries is opening the possibility of measuring classroom-level learning gains for individual teachers, as is now possible in the United States and other OECD countries. Rewarding individual teachers can provide strong work incentives and negate free-riding. But on the other hand, individual rewards can discourage collaboration and teamwork to the detriment of overall school results. They also create incentives within a school for teachers to avoid weaker students. This can complicate the work of principals; matching the more effective teachers with more difficult students might be
in the best interest of those students and the school’s overall results, but it would impede those teachers’ ability to earn fair rewards for their talent and effort.

The question for policy makers is whether focusing incentives at the school level creates benefits—in terms of stimulating collaborative work among school personnel—that outweigh the free-rider concerns. Hybrid approaches (which reward school-level outcomes but adjust the award received by individual employees within the school for some measure of their individual effort, such as attendance) are an interesting alternative. Most of the experimental teacher incentive programs in developing countries to date have been in settings with very small schools (three to five teachers, on average). Therefore, the new wave of research on the Brazilian programs (where urban schools can have 40 or more teachers) that is focusing research attention on the measurement of free-rider behavior may be a useful contribution.

**Predictability in Incentive Program Design**

For an incentive to be effective, agents must perceive realistic scope to achieve it. If the likelihood of achieving a bonus is close to zero, teachers will have little incentive to exert effort. On the other hand, teachers with near certainty of receiving a bonus will also have little incentive to increase effort. Theory predicts that if performance-based contracts result in awards distributed to a share of workers nearer the midpoint of 0 and 1, rather than at the extremes, they will elicit the most effort. Dampening or enhancing this effect, however, would be the agents’ assessment of controllability or risk and the amount of payoff expected.

For incentive programs that include sanctions—such as the prospect of contract nonrenewal for poorly performing contract teachers—the strength of the incentive is proportional to the likelihood of enforcement.

**Bonus Size in Incentive Program Design**

Theory predicts that larger bonuses elicit stronger responses, conditional on the degree of risk inherent in the contract, the chance of earning the bonus or being sanctioned, and the cost of the effort. Even if teachers are intrinsically motivated, incentive contracts must still offer payoffs that exceed the marginal cost of effort, although this cost might be lower than it would have been for less intrinsically motivated individuals. The incentive power of a given financial bonus is, of course, relative to the base wage and price levels in a given country context. Therefore, we attempt to standardize bonus size in relation to average monthly wages. We find significant divergence across programs, with annual bonuses ranging from 35 percent of a
monthly wage (Andhra Pradesh, India) to 150–200 percent (Brazil) and even to 300 percent in one of the programs in Israel.

Table 4.5 categorizes the incentive programs reviewed for which there is some impact evidence. Categorizations of this kind always involve a degree of subjectivity, and the small number of programs makes this typology preliminary and illustrative only. But this framework and the associated impact evidence may assist policy makers in thinking through the design features of new pay-for-performance programs. The impacts presented are typically for student learning outcomes, expressed as the average improvement in learning for the treated schools or classrooms compared to control groups, as a proportion of one standard deviation in the test score distribution. Our ratings of program design elements are on a standard Likert-type scale, from 1 (very low) to 5 (very high). The annex at the end of the chapter explains in further detail how the ratings for each feature were assigned.

Core Design Features

The strength of a program’s core design features (controllability and predictability) is, to some extent, complementary to the financial size of the bonus. In other words, an incentive program with high controllability and predictability—such as an individual incentive that teachers have a reasonable chance of attaining and that is based on student learning progress on a well-designed test—will induce stronger performance than a program offering an equivalent financial reward with lower controllability and predictability. The implication for policy makers is that what matters is the balance of these factors. The fiscal costs of an incentive program can to some extent be reduced by attention to the complementary design features of controllability and predictability. To put it another way, for every pay-for-performance program, there is an optimal design that bundles these features to stimulate the maximum performance response from teachers and schools at minimum fiscal cost.

We explore this hypothesis in table 4.5 by averaging the ratings for a program’s noise, locality, and predictability features into a separate score, labeled “core design” (in column 4)—which is a simple average of the three preceding subscores. We compare programs’ scores on the strength of their core design with the size of the bonus. We then generate a combined measure that averages the scores for all of the program design features, including the size of the bonus. This overall score is the program’s “predicted overall incentive strength.” We compare this predicted incentive strength with the observed impacts that the different bonus programs had on desired educational outcomes.

In table 4.5, the programs are ordered by the strength of their “predicted overall incentive strength,” from lowest to highest. Comparing
Table 4.5 Pay-for-Performance Programs by Core Design Features and Effect Size
Likert-scale ratings (1–5) except where otherwise specified

<table>
<thead>
<tr>
<th>Program</th>
<th>Noise</th>
<th>Locality</th>
<th>Predictability</th>
<th>Predicted Strength, Core Design</th>
<th>Bonus Size</th>
<th>Average Bonus Value (% MW)</th>
<th>Predicted Overall Incentive Strength</th>
<th>Observed Effect Size (Max.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Israel (group) (Lavy 2002)</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>2.7</td>
<td>2</td>
<td>40</td>
<td>2.5</td>
<td>0.13</td>
</tr>
<tr>
<td>Chile SNED (group) (Rau and Contreras 2009)</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>40f</td>
<td>2.8</td>
<td>0.12</td>
</tr>
<tr>
<td>Brazil (group) (Ferraz and Bruns, forthcoming)</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>180</td>
<td>3.5</td>
<td>0.31–0.57g</td>
</tr>
<tr>
<td>Kenya (group) (Glewwe, Illias, and Kremer 2010)</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>43</td>
<td>3.5</td>
<td>0.14</td>
</tr>
<tr>
<td>India (group) (Muralidharan and Sundararaman 2009)</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4.3</td>
<td>2</td>
<td>36</td>
<td>3.8</td>
<td>0.16</td>
</tr>
<tr>
<td>India (individual) (Muralidharan and Sundararaman 2009)</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4.7</td>
<td>2</td>
<td>36</td>
<td>4</td>
<td>0.27</td>
</tr>
</tbody>
</table>

(continued next page)
Table 4.5 Pay-for-Performance Programs by Core Design Features and Effect Size (continued)

<table>
<thead>
<tr>
<th>Program</th>
<th>Noise</th>
<th>Locality</th>
<th>Predictability</th>
<th>Predicted Strength, Core Design</th>
<th>Bonus Size</th>
<th>Average Bonus Value (% MW)</th>
<th>Predicted Overall Incentive Strength</th>
<th>Observed Effect Size (Max.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>India (attendance) (individual) (Duflö, Hanna, and Ryan 2010)</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4.7</td>
<td>2</td>
<td>30</td>
<td>4</td>
<td>0.17</td>
</tr>
<tr>
<td>Israel (individual) (Lavy 2009)</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>4.3</td>
<td>5</td>
<td>300</td>
<td>4.5</td>
<td>14% higher pass rates and 10% higher test scores</td>
</tr>
</tbody>
</table>

Source: Authors’ compilation.

Note: MW = monthly wage. SNED = National System for Performance Evaluation of Subsidized Educational Establishments (Sistema Nacional de Evaluación del Desempeño de los Establecimientos Educativos Subvencionados).

a. Some programs used multidimensional performance measures. In these cases, each measure was rated separately and then averaged out.

b. Average group sizes are as follows: Brazil: 43 teachers; Chile: 18 teachers (figures from 2004); India: 3 teachers; Israel: 80 teachers; and Kenya: 6 teachers.

c. Predictability was based on actual award rates as follows: In Brazil, 51 percent of schools got the bonus in 2009. In Chile, 20–28 percent of schools got the awards (est. 2004). In the India programs, most teachers under all programs got some award. In Israel (both programs), about 33 percent of teachers got awards. In Kenya, about 48 percent of participating schools got the award.

d. See annex for description of Likert scale used for bonus size.

e. Effect size typically represents the increase in learning caused by the program expressed in terms of standard deviation. For the Israel group incentive program, the authors converted Lavy’s (2002) reported impact into standard deviations, drawing on descriptive data presented. For the Israel individual bonus, this was not possible, so the outcomes presented are not strictly comparable to the others in the table.

f. Average award during the initial years of the program, which was the period evaluated. More recently, the average award is 70–80 percent of MW.

g. Impact estimates for Pernambuco, Brazil, are not rigorously estimated and thus not directly comparable to other estimated program impacts.
the scores for “core design” features with the bonus size does suggest some complementarity. Except for the 2002 individual bonus program in Israel, in all cases where programs’ core design features rated a (strong) score of 4 or higher, the size of the bonus adopted was relatively low.

However, programs with core design features rated lower than 4, such as Chile’s SNED, the Brazilian bonus program, and the Israel group incentive, show no clear pattern. In all three cases, the performance measures on which the bonus is based are rated highly for being designed in a way to minimize noise; Chile’s SNED pays careful attention to schools’ socio-economic contexts, and Brazil’s target-based system effectively focuses each school on improvement relative to its own prior-year performance. But locality and predictability ratings are relatively low because these bonuses are school-based rewards and typical school size is relatively large. In the Israel and Chilean cases, moreover, the tournament-style programs reward a relatively small share of schools.

In the Brazil case, a relatively large bonus size (average awards more than one month’s salary) compensates for the relatively weaker incentives created by these core design features. It is interesting to note that Chile’s SNED program has increased the average bonus size significantly over the years as well—from an initial 40 percent to a current 80 percent of the average monthly teacher wage. The Israel group bonus program appears exceptional in this regard. It is the only program in which the bonus size does not compensate for a relatively low score on core design features; instead, the relatively small maximum bonus further lowers the predicted strength of the incentive.

Both of the Israel programs, in fact, are striking in that they are the only two programs where the bonus size does not exert a moderating effect on the strength of the “core design” but instead exacerbates it—making a rather weak incentive program weaker in the case of the group bonus and making an already strong incentive stronger in the case of the individual design. One might ask what is wrong with making a strong design stronger. In the context of our comparators, the answer appears to be that it makes a bonus program more expensive than it needs to be—an important concern from a policy standpoint.

The overarching question is whether the “predicted strength” of a bonus program based on these design features is correlated with its impact. All of these programs were associated with positive outcomes and, in many cases, causality has been rigorously established. But the small number of programs and the limited years of observation for most of them make the comparisons here illustrative only. Figure 4.2 provide a simple visual mapping of how program impacts compare with estimates of predicted incentive strength.
Programs are presented below the axis along the scale derived from table 4.5, which predicts the overall strength of the incentives. Note that the scale is purely illustrative. Programs are presented above the axis in terms of the reported size of their effects on student outcomes. This continuum ranges from low effects (less than 0.1 of a standard deviation) to high effects (close to 0.3 of a standard deviation) relative to those usually seen in the economics of education and development literature. The Brazil program is not included here as it does not yet have equivalently rigorous impact estimates. The estimated impacts on student learning range from a 0.12 standard deviation for Chile’s SNED to a 0.27 standard deviation for the Andhra Pradesh, India, individual bonus program by the second year of implementation.

A few cautious observations may be made. First, these pay-for-performance programs have had positive effects on student learning, of roughly a 0.15 standard deviation. In comparison with the evaluation evidence on other types of education interventions, it is noteworthy that these effects are statistically significant and consistently positive across different bonus program designs and very different country contexts. But the effects
are not huge. The one exception is the second-year impact observed in the Andhra Pradesh, India, program, which offered schools individual teacher incentives. It stands out, having produced a very large 0.27 standard deviation increase in average math and language learning outcomes compared with randomly selected schools that were not eligible for an incentive.

Second, according to the metrics constructed for predicting the strength of a pay-for-performance incentive program, most of the programs cluster fairly closely around a rating of between 3 and 4 in our schema. The two outliers are the Israel programs. Across the entire set of programs, however, there does not appear to be any obvious correlation between the “overall incentive strength” of a program and its impact—at least in our simple ratings scheme and for this small sample of programs. In connection with our earlier observation that strength of these programs’ core design features tended to be dampened or balanced by the decisions on the average bonus size, there is an intriguing suggestion that there may be some kind of threshold incentive strength for a pay-for-performance program to “work.” However, beyond this point, there may not be any payoff to making programs more generous. At least in the two Israel cases, there is no evidence that the considerably stronger and more expensive individual bonus scheme produced better results than the “weaker” group bonus program.

These observations, however, are highly speculative. The number of evaluated programs is small, and most of the impact data come from two years or less of program experience. Both positive and negative features of performance-pay programs develop with experience and can exert potentially strong effects on their impact. On the positive side, agents’ awareness of programs and the experience of actually seeing benefits paid out will increase the credibility of the incentive and presumably its strength. Indeed, interviews with teachers and school directors in the first year of Pernambuco’s bonus program showed that schools were not universally aware of their performance targets and had quite uncertain expectations about the potential size of performance bonuses. In the second year of the program, awareness was, as expected, much higher.

On the other hand, schools’ and teachers’ ability to “game” the system also rises as their understanding of the rules of the game deepens with time. Evidence of gaming, cheating, scandals in test administration or the other performance measurement processes, or general perceptions of unfairness in the mechanics underlying a bonus program will undermine its credibility, in turn dampening its incentive effects.

There is much to be gained from additional, careful evaluations of new pay-for-performance programs in education—perhaps especially in developing countries, where the challenges of improving teachers’ performance are most acute. The framework presented here is simple, but it is grounded
in key elements of principal-agent theory. It appears to map well onto the types of pay-for-performance programs being developed in experimental pilot studies as well as systemwide education reforms. If this framework can assist policy makers in thinking through key design issues and encourage the systematic analysis of future programs, it will serve a purpose.

**Summary and Conclusions**

Teachers and school-level administrators are the front-line providers of education services. Their work determines the quality of services delivered and their results—above all, how well students learn. New research shows convincingly that student learning outcomes drive the overall benefits from education investments. If rising access and schooling attainment do not result in higher learning levels on globally benchmarked assessments, national investments may be largely wasted.

Recent research has also established that individual teachers vary widely in their ability to help their students learn. This evidence is leading countries to reexamine how they select, groom, and motivate teachers. It is also raising awareness of disconnects between the incentives teachers face and the results school systems want. Most school systems recruit teachers based on criteria that are not correlated with real effectiveness in the classroom, fail to reward the capacities and behaviors that are correlated with effectiveness, and are unable to sanction teachers who do not perform. A disconnect between teacher incentives and accountability for performance occurs in OECD countries as well as in low- and middle-income countries. But the failures are deepest in the developing world, manifested in high teacher absenteeism and persistently low learning results by international standards.

Countries are not standing still, however. Innovative, sometimes radical, reforms of teacher contracting and compensation policies are springing up in developing countries—reforms aimed at linking teacher incentives more closely to performance. An encouraging number of reforms are being rigorously evaluated. Although the number of country cases is still small and most programs are still in the initial years of implementation, evidence is beginning to accumulate about their impact.

This chapter reviewed the experience with two types of policies that can make teachers more accountable for results: contract tenure reforms and pay-for-performance reforms. Both have the advantage of overcoming the rigidity of existing teacher policies without requiring wholesale reform. Alternative contracting establishes different rules of the game for a new teacher cadre that works alongside the existing teacher stream. The new
teacher cadre is hired on fixed-term (usually one-year) contracts without the job stability that regular teachers enjoy.

Pay-for-performance programs leave core salary policies intact but create an incentive at the margin with the offer of an annual bonus based on some measure of teacher performance—be it an input measure, such as teacher attendance; an outcome measure, such as school or student results; or a combination.

The most rigorous of the seven available evaluations of contract teachers all found them to be more cost-effective than regular civil service teachers, and in India dramatically so. In both Kenya and India, randomized trials have found learning outcomes for students of contract teachers to be equal to or better than those of civil service teachers, despite contract teachers' much lower salaries. Nonexperimental studies in two additional states in India have found similar results. Earlier evidence on community-hired teachers in Central America (not included here but summarized thoroughly in Vegas 2005) was less robust but also suggested that contract teachers achieve similar or better student grade progression and learning outcomes (controlling for student background) at lower cost.

Although contract teachers usually work for lower salaries than their civil service counterparts, the cost-effectiveness of a contract teacher policy is likely to depend on country characteristics and the level of education involved. All of the evaluated cases involved contract teachers at the primary level, for example, where the supply of potential teachers with adequate skills is not as likely to be constrained as at the secondary level, or for specialty subjects such as sciences and math. Second, there are nagging questions about the sustainability of this policy over time. Many of the evaluated cases suggest that contract teachers may accept the lower salaries and insecure tenure because they are queuing for civil service positions. Teachers’ unions have also aided contract teachers in some African countries to press successfully for tenure and, in many cases, a large share do end up entering the civil service.

Nonetheless, the new wave of evidence on the short-term impacts of contract teacher reforms is fairly strong and consistent: the use of contract teachers can strengthen the scope for local monitoring of teacher performance, which results in higher teacher effort, which produces equal or better student learning outcomes than for regular teachers. And all of this is achieved at lower costs per student.

Teacher bonus initiatives in developing countries have proliferated over the past several years, both as small-scale experiments and as high-profile, systemwide reforms. In contrast to the situation just five years ago, and in contrast to recent U.S. evidence, a growing body of developing-country studies suggests that bonus pay incentives can work—at least in contexts characterized by
• **Weak systems for performance monitoring and accountability**—evidenced by relatively high teacher absence rates, low teacher dismissal rates, and low student learning performance

• *Relatively weak teacher professionalism*—evidenced in most cases by low standards for entry

• *Relatively large bonus size*—for example, an annual bonus of 30–300 percent of monthly salary

• **Focused performance metrics**—emphasis on a small number of key, measurable results, notably student learning improvements or relatively easily measured teacher “inputs” such as monthly attendance, rather than more complex, subjective, and comprehensive performance evaluations

• *“Fair” performance metrics*—rewards to schools on a value-added basis (for progress relative to their starting point) or compared with schools with similar geographic and student socioeconomic conditions, not for absolute levels of performance

• **Rewards clearly linked to prior period results**—annual bonuses directly linked to test or other results for the previous school year or monthly bonuses for input measures monitored over the previous month, such as teacher attendance.

These programs “work” in the sense that student learning outcomes improve in the presence of the bonus. Across the eight most carefully evaluated cases, the bonus program raised average learning outcomes in incentive schools relative to control schools by about a 0.15 standard deviation; in the highest case, learning outcomes by the second year of the program were a 0.27 standard deviation higher. Although not huge, effects on this order of magnitude are relatively rare across other types of education interventions, and the consistency of positive impacts, across a wide variety of country contexts, is noteworthy.

Our understanding of the mechanisms through which bonus pay improves student outcomes is still weak, however. In Brazil, classroom observations found evidence of more efficient use of instructional time by teachers in schools that subsequently earned the bonus. Teachers in these schools also lost less time due to late arrival or early departure, made more intensive use of learning materials, and kept a higher share of students “on task.” These patterns are consistent with Lavy’s observations in Israel. But in some randomized trials in Kenya and India, at least over the short term, the bonus program did not induce any reduction in teacher absence rates, which is one of the most obvious ways teachers can increase their effort in response to an incentive. These teachers did nonetheless produce statistically significant improvements in their students’ learning outcomes relative to comparison groups of teachers who were not offered a bonus; the likely
Making Teachers Accountable

channel is through offering extra homework and class work and conducting out-of-school sessions for test preparation.

In Brazil, where bonuses were available statewide, there is intriguing evidence that more-ambitious targets stimulated larger increases in student learning than in comparable schools with less-ambitious targets. This evidence suggests that in the presence of an attractive performance award, schools focus on and are motivated to try to achieve specific targets. However, the fact that under this program 50 percent of all schools earned the bonus during the first year, and 70 percent in the second year, suggests that the targets, overall, were relatively accessible.

Looking across pay-for-performance programs, several features derived from principal-agent theory appear important for the design of effective incentives. These include the controllability of the specific performance measure(s) being incentivized, from the standpoint of individual teachers; the predictability of the incentive (that is, what share of those eligible will achieve it); and bonus size.

Controllability, in our framework, is affected by two dimensions: “noise” in the performance measure and “locality” (that is, whether the bonus is an individual or group-based incentive). For group incentives, school size appears important. If a school has a large number of teachers, yet the bonus is based on results produced by only a handful of its teachers (typically, those teaching subjects and grades that are subjected to annual testing), there is room for free-riding—in other words, some teachers coasting on the work of the others.

The experience to date lends some support to the theory that core design features such as controllability and predictability affect the strength of the incentives in a pay-for-performance program. Particularly interesting from a policy standpoint is that these features appear complementary to the financial size of the bonus. In other words, bonus programs that pay attention to the design of performance measures that are perceived to reflect teachers’ work fairly (for example, student test scores for schools are adjusted for socioeconomic differences) and are reasonably predictable may (all other things being equal) achieve stronger impacts. Much longer-term and deeper analysis of experience with these and other education pay-for-performance programs is needed for any conclusive observations, but the experience to date at least suggests the usefulness of considering a set of core elements systematically in the design of new programs.

In conclusion, a growing number of school systems across the developing world are adopting two specific reforms aimed at strengthening the rewards and sanctions for teacher performance: the use of contract teachers and bonus pay linked to teacher performance. These programs have different designs, costs, and primary objectives, but both address a deep issue of weak incentives for performance in education systems across the developing world.
Our review of the best-evaluated cases to date shows that the design of effective incentives presents challenges and that the impacts vary significantly. An encouraging number of recent studies show positive impacts of incentive reforms on student learning outcomes. The number of developing-country reform experiences is still small, and most are fairly recent. But they permit a preliminary typology of examples that merit consideration by any education policy maker concerned with raising student learning as well as a framework for the generation of further research evidence on “what works” to make schools more accountable for results.

Annex: Rating the Design Features of Pay-for-Performance Programs

Controllability Ratings

We measure each of the two elements of controllability—noise and locality—on a 5-point scale, ranging from very low (1) to very high (5), as described below. In some cases, programs measure results on more than one indicator, such as test scores and matriculation rates or grade progression rates. In these cases, we constructed a simple average of the noise and locality ratings for each of the different elements.

Noise ratings

1. *(very low controllability)*: programs whose performance measures are likely to exhibit high noise (for example, programs using only cross-sectional learning data with no adjustments for schools’ socioeconomic context or other factors) and programs that measure performance based on outcomes for a very small group (for example, test scores calculated over fewer than 10 students)

2–4. *(moderate controllability)*: programs with measures less likely to exhibit noise

5. *(very high controllability)*: programs with performance measures designed to be as noise-free as possible—for example, programs using longitudinal data that are rich enough to produce value-added measures of teacher performance at the classroom level; programs that include adjustments for schools’ socioeconomic status and other contextual factors; programs with performance measures that are inherently controllable (such as seniority, attendance, and participation in professional development); and programs that avoid measuring performance of very small groups of students (less than 30) to reduce the potential for exogenous shocks or variation.
Locality ratings

1. *(very low locality)*: programs based on group performance in very large groups of more than 50 teachers
2. *(low locality)*: programs based on group performance in groups of 25–50
3. *(moderate locality)*: programs based on group performance in groups of 10–25
4. *(high locality)*: programs based on group performance in small groups of 2–10
5. *(very high locality)*: programs based on individual performance

Predictability Ratings

Predictability refers to the degree to which the agent or teacher can expect to earn the reward. A predictability rating thus ranges from 0 percent (total certainty that the bonus will *not* be attained) to 100 percent (total certainty it *will* be obtained). Theory predicts that if an individual’s chance of attaining a bonus is at either of these extremes, the motivation to perform will be weaker. In the one case, there is no point; in the other, there is no need. Not surprisingly, few bonus programs in reality operate at these extremes. Most—including all of the programs reviewed here—have rules of the game that ensure that somewhere between 20 percent and 80 percent of eligible schools or teachers achieve the bonus. One might hypothesize that, all other things being equal, programs near the mid-point of predictability (about 50 percent of participants earn the bonus) would exert the strongest incentive effects, but there is little empirical evidence on this.

The programs we reviewed use two different models of bonus assignment: “piecewise” formulations and “tournaments.” In piecewise formulations, each individual school or teacher’s award is determined by the value of the results they produce. In the case of Andhra Pradesh, India, teachers were rewarded for every student whose test scores increased by more than 5 percent over the school year—and the higher the scores, the higher the reward. Under the tournaments used in the Israel programs and Chile’s SNED, awards are restricted to a certain percentage of schools or individuals, ranked in order of their performance. In the case of Chile’s SNED, the top 25 percent of schools in each socioeconomic band are rewarded, whether average scores go up or down.

Both the average share of schools or teachers rewarded and the specific model will, in practice, enter into an agent’s estimation of the “predictability” of a program—that is, his or her chance of actually earning the bonus.
In other words, piecewise and tournament-style bonus programs might both result in the same numerical share of all teachers earning the bonus, but they may incentivize agents slightly differently ex ante because of differences in each individual’s levels of realism and risk aversion. Also, the evolution of a piecewise program over time will enable agents to refine their perceptions of predictability as evidence about the annual share of “winners” and average award size accrues.

While the theoretical literature predicts that tournament models will induce all individuals to make the same effort, the underlying assumptions are quite strong (that all agents have symmetric information, risk neutrality, and equal abilities) (Lazear 2000). Other theoretical literature speculates that piecewise (linear) bonus contracts have the advantage of being more resistant to gaming—but again, under strong assumptions of exponential utility for the agent and normally distributed noise (Holmstrom and Milgrom 1987). Again, the empirical evidence is limited.

For the purposes of this exercise, therefore, we use a limited but straightforward metric for assigning predictability ratings: the observed share of teachers or schools that accessed the bonus, on average, each year. For tournament programs, of course, this is a stable share. For the piecewise programs, the share will typically vary, but for many programs in our sample, only one year of experience was available. Figure 4A.1 indicates how...
we assigned Likert-scale ratings from 1 (low predictability) to 5 (high predictability) for these programs, in line with reported data on the average share of participants (whether schools or teachers) that actually received the award. These percentages along map onto a 100-percentile scale.

**Size-of-Bonus Ratings**

We express the size of all bonuses in terms of their proportion of the average monthly wage. Since most programs award a range of bonuses depending on the outcome achieved, we take the maximum bonus size as the main reference point to assign ratings. In doing this, we assume that all teachers have some ex ante awareness of the maximum size of the bonus and, further, that they perceive the maximum bonus as attainable. As shown in figure 4A.2, we divide the 5-point rating scale into quartiles, with a bonus of 0–25 percent of a monthly wage rated as 1 (very low) to 90 percent or more of a monthly wage rated as 5 (very high). For reference, available data on the actual maximum and average bonus size for each program are also presented.

**Notes**

1. When describing this experiment, all Indian rupee-to-dollar conversions use the exchange rate reported by the authors: Rs 48 per $1.
2. The bonus payment in the group-incentive schools ended up being lower than the individual incentives because the treatment effect was smaller (that is, the percentage gain in average scores was lower) and also because classes with scores below their targets brought down the average school gain, while teachers with negative gains (relative to targets) did not hurt teachers with positive gains in the individual-incentive schools (Muralidharan and Sundararaman 2009).
3. In fact, the authors argued that these kinds of behaviors would not be readily captured during classroom observations since many of these activities would be taking place after school or outside of the classroom.
4. More specifically, the material received by the “feedback” schools consisted of an independently administered baseline test at the start of the school year; a detailed written diagnostic feedback report on the performance of students on the baseline test; a note on how to read and use the performance reports and benchmarks; an announcement that students would be tested again at the end of the year to monitor progress in student performance; and low-stakes monitoring of classrooms during the school year to observe teaching processes and activity. In the feedback experiment, 100 schools received only the feedback treatment, while another set of randomly selected schools in the main experiment (those receiving incentives) also received feedback (Muralidharan and Sundararaman 2010b).

5. Because these are the results of the main experiment, the magnitude corresponds to the effects discussed above in Muralidharan and Sundararaman (2009).

6. Other subjects (Hebrew and Arabic) were tested as well, and their teachers also participated in the incentive program.

7. For the baseline, test scores in 1996 (Year 0) were used because there were no test scores for 1997. The first program year was 1998, and the second year was 1999.

8. A different model used to instrument the targets for the discontinuities also showed evidence that higher targets—all other things being equal—led to larger improvements in learning results over the short term.

9. To be eligible for the program, schools had to be comprehensive (offer grades 7–12) and be the only one of its kind in the community.

10. Students in Israel high schools must take matriculation exams in core and elective subjects for each grade of high school (grades 10–12) but have the option of taking all of the tests in their last year of high school, which most students choose to do. Matriculation exams are national exams that award credits for a particular subject. A minimum of 20 credits is required to qualify for a matriculation certificate, which is a necessary, although not sufficient, requirement for admission to a university (Lavy 2002).

11. The average annual starting teacher salary of a high school teacher in Israel is $20,000. Mean high school teacher salaries are $30,000 (Lavy 2002).

12. “Homogenous” groups are constructed based on geographical locations, educational levels, and student socioeconomic characteristics (such as parents’ education levels). The result is about 100 homogenous groups composed of schools that compete against each other (Romaguera 2008).

13. Schools ranked between 25 percent and 35 percent receive 60 percent of the bonus.

14. There is significant literature documenting undesired behaviors associated with high-stakes testing and teacher incentives programs in developed countries, notably in the United States, including cheating on exams (Jacob and Levitt 2002), increasing student caloric intake on the day of the exam (Figlio and Winicki 2002), and removing low-achieving students from the classroom (Murnane and Cohen 1986). The developing-country literature on these issues is less extensive, but the Kenya and India evaluations reviewed in this chapter documented increases in test preparation tutorials offered by teachers after
school—rather than improvements in pedagogy during school hours—as one of the main teacher behaviors stimulated by pay for performance.

15. First, a significant number of these programs have been run on a pilot scale, as experiments managed closely by academic researchers. Under these conditions, the scope for poorly designed tests of student learning or abuses in test administration are minimal. Second, of the government-run initiatives being implemented at scale, several are very new. There has been little time for inconsistencies in administration to surface that may undermine the credibility and power of the incentives. Nor has there been time for system actors to develop the perverse behaviors that both theory and empirical experience predict will arise as school personnel gain better awareness of the stakes and familiarity with the rules of the game.

16. There are arguments against noise being such an important measure because it could average out in the long run. However, our rankings are based on gross differences across programs, such as having baseline exams, adjusting for student socioeconomic status and other characteristics, or having a value-added framework that, even if not perfect, is technically more sound than a simple cross-section performance measure with no baseline to use for comparison.

17. Note that rating a measure as relatively noise-free does not imply that it is not susceptible to cheating or gaming. Attendance records, for example, are notorious for their manipulability.

18. It is likely that if the experiments or programs are repeated over time, teachers can adjust their perceptions based on previous results. But most of these programs were present for only one year or two years at the most (with the exception of Chile’s SNED, which is ongoing).

References


