The Promise of Performance Pay?
Reasons for Caution in Policy Prescriptions in the Core Civil Service[i]

Zahid Hasnain[ii], Nick Manning[iii], and Jan Henryk Pierskalla [iv]

There is a vast body of literature on performance-related pay (PRP), with strongly held views from opponents and proponents. This study reviews this literature, disaggregating the available evidence by the different public sector contexts, particularly the different types of public sector jobs, the quality of the empirical study, and the economic context (developing country or OECD settings), with the aim of distilling useful lessons for policy-makers in developing countries. The overall findings of the review are generally positive across these contextual categories. In particular, the findings from high quality studies, based on a simple scoring method for internal and external validity, of PRP in public sector-equivalent jobs show that explicit performance standards linked to some form of bonus pay can improve the desired service outcomes, at times dramatically. This evidence primarily concerns "craft" jobs, such as teaching, health care, and revenue administration, apparently negating (at least in the short term) the behavioral economics concern about the crowding out of intrinsic incentives. The available evidence suggests that if policy-makers are sensitive to design and vigilant about the risks of gaming, then PRP may result in performance improvements in these jobs in developing countries. However, it is difficult to draw firm conclusions from the review about the effect of PRP in core civil service jobs for three reasons. First, there are very few studies of PRP in these organizational contexts. The work of senior administrators in the civil service is very different from that of many private sector jobs and is characterized by task complexity and the difficulty of measuring outcomes. Second, although some studies have shown that PRP can work in even the most dysfunctional bureaucracies in developing countries, there are few cases illustrating its effectiveness or otherwise outside OECD settings. Finally, few studies follow PRP effects over time, providing little information on long-term effects and adjustments in staff behavior. We conclude that more empirical research is needed to examine the effects of PRP in the core civil service in developing countries.

---

[i] This paper is a revised version of World Bank Policy Research Working Paper 6043 (Hasnain et al., 2012). The paper benefited from several suggestions and comments, and the authors would like to particularly thank Mike Stevens, Willy McCourt, Mariano Lafuente, Gary Reid, and Svetlana Proskurovska. The views expressed are those of the authors and do not necessarily reflect those of the World Bank or its affiliated organizations.

[ii] World Bank (zhasnain@worldbank.org).

[iii] World Bank (nmanning@worldbank.org).

[iv] The Ohio State University (pierskalla.4@osu.edu).

Performance-related pay (PRP) is increasingly used in public sector organizations across the world. Currently, 28 of the 34 OECD countries have introduced PRP in some form (OECD 2011). Similar movements are underway in middle-income countries and, more sporadically, in lower-income countries in the health and education sectors. A vast theoretical and empirical body of literature has analyzed various dimensions of PRP, and there is now a small but growing body of robust evidence on the impact of PRP that is shedding new light on what is achievable and under what specific conditions.

The objective of this paper is to provide a review of this literature on PRP with relevance for the public sector, spanning the fields of public administration, psychology, economics, education, and health, with the aim of distilling useful lessons for policy-makers in developing countries. This is by no means the first comprehensive review of this literature, but it is, to our knowledge, the first that aims to disaggregate the available evidence by (i) the quality of the empirical study; (ii) the different public sector contexts, particularly the different types of public sector jobs; and (iii) country context (developing country or OECD settings). Although large parts of the existing literature address performance pay in the private sector or the OECD country context, this disaggregation allows us to identify more nuanced lessons for the application of performance pay in developing countries' core civil services.

PRP can be defined as a compensation arrangement in which the final salary of an employee is a function of some form of measured "performance". How performance is measured, who measures it, and how it is linked to salary can all vary considerably. Performance can be based on qualitative assessments or quantitative measures of inputs, outputs, or outcomes. Salary can be either wholly a function of performance, such as piece-rate pay in a manufacturing setting, or a combination of base pay and one-off bonuses or merit increases of base pay. Bonuses and merit increases can be awarded on an individual, small team, or larger departmental basis. Evaluations can be implemented by direct supervisors, human resource specialists, peer panels, or outside agencies. Once performance has been measured, it must be evaluated against a performance standard. This standard can be based on individually pre-agreed goals, absolute performance against minimum or scaled standards, relative improvements against past performance, rank-order of performance in a tournament evaluation, or relative performance measured against co-workers, other teams of co-workers, or other schools or agencies nationally or regionally.

Drawing on contract theory and the problems of moral hazard, a significant part of the academic literature has examined the impact of PRP on increasing effort and reducing shirking. In work environments where effort is unobservable, fixed pay contracts provide little ability for employers to influence employee effort. This is a particular problem in traditional civil service jobs that are characterized by uniform pay for jobs in similar grades, pay increases based largely on seniority, and negligible probability of termination. Contracts that tie observable outputs, which are correlated with unobservable effort, to desired pay incentives can theoretically mitigate these difficulties. Contract theory also suggests that PRP can help address the problem of adverse selection by encouraging high ability individuals who expect to do better under a performance pay scheme to join the agency and discouraging low ability individuals (the "sorting" effect).

Critics counter that PRP cannot work when outputs are difficult to measure and when tasks are multi-dimensional because it results in "gaming" behavior, whereby effort is only allocated toward activities that are observed and measured, which may not improve overall outcomes. The psychological and behavioral economics literature also argues that individuals are motivated by intrinsic concerns about the inherent social value of the job, particularly in the public sector, and that PRP, which explicitly focuses on extrinsic benefits, might crowd out intrinsic motivation and reduce worker productivity.

Our evaluation of existing studies finds a mounting body of evidence that PRP *can* increase worker effort in specific organizational contexts, particularly in those jobs where outputs are more easily measured. The evaluation proceeds in stages, first reviewing the evidence concerning PRP in general, regardless of whether the concerned jobs are in the public or private sectors. Then, drawing from the seminal work by J. Q. Wilson (Wilson 1989), a distinction is made between jobs with observable or unobservable production processes as a rough proxy for distinguishing between jobs that tend to be found in the private or public sectors. A number of increasingly rigorous studies document that performance pay can improve desired outcomes for teaching or health care provision in general, and in developing countries in particular, if it is carefully designed and implemented. Next, further distinctions are drawn between public sector jobs in which the outputs are measurable and those in which they are not as a further and likely more precise proxy for distinguishing between broad public service jobs and those in the core administration and between high-quality studies with high internal and external validity and others.[1] Although the universe of studies under consideration draws on all relevant studies regardless of their location, this winnowing reveals that there are no high-quality studies of PRP relevant to tasks in the core civil service outside of the OECD. Thus, with the current state of knowledge, it is not possible to infer much about the effectiveness of PRP for core civil service functions in developing countries.

<<A>>2. Theoretical debates

Theoretical debates on PRP have been evolving in the context of private businesses (Prendergast 1998; Prendergast 1999), the general public sector (Dixit 1999; Burgess and Ratto 2003; Perry et al. 2006), and specific occupations, such as teaching (Neal 2011). These debates can be roughly divided into psychological theories on human motivation and training, which are popular in public administration research, and core economic theories on incentive structures and principal-agent problems, with behavioral economics building a bridge between the two.

<<B>>Expectancy and reinforcement theory

Public administration research on PRP usually relies on what is often called "expectancy" (Vroom 1964; Porter and Lawler III 1968) and "reinforcement" theory (Skinner 1969; Luthans 1973). In its simplest form, the theory suggests that explicit incentives in the form of performance pay work under two conditions: first, employees need to believe that increased effort leads to increased performance; second, increased performance leads to desired outcomes and is recognized by management. If these two conditions are met, employees form a behaviorally salient expectation about a future reward and adjust their work effort upward. Reinforcement theory stresses the effect of cultivating a behavioral norm of high work effort by reinforcing behavior with positive rewards.

In addition to this direct link between PRP and individual effort, advocates in the field of public administration highlight the secondary effects of PRP: it helps recruit and retain highly skilled and/or motivated staff who presumably would do better under the scheme; it makes managers more committed to the strategic objectives and core organizational goals of the agency and increases the link between individual and organizational goals; it weakens the power of public sector unions to impose restrictive working practices; and it reduces the overall wage bill by moving away from automatic pay increases (Marsden 2004; OECD 2005b; Marsden 2009).

Critics of performance pay suggest that it is difficult to design performance pay schemes that meet the two conditions of expectancy and reinforcement theory (Kerr 1975). They argue that people do not

always approach work effort and the assessment of salary in an entirely rational way. In addition, many core public servants perform services that are difficult to measure or are non-measurable, or they produce outputs that are not market-priced. For example, early critics of test-score-based school and teacher evaluations argued that teacher performance cannot be neatly summarized by mechanical student test scores and that such practices invite behavior that contradicts the overall goals of the teaching profession (Murnane and Cohen 1986). The use of explicit and objective performance measures can induce tunnel vision, myopia, and measure fixation (Propper and Wilson 2003). Additionally, civil servants often work in large teams under the supervision of multiple managers, complicating the attribution of performance and the responsibility of performance evaluation. Some authors see the presence of high levels of trust and transparency between employees and management as a necessary condition for the effectiveness of PRP to avoid arbitrary implementation and worker dissatisfaction (Kellough and Lu 1993).

A different strand of criticism focuses on other motivations underlying public servants' efforts. Civil servants, it is argued, are motivated by notions of altruism, prosocial behavior, and commitment to institutional goals (Perry and Hondeghem 2008), which may compete or even conflict with explicit monetary incentives.

<<B>>Incentive and principal-agent theory

The most basic argument for incentive pay is based on a simple microeconomic principal-agent model of labor relations, in which a principal (the employer) wants to induce an agent (the employee) to perform a certain task. Such principal-agent relationships are commonly affected by two problems (Dixit 1999): moral hazard and adverse selection.

Moral hazard describes a scenario in which the agent's actions affect the principal's payoffs, but the action is not directly observable to the principal. In the workplace, the employee's effort is not directly observable, but it influences productivity and outcomes about which the employer cares. Under these conditions, offering fixed pay contracts to workers gives the employer little leverage to influence employee effort after hiring decisions have been made, a problem that is exacerbated if employees are difficult to fire. PRP can address this problem of moral hazard by tying observable outputs, which are correlated with unobservable effort, to pay.

In the case of adverse selection, the agent has access to private and valuable information at the time of contract signing. Adverse selection in the public sector plays an important role in recruitment, where low- and high-skilled applicants are difficult to distinguish based on public information. Public agencies need to offer contracts that induce high-quality applicants to apply and deter low quality applicants from misrepresenting their qualifications. PRP can alleviate this sorting problem because higher-quality workers will expect to perform better under this system of pay and therefore will be more likely to apply for a job opening (Delfgaauw and Dur 2008).

A well-known criticism of performance pay arrangements is that when tasks are multi-dimensional, incentivizing only tasks that are observable and measurable will not necessarily improve overall outcomes but rather will lead to a substitution of effort from unobservable to observable tasks, which can lead to worse outcomes (Holmstrom and Milgrom 1991). For example, the task of teaching can involve both instruction based on sound curricula and coaching on test-taking strategies, and poorly designed incentive schemes can encourage teachers to re-allocate effort to the latter and away from the former (known as “teaching to the test”), to the detriment of human capital accumulation.

The problem of selecting appropriate performance measures to address this problem has spawned its own theoretical and empirical debates (Courty et al. 2005). A problem related to the multi-tasking argument

addresses the issue of gaming or cheating incentive systems. Typical examples are the outright manipulation of results, whether through cream-skimming (i.e., the manipulative selection of clients to improve program effects; Heckman et al. 1997) or other forms of manipulation, such as the provision of high-caloric food to students on test days (Figlio and Winicki 2005).

The problem of gaming performance standards argues for ongoing adjustments in targets and metrics by the principal (Courty and Marschke 2003). To counteract the excessive gaming of incentive schemes, it has been suggested that evaluation systems should be used independently of output measurements (Neal 2011). Additionally, relative performance schemes in which employees are ranked against each other, potentially in a formal tournament setting, are much more difficult to manipulate (Barlevy and Neal 2011; Neal 2011). However, (Marsden 2009) notes that the gaming is not necessarily restricted to ingenious behaviors on the part of staff. Managers might be tempted to "collude with their subordinates: to go through the motions and fill in the forms for goal setting and appraisal, but not to worry about the reality" (Marsden 2009, 5)

Rewarding team performance can have certain advantages, ranging from reduced evaluation costs to the avoidance of harmful competition between employees. However, basing rewards on team outputs can also lead to problems of free-riding, where some team members willfully reduce their efforts in the expectation of relying on the work of others (Dixit 1999).

Choosing the correct bonus size brings its own challenges. Small bonuses will have little incentive effects and fall short of expectations, whereas large bonuses can lead employees to treat incentive schemes as pure lotteries, especially if outcomes are strongly stochastic (e.g., student test scores (Neal 2011)).

<<B>>Behavioral economics —intrinsic versus extrinsic motivation

When staff perform well because of the inherent characteristics of the job they are doing, e.g., its interest or perceived social value, they are said to be "intrinsically" motivated. When staff performance stems from rewards that are unrelated to the nature of the job, staff are said to be "extrinsically" motivated. Intrinsic motivations are of particular significance in public service (Banuri and Keefer 2013). Behavioral economists have argued that PRP can led to a reduction in effort by crowding out intrinsic motivation as employees change their perceptions about the nature of their work. Several authors (Kreps 1997; Benabou and Tirole 2003; Benabou and Tirole 2006) have developed formal treatments of the trade-off between extrinsic and intrinsic motivation, building on the literature from psychology. Crowding out can be especially salient if performance pay is introduced using antagonistic framing and can stifle creativity and collaboration (Frey and Osterloh 1999). More generally, Pink (2009) argues that monetary and other extrinsic incentives are both counterproductive (because they frequently undermine intrinsic incentives) and unnecessary (because intrinsic incentives can be harnessed and used to maximize individual productivity). Drawing on empirical work (Ryan and Deci 2000; Chirkov et al. 2003; Sauermann and Cohen 2008; Niemiec et al. 2009), his theory suggests that tasks can be constructed to maximize an individual's sense of (i) *autonomy;* (ii) *mastery* (continuous incremental learning and improvements rather than distant targets); and (iii) *purpose*, improving overall performance.

Another psychological argument, known as the "Yerkes-Dodson law", highlights the phenomenon of "choking under pressure" (Ariely et al. 2009). The argument is that performance has an "inverse U" relationship with the level of the incentive payment, with performance improving at low and moderate levels of incentive payments compared to no payments but becoming worse at very high levels of payment compared to moderate, low, or even no payments.

Finally, behavioral economists have identified the possibility of satisficing instead of maximizing behavior. Employees might exert effort until a certain minimum level of reward is reached and then substitute additional labor supply for increased leisure or idle time (Camerer et al. 1997).

<<A>>3. Organizing the empirical evidence

<<B>>The type of job

The hypothesized effects of PRP are heavily contingent on organizational context, particularly the nature of the job in which PRP is introduced. Borrowing James Q. Wilson's typology (Wilson 1989), jobs can be characterized by whether the job's outputs are easily measurable and whether the actions in the job to produce the output or the internal production process are observable.[2] Table 1 provides a framework to organize the empirical evidence by job type, with the simplifying assumption that jobs with multiple dimensions are located within the cell that represents the most complex of those dimensions. The top left box describes "Production Jobs", in which outputs are easily measurable, the production process consists of repeatable, mechanical tasks that are observable to an outside monitor, and controllability is likely to be high. Typical examples are manufacturing factory-floor jobs and municipal services, such as garbage collection. If the production process is not directly observable but outputs remain measurable, such jobs are termed "Craft Jobs". With recent advances in measuring learning outcomes, teaching can be classified as a job in which the exact process of production is difficult to ascertain, but, at least to a certain degree, desired outputs are quantifiable. Similarly, some of the outputs of healthcare, particularly in preventative services such as child immunization, are more measurable. Other examples include tax collection, job placement services, and auditing.

In the bottom row are "Procedural Jobs" and "Coping Jobs". Both are characterized by difficult-to-measure outputs, but they differ in the observability of the production process to an outsider. Procedural jobs such as the military have clearly defined inputs, whereas policy jobs in the core civil service neither produce easily measurable outputs nor have transparent production processes. Coping jobs present the most challenging functional contexts for PRP.

<<Table 1 about here>>

We use the distinction between jobs with observable and unobservable production processes (center and right-hand columns of Table 1) as a rough proxy for distinguishing between jobs that tend to be found in the private or public sectors. Within the latter, we use the distinction between public sector jobs in which the outputs are measurable and those in which they are not (center and bottom rows in Table 1) as a further and likely more precise proxy for distinguishing between broad public service jobs and those in the core administration. The universe of studies considered for this evaluation were drawn from anywhere in the matrix, but we consider studies of PRP in coping jobs to be likely the best measure of the impact of performance-related pay in the primarily policy-jobs in the core civil service.

Using this classification system allows us to draw on results from studies of private sector jobs if the jobs themselves are essentially similar to those found in the public sector. The general literature on performance pay in the private sector has focused on activities classifiable as production jobs and has found fairly encouraging results (Stajkovic and Luthans 2003). The task of this evaluation was to determine what survived of this general finding when the studies were winnowed down to coping jobs, specifically to high-quality studies of PRP in coping jobs.

<<B>>Methodological approaches

The vast empirical literature on PRP utilizes a range of methodological approaches. Early studies on performance pay for administrative civil service, teaching, and healthcare were largely observational. Studies that focused on administrative jobs were either qualitative case studies or used convenience samples of employees and senior management to collect data on self-assessed motivation and satisfaction with newly introduced performance pay. Studies of performance pay for teachers were more quantitative, assessing the impact of bonus schemes on actual public service outcomes (e.g., drop-out rates, students' grades, or test scores).

Although an improvement, many of these quantitative observational studies and similar work on private companies and the public service fall short of an ideal research design for causal inference on program effects. The gold standard for program evaluation is the use of randomized-controlled trials (RCTs), in which treatment assignment to subjects is randomized and unrelated to other observable and unobservable characteristics. This randomization allows the estimation of the treatment effect by comparing the treated and control units. Observational studies, in contrast, rely on treatment and control groups created not by controlled random assignment but by real-world social processes. Issues of selection bias and confounding factors can undermine the internal validity of such studies.

Utilizing the power of a randomization framework, several behavioral economists have used laboratory experiments to test hypotheses with regard to incentive schemes. Laboratory experiments offer at least two distinct advantages over observational studies: the researcher can use randomization to ensure the identification of treatment effects, and researchers can design their experiments to directly relate to the theoretical questions at hand. However, laboratory experiments often use notoriously small samples and student subjects who share few characteristics with actual workers or public servants, and they cannot replicate real workplace settings or offer bonus schemes that remotely approach the bonus sizes common even in moderately incentivized performance pay schemes.

The most recent attempts to address the issue of proper causal inference on performance pay and to increase the representativeness of results are RCTs in the field. In a field RCT, researchers are able to randomize key features of an actual policy program that services the population of interest. The advantage of such a field experiment is the similarity of the target population, the structure of the incentive program, and the randomization of treatment. Although field RCTs are time- and resource-intensive studies, several teams of researchers have implemented similar studies in different contexts around the world, adding considerably to the empirical understanding of performance pay.

<<A>>4. Assessing the evidence

<<B>>Classifying the universe of studies

In total, 153 empirical studies of PRP were considered in this review (see Hasnain et al. (2012) for the full list), of which 110 were for craft and coping jobs (Table 2), our proxy for general public sector jobs, with 17 of these for coping jobs specifically (our proxy for jobs in the core public administration). The research to date on the subject has largely focused on advanced countries; in the review, 127 studies are in OECD contexts, and only 26 are in developing country settings. The literature has also focused largely on craft jobs and production jobs, with no experimental studies to date on coping jobs.

<<Table 2 about here>>

The 153 studies that were reviewed were grouped into three categories to capture the effect of PRP that they revealed: *positive* if their findings provided positive evidence of the effectiveness of incentive schemes;[3] *neutral* if the study was largely descriptive or found contradictory evidence; and *failed* if the evidence indicated no effect or a negative effect of performance pay. Figure 1 shows the overall frequency of results. A majority of studies (93 of the 153) presented supportive evidence for some type of effect of performance pay schemes, with experimental studies showing more positive findings than observational ones.

<<Figure 1 about here>>

In drawing lessons, however, it is important to distinguish the findings more systematically by their research quality. Study quality was ranked in two different ways. First, each study was assessed for its "internal validity", or strength of the causal arguments, using a five-point ranking (from weak to strong) as follows:

1. no empirical study or a faulty research design[4]
2. descriptive; small sample size[5]
3. secondary data analysis and/or descriptive data analysis; small sample size; some statistical analysis[6]
4. quasi-experimental design; reasonable sample size; conclusions based on statistical analysis[7]
5. laboratory experiments; randomized controlled trial; large sample size; strong statistical analysis; strong conclusions.

Second, studies were evaluated on the dimension of "external validity", or the extent to which the causal connections in the specific context of the study would remain valid if replicated in other contexts. For example, lab experiments and RCTs offer very strong evidence of causality (high internal validity), but in a specific context: they tell us the average impact of a particular intervention in a particular location with a particular sample at a particular point in time. They are often accused of being low in external validity because the study subjects (usually college students, in the case of laboratory experiments) are not representative of the general population or, in this case, the population of interest (civil servants) and the requirements of the experiment imply very particular conditions that may not approximate real-world settings.

<<Figure 2 about here>>

Applying these two quality filters to the 153 studies resulted in 72 high-quality studies (ranked 4 or 5 on the internal validity scale and "high" on the external validity scale).[8] Of these, 54 high quality studies found positive effects of PRP (Figure 2).

<<B>>General observational and experimental studies

Previous meta studies have attempted to aggregate evidence concerning performance-related pay from observational and experimental studies across different task types (Jenkins et al. 1998; Condly et al. 2003; Weibel et al. 2009). General literature reviews (Petersen et al. 2006; Eldridge and Palmer 2009) highlight the degree to which studies have focused on OECD country experiences.

These overviews are generally positive regarding the impact of performance pay schemes. Most meta studies confine themselves to a broad qualitative assessment of the overall impact of PRP. Only two claim to offer quantitative data on the change in output per worker. Based on an analysis of 47 studies, (Jenkins et al. 1998) concluded that these incentives resulted in a 12% improvement in performance quantity and a

negligible effect on performance quality. Based on a meta-analysis of 64 field and laboratory experiments as well as observational studies, Condly et al. (2003) concluded that employees and other research participants who received performance incentives achieved an 22% average increase in work performance. In a meta-analysis of 46 high-quality empirical studies covering both simple and complex tasks and with both quantitative and qualitative outcome measurements, Weibel et al. (2009) found a statistically significant and positive effect of PRP on performance. The findings were significantly positive for simple tasks and significantly negative, albeit smaller, for complex tasks. The authors argued that this negative effect was because of the reduction in intrinsic motivation brought about by the incentive scheme.

Observational studies on performance-related pay are particularly suggestive that PRP has a positive impact, but with many caveats. Belfield and Marsden (2003) use panel data from a large UK workplace survey, which include both piece-rate jobs and "knowledge work" jobs in which performance is based on the achievement of previously agreed goals. They find strong effects of individual pay-for-performance, but only conditionally on the monitoring regime. Another study uses the British Household Panel Survey to distinguish the productivity and sorting effects of performance pay, finding that jobs with performance-related pay attract workers of higher ability and induce workers to provide greater effort (Booth and Frank 1999).

In contrast, Beer and Cannon (2004) study the failure of 13 incentive plans at Hewlett Packard using interviews and internal documents. They find that managers abandoned the programs because of the perceived costs. In a worldwide survey of 205 top managers, Beer and Katz (2003) document the weak support of incentive schemes among management.

These findings are in distinct contrast to the conclusions from qualitative observational OECD reports and discussion papers that report a negative effect of PRP on staff motivation, such as perceptions of unfairness, the difficulties of implementing such schemes given the tendency in bureaucracies to rate most employees as performing satisfactorily, and the political and operational difficulties of introducing any major pay reform within the public service (OECD 1993; OECD 1996; OECD 1997; Burgess and Ratto 2003; OECD 2004; OECD 2005a; OECD 2005b; Perry et al. 2006; Ketelaar et al. 2007; Rexed et al. 2007; OECD 2008; OECD 2009; Perry et al. 2009). This negative finding is mirrored in the smaller set of observational meta studies concerning developing countries (World Bank 1999; Kiragu and Mukandala 2003; Independent Evaluation Group 2008).

The evidence from experimental studies is largely from OECD countries. In a field experiment from the private sector (Bandiera et al. 2006), some managers were treated with the introduction of a performance-pay system, and the productivity of lower-tier workers was used as an outcome measure. The study finds evidence of both an incentive and sorting effect: managers supported their high-productivity workers and fired the least qualified employees. An experimental treatment of monitoring efforts by management in a call center found that employees largely behaved according to a rational-cheater model of human behavior, highlighting the importance of performance measurement as well as a substantial portion of employees that remains unaffected by monitoring attempts (Nagin et al. 2002).

Laboratory experiments have enabled the exploration of specific aspects of performance pay, including the functional relationship between bonus sizes and performance, the incentive and sorting effects of performance pay, the impact of different types of bonuses, and the possible tradeoffs between extrinsic and intrinsic motivation. Ariely et al. (2009) explore the effect of bonus size on performance in laboratory experiments using subjects in the US and India. Participants had to solve cognitive tasks under time pressure and were incentivized with bonuses that varied from small to large relative to their normal pay.

The authors find evidence of an "inverse-U" relationship between bonus size and performance, with a "choking-under-pressure" effect in which bonuses at very high levels lead to a worsening of performance.

Cadsby et al. (2007) find support for the incentive and sorting effect of performance pay and found that subjects with higher levels of risk aversion avoided pay-for-performance, suggesting important unintended side effects.

Addressing the problem of the multi-dimensionality of many tasks, Fehr and Schmidt (2004) conduct an experiment with university students to understand the effects of varying bonus schemes on effort provision on two distinct tasks, only one of which is contractible. They find that simple piece-rate contracts led to a focus on the contractible task, whereas bonus arrangements designed to be more encompassing and to explicitly address the multi-tasking problem also induce participants to spend time on the second task.

In a laboratory experiment by Gneezy and Rustichini (2000), high school and university students in Israel were offered different size of bonuses for specific tasks. The results suggest that subjects showed higher levels of productivity when offered large rewards but that small awards led to worse performance than did offering no monetary reward at all. This finding suggests the importance of framing of performance pay. If bonuses adequately communicate the importance of performing assigned tasks well compared to the overall goals of the organization, they can work; however, if bonuses trigger a change in the evaluation of the worker relationship, crowding out of intrinsic motivation can worsen productivity.

<<B>>Studies on jobs that are representative of the general public sector

This paper examines the effect of PRP in craft and coping jobs because these jobs most closely resemble public sector organizational contexts. Figure 3 presents the evidence for all of the reviewed studies for craft and coping jobs. Overall, 65 of the 110 studies found positive effects of PRP, with stronger evidence for the relatively few studies in developing countries.

<<Figure 3 about here>>

Many of these studies are in the health sector. The British NHS introduced performance-pay elements into the remuneration of primary care physicians in 2004. Several studies (Campbell et al. 2005; Campbell et al. 2007; Steel et al. 2007; Vaghela et al. 2009; Chalkley et al. 2010) report positively on the impact of these incentives. In contrast, in the US, Hillman et al. (1991) report more mixed findings of similar schemes for physicians, and Shen (2003) and Doran et al. (2006) find evidence of "gaming".

The few studies of PRP in the health sector in low- and middle-income countries have generally found positive results but illustrate data problems. McNamara (2005) discusses six cases of payment for quality in the health services sector across developed and developing countries, with cases in Nicaragua and Haiti having a positive effect. The Nicaraguan reform efforts combined the decentralization of decision-making authority and increased local accountability with explicit performance agreements. They were considered to have led to an overall improvement (Jack 2003), but it is difficult to disentangle the effects of each reform element. Similarly, in a recent study by Witter et al. (2011), a pay-for-performance arrangement in a NGO-led health project in the Battagram district of Pakistan was found to have improved general services provision, but with an unclear effect of the performance-based elements.

Other studies are clearer. Meessen et al. (2007) evaluate the performance of 15 health centers in Kabutare, Rwanda, and document a sharp increase in staff productivity after the introduction of output-based bonuses. Soeters et al. (2006) highlight the potential applicability of the Rwandan experience in sub-

Saharan Africa more generally. In a similar vein, efforts to improve health services provision in Haiti using performance-based payment for NGOs in a USAID pilot project showed encouraging effects on immunization coverage and organizational behavior (Eichler et al. 2001).

Outside the health sector, revenue authorities have been studied extensively, most likely because they provide examples of craft jobs in which, although the methods of work are difficult to observe, the outputs (the number of audits conducted and tax fines collected) are more easily measurable. Kahn et al. (2001) examine a 1998 Brazil incentive scheme and found that it resulted in a 75% increase in fines per inspection. The World Bank (2001) concludes that "circumstantial evidence" suggests that bonus systems do seem to have an impact on organizational effectiveness in revenue administrations.

<<Figure 4 about here>>

Overall, as Figure 4 illustrates, PRP for craft jobs is generally found to have a positive effect.

<<B>>Narrowing by quality

The higher-quality studies for craft and coping jobs that were reviewed were primarily in the education and health sectors. There is an extensive and growing body of literature on performance pay for teachers. In the US, most observational studies have primarily examined the impact of performance incentives on student test scores, although a few studies (Clotfelter et al. 2004; Clotfelter et al. 2007; Clotfelter et al. 2008) use detailed data from North Carolina schools to show that accountability and performance pay systems can also contribute positively to retaining quality teachers. With regard to student test scores, the evidence is mixed (Eberts et al. 2002; Dee and Keys 2004). A number of studies identify problems of gaming, such as outright cheating (Jacob and Levitt 2003; Jacob 2005) or, more subtly, the adjustment of the caloric content of school lunches to improve cognitive ability on test days (Jacob and Levitt 2003; Figlio and Winicki 2005; Jacob 2005).

Outside the US, an analysis by Atkinson et al. finds clear positive effects of performance pay for British schools (Atkinson et al. 2004). A set of observational studies (Lavy 2008; Lavy 2009) use data from an Israeli policy experiment with tournament-based teacher competition for bonuses and find significant gains in student achievements.

A number of field experiments have evaluated the impact of performance pay for teachers on reducing absenteeism and improving learning outcomes. The findings are generally mixed. Duflo et al. (2010) show that random assignment for monitoring and financial incentives for teachers in rural India led to a strong reduction of teacher absenteeism and increased students' test scores. Kremer and Chen (2001), in contrast, show that subjective monitoring arrangements by an individual in the institutional hierarchy (such as the headmaster of a school) may not work in developing country settings because the monitor might shirk, attempt to avoid confrontation, or collude with the workers. These studies suggest that impersonal, external monitoring by a camera coupled with a clear, credible, and automatic threat of punishment and promise of reward was the key design feature for program success.

A field experiment in 50 Kenyan schools linking teacher salaries to student test scores failed to find lasting effects (Glewwe et al. 2010). Teacher attendance did not improve, and teachers did not adjust their teaching methods or conduct more preparation sessions. Students in treated schools performed better during the program duration, but these gains did not extend beyond the study period. A field experiment conducted in NYC public schools also failed to find statistically significant effects of team incentives for teachers on student outcomes (Fryer 2011). A related study that assessed the effects of the NYC group incentive program on classroom activities and teacher turnover and qualification, apart from test scores

and teacher effort, similarly found no effects (Goodman and Turner 2010). A three-year experimental evaluation of the Project on Incentives in Teaching in Metropolitan Nashville schools also found no significant effects of bonus incentives on student test scores (Springer et al. 2010).

In contrast, a large-scale field experiment in a representative sample of 300 government-run rural primary schools in India found that bonus pay linked to the mean improvement of student test scores in an independent learning assessment led to a statistically significant and substantively meaningful improvement of student outcomes (Muralidharan and Sundararaman 2009).

In the health sector, a number of randomized-controlled trials have been implemented to determine the role of performance pay on health worker productivity, patient treatment, and outcomes. Similar to studies on health care relying on observational data, the majority of studies assess these questions in the context of OECD health care systems. Kouides et al. (1998) implemented a randomized-controlled trial, offering a randomly selected set of primary care physicians financial incentives based on influenza immunization rates of the elderly as part of a Medicare demonstration project. Doctors in the treatment group performed more immunizations. Hillman et al. (1998) and Hillman et al. (1999) used two RCT designs to incentivize cancer screenings for women aged 50 and above and pediatric immunizations, respectively. In both studies, the authors documented no significant difference between the treatment and control groups. Similarly, an RCT implemented by Grady et al. (1997) found no clear effects of financial incentives on mammography referrals by primary care physicians.

In contrast, a set of studies (Fairbrother et al. 1999; Fairbrother et al. 2001), also focusing on pediatric immunizations, found that performance incentives increased immunizations rates by several percentage points compared to the control group. A randomized field trial at the clinic-level found that financial incentives improved the treatment of smoking cessation outcomes (Roski et al. 2003). Work on performance pay for cognitive services interventions by pharmacists has also demonstrated positive effects (Christensen et al. 2000).

To our knowledge, the only two available randomized-controlled trials on performance pay in health care in a low-income country are a study by Basinga et al. (2010) in Rwanda and a study by Singh (2010) in India. Basinga et al. used an RCT design to evaluate performance pay in Rwandan primary health care centers. The authors took advantage of a sequenced roll-out of the scheme across Rwandan health care facilities, collecting data on child preventive care and prenatal delivery. To isolate the performance-pay effect from a general increase in resources, comparison facilities received an equivalent increase in their budgets. The study used information from 166 facilities and 2158 households. The authors found large effects on all central outcome measures, with particularly striking effects for services with the highest payoffs and smallest necessary staff effort.

Singh (2010) treated three groups of mothers and the staff providing child care and nutritional advice to them in Chandigarh, India. In one group, the workers received performance pay; in a second group, the workers had no performance pay, but the women they worked with were separately given factual information about nutrition; and the third group received both treatments. The study found that children's weights improved only in the third group compared to the control group.

It is noteworthy that nearly all of the identified studies on the health care sector focus on fairly narrow types of performance pay and specific, single outcome measures in preventative care, not necessarily overall multidimensional patient treatments and outcomes.

<<Figure 5 about here>>

Outside these sectors, Burgess et al. (2010) used an RCT to examine the impact of a pilot team-based incentive scheme introduced in 2002 on the indirect tax assessment and collection agency of the UK government. The tax yield increased for both the treatment teams relative to the control group, with the increases occurring because more time was spent auditing, which resulted in the recovery of greater tax revenue. Bertelli (2006) found that in the Internal Revenue Service in the US, the incentive scheme crowded in intrinsic motivation at the lowest pay levels and crowded it out at the highest levels. A set of studies of performance incentives for agencies with responsibility for training and recruitment found considerable evidence of gaming among the agency staff in the choice of the termination date of the training for the participants (Asch 1990; Heckman et al. 1997; Courty and Marschke 2004).

Other interesting findings from this group of high quality studies of craft and coping jobs were from Straberg (2010), whose empirical study concerning the perceived impact of performance-related pay in Sweden showed that men were much more likely than women to see the arrangement as fair and reasonable. A laboratory experiment in India assessed teacher efforts when rewards were a function of average student test scores and revealed that poorly designed incentive plans led to the misallocation of teacher effort, which produced an unequal distribution of effort across student groups. However, properly designed incentives could mitigate such behavior (Jain and Narayan 2011).

Overall, raising the quality bar still produces a positive result for craft and coping jobs (Table 3). Thirty-seven of the 53 high-quality studies (4 or 5 on the measure of internal validity and "high" on the measure of external validity) show positive findings. However, it must be emphasized that the number of studies for developing countries is low and that, as noted, the majority of the studies are in the health and education sectors (Table 3).

<<Table 3 about here>>

<<B>>Narrowing further – high-quality studies in relation to coping jobs

Coping jobs are our proxy for jobs within the core civil service. We found very few high-quality studies that addressed these jobs, and none within developing countries (Table 3). Of the three studies that were reviewed, two were on performance pay for managerial positions in the private sector, and only one was on core administrative jobs in the public sector. All of these studies showed positive effects of the performance incentive. Hochberg and Lindsey (2010) reviewed the impact of stock options on company rank-and-file on firm performance (as opposed to the impact of options on top executives, on which there is a large body of literature), finding a positive effect on firm performance. However, although Aboody et al. (2010) similarly showed that firms that repriced their options had a larger increase in operating income and cash flows compared to non-repricers, they also found that this impact was entirely because of executive stock options. In examining performance-related pay for managers in the UK National Health Service, Dowling and Richardson (1997) found higher staff views on the performance-related pay system in these jobs.

<<A>>5. Summary and Implications for Developing Countries

Figure 6 summarizes the major findings of this review with the pool of studies filtered based on job type and quality of the empirical study. Figure 7 summarizes the public sector equivalent studies by country type.

<<Figures 6 and 7 about here>>

The overall body of evidence paints a generally supportive picture of performance pay in most jobs and in craft and coping jobs that reflect the tasks found in the public sector. This conclusion is strengthened when the sample is narrowed to exclude lower-quality studies and is surprisingly even stronger when the sample is limited to high-quality public sector equivalent jobs in developing countries. However, when focusing narrowly on coping jobs, which more closely resemble those found in the core civil service, the number of studies becomes trivial, and there are none that assist in understanding the relevance of PRP to such jobs outside of the OECD.

At the same time, several observational studies identify problems with unintended consequences, generically subsumed under "gaming" the incentive scheme, which can run counter to the original intentions of the reforms. With the current evidence, however, it remains unclear whether incidents of gaming have a net negative effect in the presence of increased productivity. Furthermore, although explicit incentive schemes certainly increase the opportunity for gaming, standard civil service arrangements have their own unintended incentive effects (i.e., employees will engage in behavior that increases the chances of easy work assignments or promotions). It is simply unknown whether the existing forms of gaming are worse than similar behavior under performance pay. In addition, important cultural differences might exist in the prevalence of gaming performance standards in the public sector between developed and developing countries. Although, to our knowledge, no explicit research on this question exists, work on the prevalence of corruption, behavioral norms, and the effectiveness of anti-corruption efforts suggests that gaming might be more problematic in highly corrupt bureaucracies.

Moving to studies that attempt to fulfill the gold standard of experimental design, the evidence overall again supports the potential utility of performance pay for craft jobs. Comparing various laboratory experiments, the results suggest that explicit performance incentives can work, but the studies employ easily measurable performance indicators and use fairly unrepresentative subject pools. Both concerns should caution policy makers against accepting the results independently of other research. In contrast, similar results have been found across a varied set of experimental settings, test locations, and subject pools, and the overall findings resonate with the observational literature, improving the overall credibility and external validity.

The strongest form of evidence comes from field experimental studies for craft jobs that neatly address concerns of internal and external validity. Here, the evidence is somewhat more mixed. Several studies of teacher incentive programs have found no or transient effects of bonus pay systems in the context of US schools, but in the developing world, the evidence has been more positive. The discrepancy between teacher incentives in the developed and developing world could stem, on the one hand, from the relative magnitude of incentives compared to normal salary or, on the other hand, from higher marginal effects in the education production function in developing countries. Many factors enter the production of education, all of which are likely lacking in many developing country schools. Improving one input aspect (e.g., teacher presence and effort) could have conceivably larger marginal effects than the same input improvement in a developed country school.

What can be concluded with some degree of confidence from this evidence is that if policy-makers are sensitive to design and vigilant about the risks of gaming, then PRP can incentivize workers in both OECD and developing countries to increase effort in craft jobs in which the outputs are readily observable, such as teaching, health care, and revenue administration. The evidence confounds, at least in the short term, the behavioral economics concern about the crowding out of intrinsic incentives. The key focus for policy-makers should be on the design and implementation modalities of the PRP scheme, such

as the size of the incentive, whether it should be an individual or group-based award, the nature of the performance evaluation, and the monitoring regime.

For the core civil service, it is more difficult to draw conclusions for three reasons. First, there is very little research on PRP in these organizational contexts. Second, although some studies have shown that PRP can work even in the most dysfunctional bureaucracies in developing countries, there are too few cases to draw firm conclusions concerning its effectiveness outside of OECD settings for coping jobs. Most glaringly, the role of politicized bureaucracies has not been addressed properly. Finally, although studies do not show that problems with unintended consequences and gaming of incentive schemes result in an overall decline in productivity compared to the counter-factual, few studies follow up performance-related pay effects over a long period of time, leaving the possibility that the positive findings may occur because of Hawthorne Effects (i.e., only as a result of the subjects knowing that they are part of a study and not because of the incentive itself) and that gaming behavior may increase over time as employees become more familiar with the scheme and learn to manipulate it.

Given this problem of the measurability of outputs, PRP for core civil service jobs will largely need to draw on the subjective performance evaluations of individuals. It would be reasonable to expect that performance pay schemes in the core administration are likely to be more successful if they foster better performance dialogue between staff and their managers, based on individual-specific results agreements. Thus, prior investments in improving this performance dialogue and individual performance assessments may increase the likelihood of PRP's positive effects on worker motivation and effort. This dialogue is also more likely to function better under decentralized arrangements for human resource management in which the management of staff and decisions over the performance bonus are the responsibility of line managers and not central human resource agencies or oversight bodies.

Similarly, it can be hypothesized that PRP may have a role in coping jobs through complementing other management reforms, such as results-based management or performance budgeting, which attempt to inculcate a goal orientation in an organization. While the measurability of outputs will always be a challenge by the very nature of these jobs, even these jobs can develop some proxy measures for outputs, such as the level of satisfaction of the users of the services of these jobs, which in the case of policy and regulatory jobs, will be other the government agencies impacted by these policies and regulations. To the extent that reasonable proxy measures are developed, then the positive results from the studies from craft jobs may be more generalizable.

PRP's contribution to agency level productivity through the sorting channel — attracting higher-quality workers who are likely to do better in this scheme — has already been noted, and there is some fairly robust evidence of this in developed country contexts (Booth and Frank 1999; Bandiera et al. 2006; Cadsby et al. 2007). Consequently, it is possible that PRP may have a clearer effect on individual incentives if the sorting effect had longer to work its way through the system. To evaluate this effect, future studies need a longer time frame to assess changes in recruitment and workplace behavior.

Finally, this literature review focuses almost entirely on the individual incentive effects of PRP because this has been the emphasis of the bulk of the studies. It should be noted that there are, at least in the policy literature, potential agency-level and public sector-wide effects of PRP, such as a change in performance culture or fiscal sustainability, that should be explored in future studies.

Notes

[1] The list of studies reviewed and classified by type of job and quality is available from the authors upon request.

[2] Wilson originally used this framework to classify organizations and not jobs, with the implicit assumption that organizations were homogenous in the tasks that they performed.

[3] Inevitably, there is some subjectivity in the classification of studies. Studies were rated as *positive* if there was general evidence of the basic functionality of incentive schemes, even if the additional results qualified the effect; for example, studies on the crowding out of intrinsic motivation generally still find positive effects of explicit incentives.

[4] Interpreted as purely theoretical papers or studies with a weak research design (e.g., selection on the dependent variable only, no meaningful variation, and no explicit consideration of counter-factuals).

[5] Studies that mostly describe reforms implemented in a small number of cases without comparing to cases without performance pay.

[6] Studies based on a small number of cases but with at least an implicit consideration of a counter-factual and the use of some minimal data analysis.

[7] Studies with an explicit counter-factual analysis, representative sample of cases with and without treatment, and often explicit use of statistical techniques to limit threats to causal inference.

[8] We believe that a minimal level of internal and external validity is necessary to draw reliable conclusions from the evidence presented in a study, especially when policy recommendations are concerned. For that reason, we opted to classify studies as "high" quality only if their analysis was based on quasi-experimental methods and the analyzed sample was somewhat representative of the theoretical population under study.

References

Aboody, D., N. Johnson, and R. Kasznik. 2010. "Employee Stock Options and Future Firm Performance: Evidence from Option Repricings." *Journal of Accounting and Economics* 50 (1): 74–92.

Ariely, D., U. Gneezy, G. Loewenstein, and N. Mazar. 2009. "Large stakes and big mistakes." *Review of Economic Studies* 76 (2): 451–69.

Asch, B. J. 1990. *Navy Recruiter Productivity and the Freeman Plan*. Santa Monica: RAND Corporation.

Atkinson, A., S. Burgess, B. Croxson, and P. Gregg. 2004. "Evaluating the Impact of Performance-related Pay for Teachers in England." Working Paper No.04/113. Bristol, UK, Centre for Market and Public Organisation.

Bandiera, O., I. Barankay, and I. Rasul. 2006. "Incentives for Managers and Inequality Among Workers: Evidence from a Firm Level Experiment." Discussion Paper No. 2062. Bonn, Germany, Institute for the Study of Labor

Banuri, S. and P. Keefer. 2013. "Intrinsic Motivation, Effort and the Call to Public Service." Policy Research Working Paper 6729. Washington DC. World Bank,

Barlevy, G. and D. Neal. 2011. "Pay for Percentile" Working Paper 17194. Cambridge, MA, National Bureau for Economic Research.

Basinga, P., P. J. Gertler, A. Binagwaho, A. L. B. Soucat, J. R. Sturdy, and C. Vermeersch. 2010. "Paying Primary Health Care Centers for Performance in Rwanda." Policy Research Working Paper 5190. World Bank, Washington DC.

Beer, M., and M. D. Cannon. 2004. "Promise and Peril in Implementing Pay-for-Performance." *Human Resource Management* 43 (1): 3–48.

Beer, M., and N. Katz. 2003. "Do Incentives Work? The Perceptions of a Worldwide Sample of Senior Executives." *People and Strategy* 26 (3): 30–44.

Belfield, R., and D. Marsden. 2003. "Performance pay, monitoring environments, and establishment performance." *International Journal of Manpower* 24 (4): 452–89.

Benabou, R., and J. Tirole. 2003. "Intrinsic and Extrinsic Motivation." *Review of Economic Studies* 70 (3): 489–520.

Benabou, R., and J. Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review* 96 (5): 1652–78.

Bertelli, A. M. 2006. "Motivation crowding and the federal civil servant: Evidence from the U.S. Internal Revenue Service." *International Public Management Journal* 9 (1): 3–23.

Booth, A. L., and J. Frank. 1999. "Earnings, Productivity, and Performance-Related Pay." *Journal of Labor Economics* 17 (3): 447–63.

Burgess, S., C. Propper, M. Ratto, S. von Hinke Kessler Scholder, and E. Tominey. 2010. "Smarter Task Assignment or Greater Effort: What Makes the Difference on Team Performance?" *The Economic Journal* 120 (547): 968–89.

Burgess, S., and M. Ratto. 2003. "The Role of Incentives in the Public Sector: Issues and Evidence." Working Paper. Bristol, UK, Centre for Market and Public Organisation.

Cadsby, C. B., F. Song, and F. Tapon. 2007. "Sorting and Incentive Effects of Pay-for-Performance: An Experimental Investigation." *Academy of Management Journal* 50 (2): 387–405.

Camerer, C., L. Babcock, G. Loewenstein, and R. Thaler. 1997. "Labor Supply of New York City Cabdrivers: One Day at a Time." *Quarterly Journal of Economics and Philosophy* 111 (2): 407–41.

Campbell, S. M., D. Reeves, E. Kontopantelis, E. Middleton, B. Sibbald, and M. Roland. 2007. "Quality of Primary Care in England with the Introduction of Pay for Performance." *New England Journal of Medicine* 357: 181–90.

Campbell, S. M., M. Roland, E. Middleton, and D. Reeves. 2005. "Improvements in the Quality of Clinical Care in English General Practice: Longitudinal Observational Study." *British Medical Journal* 331: 1121.

Chalkley, M., C. Tilley, L. Young, D. Bonetti, and J. Clarkson. 2010. "Incentives for Dentists in Public Service: Evidence from a Natural Experiment." *Journal of Public Administration Research and Theory* 20 (suppl 2): 207–23.

Chirkov, V. I., R. M. Ryan, Y. Kim, and U. Kaplan. 2003. "Differentiating autonomy from individualism and independence: A self-determination theory perspective on internalization of cultural orientations and well-being." *Journal of Personality and Social Psychology* 84 (1): 97–110.

Christensen, D. B., N. Neil, W. E. Fassett, D. H. Smith, G. Holmes, and A. Stergachis. 2000. "Frequency and characteristics of cognitive services provided in response to a financial incentive." *Journal of the American Pharmaceutical Association* 40 (5): 609–17.

Clotfelter, C., H. F. Ladd, J. L. Vigdor, and R. A. Diaz. 2004. "Do School Accountability Systems Make It More Difficult for Low-Performing Schools to Attract and Retain High-Quality Teachers?" *Journal of Policy Analysis and Management* 23 (2): 251–71.

Clotfelter, C., H. F. Ladd, and J. L. Vigdor. 2007. "How and Why Do Teacher Credentials Matter for Student Achievement?" Working Paper 2. Washington DC, National Center for Analysis of Longitudinal Data in Educational Research.

Clotfelter, C., E. Glennie, H. Ladd, and J. Vigdor. 2008. "Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina." *Journal of Public Economics* 92 (5-6): 1352–70.

Condly, S., R. Clark, and H. D. Stolovitch. 2003. "The Effects of Incentives on Workplace Performance: A Meta-analytic Review of Research Studies." *Performance Improvement Quarterly* 16 (3): 46–63.

Courty, P., C. Heinrich, and G. Marschke. 2005. "Setting the Standard in Performance Measurment Systems." *International Public Management Journal* 8 (3): 1–27.

Courty, P., and G. Marschke. 2003. "Dynamics of Performance-Measurement Systems." *Oxford Review of Economic Policy* 19 (2): 268–84.

Courty, P., and G. Marschke. 2004. "An Empirical Investigation of Gaming Responses to Explicit Performance Incentives." *Journal of Labor Economics* 22 (1): 23–56.

Dee, T. S., and B. J. Keys. 2004. "Does Merit Pay Reward Good Teachers? Evidence from a Randomized Experiment." *Journal of Policy Analysis and Management* 23 (3): 471–88.

Delfgaauw, J., and R. Dur. 2008. "Incentives and Worker's Motivation in the Public Sector." *The Economic Journal* 118 (525): 171–91.

Dixit, A. 1999. "Incentives and Organization in the Public Sector. An Interpretative Review." *The Journal of Human Resources* 34 (4): 696–727.

Doran, T., C. Fullwood, H. Gravelle, D. Reeves, E. Kontopantelis, U. Hiroeh, and M. Roland. 2006. "Pay-for-Performance Programs in Family Practices in the United Kingdom." *New England Journal of Medicine* 355: 375–84.

Dowling, B., and R. Richardson. 1997. "Evaluating Performance-related Pay For Managers in the National Health Service." *The International Journal of Human Resource Management* 8 (3): 348–66.

Duflo, E., R. Hanna, and S. P. Ryan. 2010. "Incentives Work: Getting Teachers to Come to School." Cambridge, Mass., MIT (Department of Economics and J-PAL) and the Kennedy School of Government.

Eberts, R., K. Hollenbeck, and J. Stone. 2002. "Teacher Performance Incentives and Student Outcomes." *Journal of Human Resources* 37 (4): 913–27.

Eichler, R., P. Auxila, and J. Pollock. 2001. "Performance-Based Payment to Improve the Impact of Health Services: Evidence from Haiti." *World Bank Institute Online Journal* (April 2001).

Eldridge, C., and N. Palmer. 2009. "Performance-based payment: some reflections on the discourse, evidence and unanswered questions." *Health Policy and Planning* 24 (3): 160–6.

Fairbrother, G., K. L. Hanson, S. Friedman, and G. C. Butts. 1999. "The impact of physician bonuses, enhanced fees, and feedback on childhood immunization coverage rates." *American Journal of Public Health* 89 (2): 171–5.

Fairbrother, G., M. J. Siegel, S. Friedman, P. D. Kory, and G. C. Butts. 2001. "Impact of Financial Incentives on Documented Immunization Rates in the Inner City: Results of a Randomized Controlled Trial." Ambulatory Pediatrics 1 (4): 206–12.

Fehr, E., and L. Goette. 2007. "Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment." *Quarterly Journal of Economics* 97 (1): 298–317.

Figlio, D. N., and J. Winicki. 2005. "Food for thought: the effects of school accountability plans on school nutrition." *Journal of Public Economics* 89 (2-3): 381–94.

Frey, B. S., and M. Osterloh. 1999. *Pay for Performance - Immer Empfehlenswert?* Münster, Germany, Zeitschrift fur Fuhrung und Organisation.

Fryer, R. G. 2011. "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." NBER Working Paper 16850. Washington DC, National Bureau for Economic Research.

Glewwe, P., N. Ilias, and M. Kremer. 2010. "Teacher Incentives." *American Economic Journal: Applied Economics* 2 (3): 205–27.

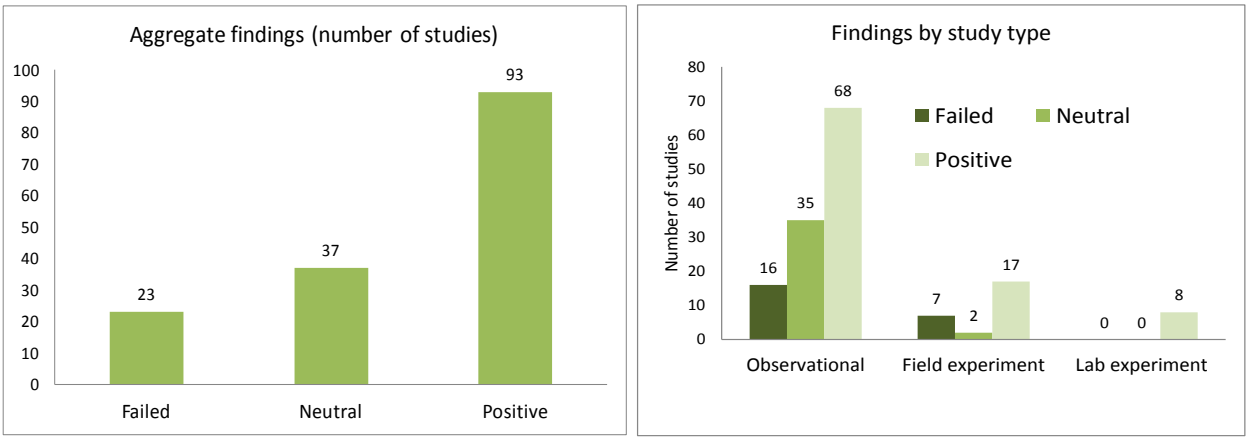
Gneezy, U., and A. Rustichini. 2000. "Pay Enough or Don't Pay at All." *The Quarterly Journal of Economics* 115 (3): 791–810.
Goodman, S., and L. Turner. 2010. "Teacher Incentive Pay and Educational Outcomes: Evidence from the NYC Bonus Program." Working Paper. PEPG Conference "Merit Pay: Will It Work? Is It Politically Viable?". Harvard Kennedy School, June 3-4.
Grady, K., J. Lemkau, N. R. Lee, and C. Caddell. 1997. "Enhancing Mammography Referral in Primary Care." *Preventive Medicine* 26 (6): 791–800.
Hasnain, Z., N. Manning, and J. H. Pierskalla. 2012. "Performance-related Pay in the Public Sector: A Review of Theory and Evidence." World Bank Policy Research Working Paper 6043. Washington DC, World Bank.
Heckman, J., C. Heinrich, and J. Smith. 1997. "Assessing the Performance of Performance Standards in Public Bureaucracies." *The American Economic Review* 87 (2): 389–95.
Hillman, A., M. Pauly, K. Kerman, and C. R. Martinek. 1991. "HMO manager's views on financial incentives and quality." *Health Affairs* 10 (4): 207–19.
Hillman, A., K. Ripley, N. Goldfarb, I. Nuamah, J. Weiner, and E. Lusk. 1998. "Physician Financial Incentives and Feedback: Failure to Increase Cancer Screening in Medicaid Managed Care." *American Journal of Public Health* 88 (11): 1699–701.
Hillman, A., K. Ripley, N. Goldfarb, J. Weiner, I. Nuamah, and E. Lusk. 1999. "The use of physician financial incentives and feedback to improve pediatric preventive care in Medicaid managed care." *Pediatrics* 104 (4 pt 1): 931–5.
Hochberg, Y. V., and L. Lindsey. 2010. "Incentives, Targeting and Firm Performance: An Analysis of Non-Executive Stock Options." *Review of Financial Studies* 23 (11): 4148–86.
Holmstrom, B., and P. Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." Journal of Law, Economics & Organization 7: 24–52.
Independent Evaluation Group. 2008. *Public Sector Reform: What works and Why?* Washington DC: World Bank.
Jack, W. 2003. "Contracting for health services: an evaluation of recent reforms in Nicaragua." *Health Policy and Planning* 18 (2): 195–204.
Jacob, B. A. 2005. "Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools." *Journal of Public Economics* 89 (5-6): 761–96.
Jacob, B. A., and S. D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics* 118 (3): 843–77.
Jain, T., and T. Narayan. 2011. "Incentive to discriminate? An experimental investigation of teacher incentives in India." Working Paper, Indian School of Business.
Jenkins, G. D., A. Mitra, N. Gupta, and J. D. Shaw. 1998. "Are Financial Incentives Related to Performance? A Meta-analytic Review of Empirical Research." *Journal of Applied Psychology* 83 (5): 777–87.
Kahn, C. M., E. C. De Silva, and J. P. Ziliak. 2001. "Performance-Based Wages in Tax Collection: The Brazilian Tax Collection Reform and Its Effects." *The Economic Journal* 111 (468): 188–205.
Kellough, E. J., and H. Lu. 1993. "The Paradox of Merit Pay in the Public Sector: Persistence of a Problematic Procedure." *Review of Public Personnel Administration* 13 (2): 45–64.
Kerr, S. 1975. "On the Folly of Rewarding A, While Hoping for B." *The Academy of Management Journal* 18 (4): 769–83.
Ketelaar, A., N. Manning, and E. Turkisch. 2007. "Performance-based arrangements for senior civil servants OECD and other country experiences." OECD Governance Working Paper, Paris.
Kiragu, K., and R. Mukandala. 2003. "Public Sector pay reform - Tactics Sequencing And Politics In Developing Countries: Lessons From Sub-Saharan Africa." Dar es Salaam: Pricewaterhousecoopers and University of Dar es Salaam.
Kouides, R., N. Bennett, B. Lewis, J. Cappuccio, W. Barker, and M. LaForce. 1998. "Performance-Based Physician Reimbursement and Influenza Rates in the Elderly." *American Journal of Preventive Medicine* 14 (2): 89–95.
Kremer, M., and D. Chen. 2001. "An Interim Report on a Teacher Attendance Incentive Program in Kenya." Mimeo. Cambridge, MA: Harvard University.
Kreps, D. M. 1997. "Intrinsic Motivation and Extrinsic Incentives." *The American Economic Review* 87 (2): 359–64.
Lavy, V. 2008. "Gender Differences in Market Competitiveness in a Real Workplace: Evidence from Performance-Based Pay Tournaments Among Teachers." NBER Working Paper No. 14338. Washington DC, Natoinal Bureau for Economic Research.

Lavy, V. 2009. "Performance Pay and Teachers' Effort, Productivity and Grading Ethics." *American Economic Review* 99 (5): 1979–2011.
Luthans, F. 1973. *Organizational Behavio*r. New York: McGraw-Hill.
Marsden, D. 2004. "The role of performance-related pay in renegotiating the "effort bargain": the case of the British public service." *Industrial and Labor Relations Review* 57 (3): 350–70.
Marsden, D. 2009. "The paradox of performance related pay systems: why do we keep adopting them in the face of evidence that they fail to motivate?" London, Centre for Economic Performance, London School of Economics.
McNamara, P. 2005. "Quality-based payment: six case examples." *International Journal for Quality in Health Care* 17 (4): 357–63.
Meessen, B., J. Kashala, and L. Musango. 2007. "Output-based payment to boost staff productivity in public health centres: contracting in Kabutare district, Rwanda." *Bulletin of the World Health Organization* 85 (2): 108–15.
Muralidharan, K., and V. Sundararaman. 2009. "Teacher Performance Pay: Experimental Evidence From India." NBER Working Paper 15323. Washington DC, National Bureau for Economic Research.
Murnane, R. J., and D. K. Cohen. 1986. "Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and Few Survive." *Harvard Educational Review* 56 (1): 1–17.
Nagin, D. S., J. B. Rebitzer, S. Sanders, and L. J. Taylor. 2002. "Monitoring, Motivation, and Management: The Determinants of Opportunistic Behavior in a Field Experiment." *American Economic Review* 92 (4): 850–73.
Neal, D. 2011. "The Design of Performance Pay in Education." NBER Working Paper 16710. Washington DC, National Bureau for Economic Research.
Niemiec, C. P., R. M. Ryan, and E. L. Deci. 2009. "The Path Taken: Consequences of Attaining Intrinsic and Extrinsic Aspirations in Post-College Life." *Journal of Research in Personality* 73 (3): 291–306.
OECD. 1993. *Pay Flexibility in the Public Sector*. Paris: OECD.
OECD. 1996. *Pay Reform in the Public Service: Initial impact on pay dispersion in Australia, Sweden, and the United Kingdom*. Paris: OECD PUMA.
OECD. 1997. *Trends in Public Sector Pay in OECD countries*. Paris: OECD/PUMA.
OECD. 2004. "Trends in Human Resources Management Policies in OECD Countries. An Analysis of the Results of the OECD Survey on Strategic Human Resources." Paper presented to the Human Resources Management Working Party. Paris: OECD.
OECD. 2005a. *Modernising Government: The Way Forward*. Paris: OECD.
OECD. 2005b. *Performance-related Pay Policies for Government Employees*. Paris: OECD.
OECD. 2008. *The state of the public service*. Paris: OECD.
OECD. 2009. *Government at a Glance*. Paris: OECD.
OECD. 2011. *Government at a Glance*. Paris: OECD.
Perry, J. L., T. A. Engbers, and S. Y. Jun. 2009. "Back to the Future? Performance-Related Pay, Empirical Research and the Perils of Persistence." *Public Administration Review* 69 (1): 39–51.
Perry, J. L., and A. Hondeghem, eds. 2008. *Motivation In Public Management: The Call Of Public Service*. Oxford: Oxford University Press.
Perry, J. L., D. Mesch, and L. Paarlberg. 2006. "Motivating employees in a new governance era: the performance paradigm revisited." *Public Administration Review* 66 (4): 505–14.
Petersen, L. A., L. D. Woodard, T. Urech, C. Daw, and S. Sookanan. 2006. "Does pay-for-performance improve the quality of health care?" *Annals of Internal Medicine* 145 (4): 265–72.
Pink, D. H. 2009. *Drive: the surprising truth about what motivates us*. New York: Riverhead.
Porter, L. W., and E. E. Lawler III. 1968. *Managerial Attitudes and Performance*. Homewood, IL: Dorsey Press.
Prendergast, C. 1998. "What Happens within Firms? A Survey of Empirical Evidence on Compensation Policies." NBER Working Paper. Washington DC, National Bureau for Economic Research.
Prendergast, C. 1999. "The Provision of Incentives in Firms." *Journal of Economic Literature* 37 (1): 7–63.
Propper, C., and D. Wilson. 2003. "The Use and Usefulness of Performance Measures in the Public Sector." CMPO Working Paper Series No. 03/073. Bristol, UK, The Centre For Market And Public Organisation.
Rexed, K., C. Moll, N. Manning, and J. Allain. 2007. "Governance Of Decentralised Pay Setting In Selected Oecd Countries." OECD Working Papers on Public Governance, 2007/3. Paris, OECD.

Roski, J., R. Jeddeloh, L. An, H. Lando, P. Hannan, C. Hall, and S. H. Zu. 2003. "The impact of financial incentives and a patient registry on preventive care quality: increasing provider adherence to evidence-based smoking cessation practice guidelines." *Preventive Medicine* 36 (3): 291–9.
Ryan, R. M., and E. L. Deci. 2000. "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being." *American Psychologist* 55 (1): 68–78.
Sauermann, H., and W. M. Cohen. 2008. "What Makes Them Tick? Employee Motives and Firm Innovation." NBER Working Paper No. 14443. Cambridge, MA, NBER.
Shen, Y. 2003. "Selection Incentives in a Performance-Based Contracting System." *Health Services Research* 38 (2): 535–52.
Singh, P. 2010. *Performance pay and information: reducing child malnutrition in urban slums*. London: London School of Economics.
Skinner, B. F. 1969. *Contingencies of Reinforcement*. New York: Appleton-Century-Crofts.
Soeters, R., C. Habineza, and P. B. Peerenboom. 2006. "Performance-based financing and changing the district health system: experience from Rwanda." *Bulletin of the World Health Organization* 84 (11): 884–9.
Springer, M. G., D. Ballou, L. Hamilton, V. Le, J. R. Lockwood, D. McCaffrey, M. Pepper, and B. Stecher. 2010. "Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching." Working Paper. Nashville, TE, National Center on Performance Incentives at Vanderbilt University.
Stajkovic, A. D., and F. Luthans. 2003. "Behavioral management and task performance in organizations: conceptual background, meta - analysis, and test of alternative models." *Personnel Psychology* 56 (1): 155–94.
Steel, N., S. Maisey, A. Clark, R. Fleetcroft, and A. Howe. 2007. "Quality of clinical primary care and targeted incentive payments: an observational study." *British Journal of General Practice* 57 (539): 449–54.
Straberg, T. 2010. "Employee perspectives on individualised pay: attitudes and fairness perceptions." PhD dissertation, University of Stockholm.
Vaghela, P., M. Ashworth, P. Schofield, and M. C. Gulliford. 2009. "Population intermediate outcomes of diabetes under pay-for-performance incentives in England from 2004 to 2008." *Diabetes Care* 32 (3): 427–9.
Vroom, V. H. 1964. *Work and Motivation*. Hoboken: Wiley.
Weibel, A., K. Rost, and M. Osterloh. 2009. "Pay for Performance in the Public Sector - Benefits and (Hidden) Costs." *Journal of Public Administration Research and Theory* 20 (2): 387–412.
Wilson, J. Q. 1989. *Bureaucracy: What Government Agencies Do and Why They Do It*. New York: Basic Books.
Witter, S., T. Zulfiqur, S. Javeed, A. Khan, and A. Bari. 2011. "Paying health workers for performance in Battagram district." *Human Resources for Health* 9: 23.
World Bank. 1999. *Civil Service Reform: a Review of World Bank Assistance: Report No. 19211*. Washington DC: OED, World Bank.
World Bank. 2001. *Salary Supplements and Bonuses in Revenue Departments (Final report)*. Washington DC: World Bank.

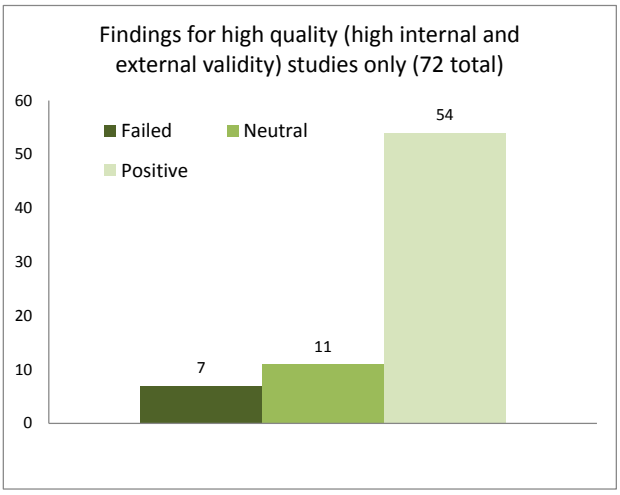
Figures and Tables

Figure 1: Aggregate Findings on Performance-related Pay



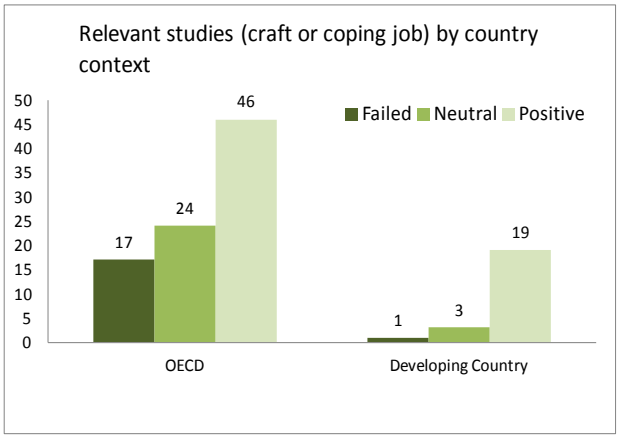
Source: Authors' own findings.

Figure 2: Findings by Research Quality



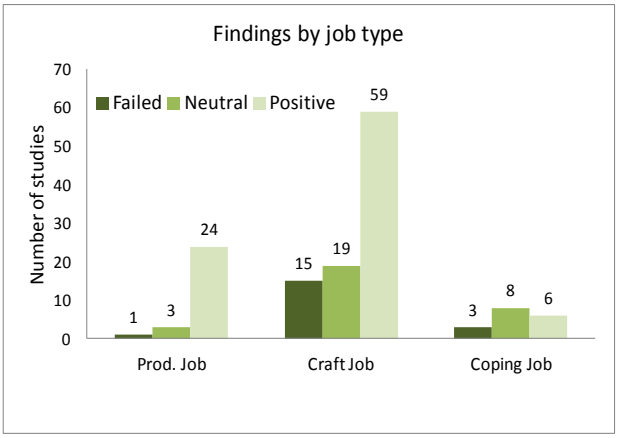
Source: Authors' own findings.

Figure 3: Findings by Country Context



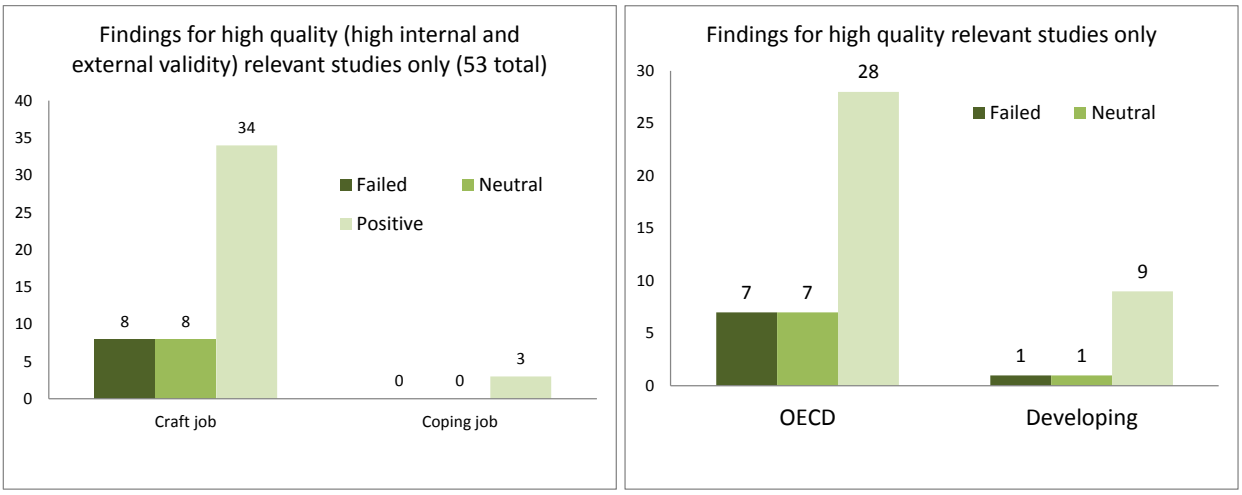
Source: Authors' own findings.

Figure 4: Findings by Job Type



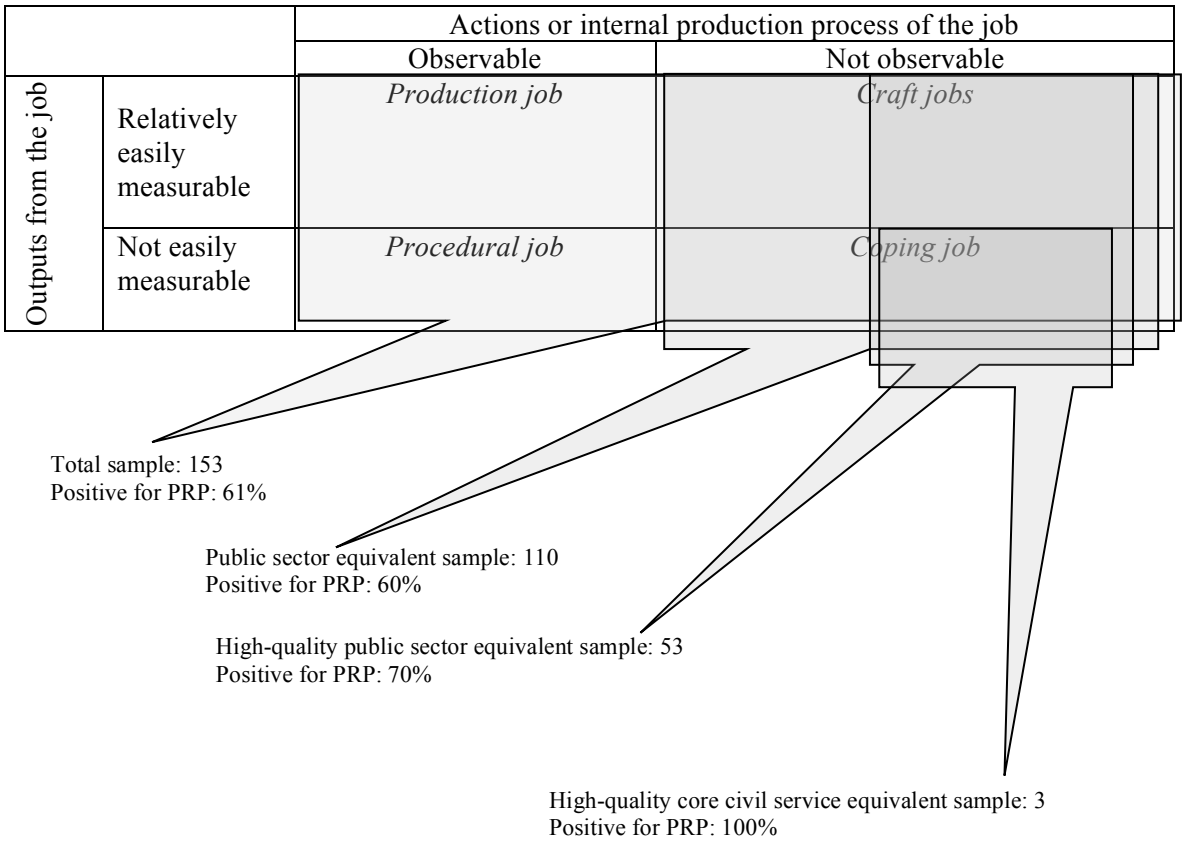
Source: Authors' own findings.

Figure 5: Findings Concerning High-quality Craft and Coping Studies by Sector and Country Context



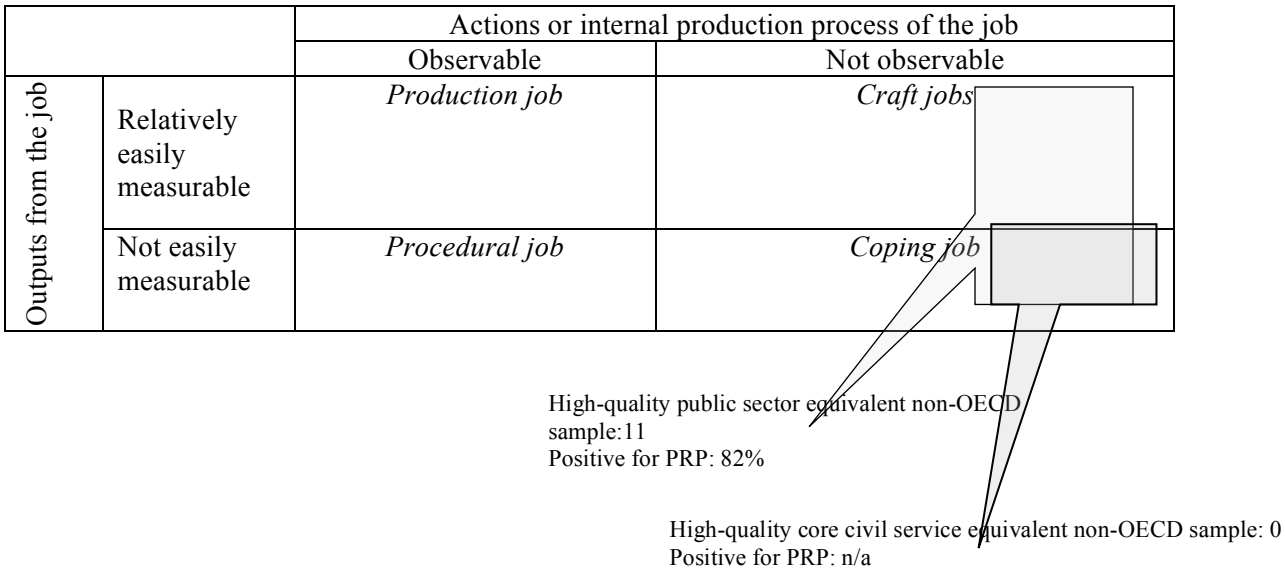
Source: Authors' own findings.

Figure 6: Numbers of Studies and their Findings, Filtered by Job Type and Study Quality



Source: Authors' own findings.

Figure 7: Numbers of High-quality Public Sector Equivalent Studies in Non-OECD Settings



Source: Authors' own findings.

Table 1: James Q. Wilson's Classification of Job Types

		Actions or internal production process of the job	
		Observable	**Not observable**
Outputs from the job	**Relatively easily measurable**	***Production job***: Simple repetitive stable tasks, specialized skills. Examples: Manufacturing, sales, simpler municipal services (garbage collection).	***Craft jobs***: Application of general sets of skills to unique tasks, but with stable, similar outcomes. Examples: Auditing; revenue collection; teaching; medical practice; job placement work
	Not easily measurable	***Procedural job***: Specialized skills; stable tasks, but unique outcomes Examples: Military	***Coping job***: Application of generic skills to unique tasks, but outcomes cannot be evaluated in absence of alternatives Examples: Core civil service; managerial jobs in large private sector organizations

Source: Adapted from Wilson (1989).

Table 2: Studies by Country Environment, Methodology, and Job Type

Country and methodology	Types of Jobs					
	Production jobs	***Procedural jobs***	***Coping jobs***	***Craft jobs***	***Unclassified***	***Total***
OECD study	**27**	**0**	**16**	**71**	**13**	**127**
Observational	14	0	16	58	13	101
Field RCT	7	0	0	13	0	20
Lab. experiment	6	0	0	0	0	6
Developing country study	**1**	**0**	**1**	**22**	**2**	**26**
Observational	0	0	1	15	2	18
Field RCT	0	0	0	6	0	6
Lab. experiment	1	0	0	1	0	2
Total	**28**	**0**	**17**	**93**	**15**	**153**

Source: Authors' own findings.

Table 3: Findings of High-quality craft and Coping Studies by Sector and Country Context

	Craft jobs				Coping jobs	
	Education	*Health*	*Tax*	*Other*	*Public*	*Private*
OECD	**19**	**13**	**2**	**5**	**1**	**2**
Positive	12	8	1	4	1	2
Negative or neutral	7	5	1	1	0	0
Developing country	**6**	**4**	**1**	**0**	**0**	**0**
Positive	4	4	1	0		
Negative or neutral	2	0	0	0		
Total	**25**	**17**	**3**	**5**	**1**	**2**

Source: Authors' own findings.