European Bank
for Reconstruction and Development

Self-worth versus net worth: image motivation and the quantity-quality trade-off

J. Michelle Brock

Abstract
In this paper we study the quantity-quality trade-off among professionals and how this trade-off may change in the presence of concerns over self-image. Concern over self-image (self-reputation) can motivate performance at work (Bénabou and Tirole, 2002), but there is little empirical work to date on self-image as a performance incentive. We use a real-effort framed field experiment to investigate how judges in Tajikistan allocate effort to quantity and quality of work. The treatments include a prime for self-image (anonymous visibility), a performance-based bonus incentive and a combination of the two. We find that subjects respond to both anonymous visibility and bonus incentives with increased quantity of work, and respond to a bonus with worse quality. However, judges are less willing to earn points from poor quality work in a bonus treatment if a peer might see it, albeit anonymously. This study provides evidence on the power of self-image, independent of external reputation, to motivate less opportunistic behaviour, even in a legal system with low accountability. More broadly, this study provides evidence of the economic importance of self-image.

Keywords: social incentives, professionalism, self-image, real-effort task, intrinsic incentives, Tajikistan.
JEL Classification: D83, K00, C91

Contact details: J. Michelle Brock, One Exchange Square, London, EC2A 2JN Email: brockm@ebrd.com.

This work was funded by a Technical Cooperation grant from the European Bank for Reconstruction and Development. We are grateful for the support of the Judicial Training Centre and the Council of Justice in Tajikistan. We acknowledge excellent research assistance from Carly Petracco and Muzaffar Ahunov. We benefited from comments from Catherine Eckel, Sergei Guriev, R'oi Sultan, Nathanial Young, EBRD OCE Seminar Series participants, and participants at several ESA conferences.

Working Paper No. 182 **Prepared in November 2015**

1 Introduction

According to the classic theory on multi-tasking, workers will put as little effort as possible into their jobs and only along the lines of what is in their contract. Effort is limited and workers will allocate effort in order to earn rewards (Holmstrom and Milgrom, 1991). While quantity may be contracted upon, it is difficult to incentivise quantity without compromising quality. But despite incomplete contracts, professionals – judges, teachers, doctors and others – regularly deliver on quality. This suggests additional rewards to effort beyond the explicit contractual ones.

In the public service professions mentioned, social rewards from acting according to professional norms can be important non-contractual incentives to motivate quality. Moreover, explicit incentives are sometimes counterproductive when social motivation is strong. For example, performing a charitable act can be a signal to oneself and others about deeply held values. If a charity offers to pay volunteers, the signal gets muddled – it becomes unclear if the volunteer is acting due to deeply held values or in order to earn money – and effort goes down (Bowles and Polania-Reyes, 2012). The same logic can be applied to professions with a social service dimension. It becomes complicated, then, to design incentive schemes that reward quality without undermining social motivation. The main contribution of this paper is to show how monetary and social incentives compare and interact in a professional setting with multi-tasking.

Peer scrutiny is a powerful example of a social incentive that can influence a professional's quantity-quality trade-off. In a principal-agent context, being watched by a manager can have negative consequences for productivity due to crowding out of intrinsic motivation (Frey, 1993). But unlike monitoring, peer scrutiny in and of itself entails no feedback or remuneration consequences. When we know our work our will be scrutinised by our professional peers, we work harder to make sure it will meet the group's standards (Akerlof and Kranton, 2008). In doing so, we reinforce our commitment to the group values as well as our own professional identity (Ryan and Deci, 2000), which has implications for our utility optimisation. Peer scrutiny and the corresponding "social rewards" thus leverage professional self-image to influence workplace

behaviour.

In this paper we use a framed field experiment to test how social rewards impact professional behaviour. For our treatments we vary visibility of work performance among peers (that is, peer scrutiny), test the relative importance of a competitive bonus and look at the interaction of increased visibility and bonus incentives. Our results provide insight into the relative importance of each incentive for behaviour among professionals. We designed the visibility treatment – a one-shot setting with no feedback and where the audience is uncertain – such that the social rewards are the result of an internal feedback process where the individual acts to conform to norms and thus understands that he has social approval, were it to be explicitly given. We chose a competitive bonus as the high-powered monetary incentive because we expect it to act primarily on volume and a backlog of cases has been identified as an important performance failure in this setting.

We explore these issues using a sample that particularly lends itself to our research aim, namely judges in Tajikistan. Implementation of (potentially) well-designed laws depends in part on how judicial professionals make trade-offs between quantity and quality of their work. Institutional elements that reinforce quality-based reputation or career concerns for judges are absent in Tajikistan, making it an ideal place to look at intrinsic motivation to perform. Preferences for quality that stem from intrinsic sources, like self-image, may be instrumental for maintaining a minimum level of quality in the system. Increasing opportunities to earn social rewards in such an environment may improve quality, by making professional image more salient. But it may also reinforce the status quo, if preferences for quality and professional values cannot produce sufficiently strong social rewards.

The role of social rewards for judges' performance, and the performance of other social service professions' (for example, doctors, teachers, researchers), has important social and economic implications. Judges are responsible for upholding legal institutions and their decisions impact firms' ability to thrive. Ongoing debates about performance in health care and academia centre

around the question of how to achieve a desired volume of output – patients treated or research papers published – without compromising quality. Primary school teachers perform an essential task for society and at the same time receive modest incomes (less in the developing world). Social incentives can play a large role in motivating performance in these professions. As Bowles and Polania-Reyes (2012) suggest, understanding how to leverage social incentives effectively can not only reduce crowding out of intrinsic motivation, but can also complement monetary incentives. This paper contributes to the debate by looking specifically at peer scrutiny as a tool for motivating compliance with existing professional norms for quality.

The paper proceeds as follows. In section 2 we discuss crowding out and self-image in more depth. We include a descriptive model of how judges determine quality in the presence of concerns for self-image. In section 3 we provide a brief outline of the structure of the Tajik judicial system and describe the sample. From there, we describe the experimental design, the empirical specification and the results. A final section concludes.

2 Crowding out and self-image

2.1 Crowding out of intrinsic motivation

This study contributes to the growing literature comparing social rewards with monetary incentives, where image concerns feature heavily. Monetary incentives may crowd out intrinsic motivation for a variety of reasons: they may act as signals from principals to agents about the quality of a task, they may come from distrust between the principal and the agent, or they may reflect image concerns (Bowles and Polania-Reyes, 2012; Frey and Jegen, 2000; Frey and Oberholzer-Gee, 1997; Gneezy et al., 2011). The question is whether monetary incentives to "do well" crowd out intrinsic incentives to "do good" in activities that help others at one's own cost (Ariely, Bracha, and Meier, 2009). Effectively, the incentive structure frames an exchange and may reduce intrinsic motives (Heyman and Ariely, 2004). Thus monetary incentives can act both directly, as an incentive to perform a task, and indirectly, acting to undermine intrinsic motivation.

The bulk of the empirical literature on crowding out examines effort expended doing pro-social activities such as helping a charity or donating blood. The consensus is that image concerns and social rewards matter for how people respond to monetary incentives. One of the most widely cited studies of crowding out explores how introducing a fine changes parents' punctuality in picking up their children from school (Gneezy and Rustichini, 2000). Counter-intuitively, the fine increased late pick-ups. The authors interpret this as crowding out of intrinsic motivation – by monetising the exchange, officials implicitly gave parents permission to be late if they were willing to pay the price. In doing so, officials undermined the social rewards for being on time (or sanctions for being late).

Increased scrutiny can also crowd out intrinsic motivation and reduce effort. For example, increased monitoring by a manager can decrease individual effort by crowding out of intrinsic motivation (Dickinson and Villeval, 2008; Frey, 1993). In a laboratory experiment that uses a similar approach as this study, Ariely, Bracha, and Meier (2009) look at image motivation by

comparing effort under two levels of visibility. In their treatments, some subjects get paid in private for their charitable work while others get paid in public. Given that is a donation good and not a private good, image relies on not working harder for money. They find that those who receive money privately work harder than those who receive money publicly. They also find that subjects work harder in the public condition under no monetary incentive than if they receive a stipend privately. Thus, with the money in the public condition, we see a crowding out of intrinsic motivation.

This result is echoed in Carpenter and Myers (2010), who find evidence of crowding out among firefighters with image concerns. They further find that visibility matters disproportionately for those with image concerns, where previous studies have simply assumed this connection. Theirs is a field study on behaviour of volunteer firefighters in which they evaluate the impact of image concerns on the decision to volunteer as a firefighter (that is, undergo training, etc.) and the willingness to respond to a call (that is, to actually do the job), controlling for whether a monetary stipend is offered. Presence of image concerns is measured at the individual level and captured with whether or not a volunteer had a vanity license plate. Altruism is measured with a dictator game. Carpenter and Myers (2010) find that time spent training is correlated with altruism, while those with image concerns are more likely to show up on a call, "supporting predictions that the effect of image concerns increases with the visibility of the activity".

Our paper complements this literature by testing monetary and social incentives for a real-effort task with professionals in the laboratory. Like Ariely, Bracha, and Meier (2009) we look at how effort changes under different levels of visibility of effort and incentive schemes, but but in their paper self-image is considered a constant and independent of visibility. In contrast, we consider visibility as a tool for increasing salience of the "judge as professional" identity and thus a prime for image concerns that are not present (or salient) absent the incentive. Our set-up also differs from Ariely, Bracha, and Meier (2009) in that we study the power of social-image in a real-effort task, where we offer monetary incentives in all treatments. We pay this piece rate because paying for the task clearly sets it out as work. It builds in the inherent crowding

out between intrinsic and extrinsic motivation in a stark way – there is no way to go above and beyond without getting paid for it. Everyone will earn more money the more effort they put in. Those who are intrinsically motivated to perform will get paid well since correct answers are worth twice as much as incorrect ones. Those with extrinsic motivation may also get paid well if the relative price incentive is high enough. Thus if crowding out of intrinsic motivation occurs differently across treatments, we will see it reflected only as a decrease in total items completed (relative to the control). Crowding out in visibility treatments may also change the ratio of correct to incorrect items compared with the control; while decreasing total items completed, subjects may actually reduce the correct-to-incorrect ratio if they think it shows less ambition for getting money.

2.2 Social rewards and self-image

Understanding how self-image and group identity influence effort in the workplace is particularly important because these factors can compensate for the inability to contract over quality. Using visibility to generate social rewards from self-image is based on self-perception theory. According to self-perception theory, stimulus $\rightarrow$ action $\rightarrow$ belief (Bem, 1972). In this study, the peer scrutiny is the stimulus, the response to that is the quantity and quality effort choice and the resulting belief is the professional identity supported by that action. Self-image can thus be an important driver of professional behaviour (Akerlof and Kranton, 2000; Bénabou and Tirole, 2011; Posner, 1993). For example, self-image at work will be heavily influenced by professional norms, which provide benchmarks for performance quality.[1] When individuals internalise a norm, they will adhere to it as a way to maintain cognitive consistency with respect to their own self-image (Ryan and Deci, 2000). Highly skilled and specifically trained groups, such as the judiciary, teachers and doctors get exposure to professional norms during their training and may even belong to professional associations, which further propagate expectations of behaviour.

[1]Professional norms can be thought of as agreed upon codes of conduct that drive decision-making among the individuals in a group of common expertise, training and occupation.

Thus, people in these groups may get utility from norm conformance or disutility from not complying with norms (Akerlof and Kranton, 2000).

Providing a stimulus for self-image investment can be especially important in the presence of competing incentives or incompletely internalised professional norms (Ryan and Deci, 2000). Competing incentives may include extrinsic rewards or other norms. Individuals may even have competing elements of their self-image such that they make trade-offs between these when deciding behaviour. Psychologists refer to multiple (possibly competing) identities that together make our self-concept. The term "identity salience" describes the fact that at different points in time or in different contexts, an individual's behaviour may be more or less influenced by any one identity (Turner and Onorato, 1999). Identity salience can explain why one may not act according to professional norms in the absence of external stimulus or a reminder to do so (Cohn, Fehr, and Maréchal, 2014; Hoff and Pandey, 2014). In this case, the identity of "the person who likes to conform to professional norms for quality" may not be a salient identity absent a stimulus. Peer scrutiny can serve as such a stimulus, or reminder, to conform to professional norms. With such a reminder, the worker will generate self-image gains from norm conformance. The gains will come from reflecting on having behaved in a way that is consistent with external expectations. Izuma (2012) refers to this internal payoff as a "social reward". Increasing opportunities for such social rewards may thus increase compliance with professional norms for quality.

2.3 A model of worker effort and self-image utility

To clarify ideas, we specify a model where adhering to professional norms delivers social rewards to a decision-maker in the form of self-image utility (as in (Gleitman, 1996), (Bénabou and Tirole, 2003), (Bénabou and Tirole, 2011) and (Bowles and Polania-Reyes, 2012)). In our model, utility is a function of income, Y, effort cost, $c(q,w)$, self-image gains from effort, $S(q,w)$ and self-image gains from earnings, $M(q,w)$.

For simplicity, we take $Y(q,w)$, $S(q,w)$ and $M(q,w)$ to be additively separable functions of quantity (q) and quality (w).[2] We assume that costly effort provides disutility for all workers, but those that enjoy their work more may have different cost functions.[3] The worker's utility function is thus:

$$U(q,w) = Y(q,w) - c(q,w) + \delta(\gamma_S S(q,w) + \gamma_M M(q,w)) \tag{1}$$

where δ is a parameter describing the chance one's work may be observed by peers. γ_S and γ_M are similar to Bénabou and Tirole (2011)'s weights reflecting "the idea that people would like to appear as *prosocial* (public spirited) and *disinterested* (not greedy)". Unlike Bénabou and Tirole (2011) we will allow γ_M to be negative or positive. This is because we are looking at image gains from work effort, rather than from altruistic actions. Earning a high income may be image-positive, but due to the public service nature of the profession, it may also pay to appear disinterested. Thus, it is not clear a priori that a motive to appear disinterested should outweigh self-image gains to earning a lot of money from working. Note that self-image is not a function of reputation. It is the result of an internal assessment of one's type given one's own behaviour.[4]

In the case of judges, quality is defined as appropriate application of the legal code (effort), combined with an absence of rent-seeking (that is, inappropriate application of the code in order to obtain monetary gains). Quantity measures the volume of legal decisions passed during a fixed

[2] In many social service professions, income is a more a function of tenure than performance. In our study we allow income to be a function of performance because, even in the public sector, managers often have leeway to set income and bonuses according to performance, albeit in fixed ranges.

[3] The model is restricted to these three elements for simplicity, but can be expanded to include reputation, time spent at work and time spent at leisure, as in Posner (1993). We exclude reputation from the model to more accurately reflect the situation in the field: self-image is potentially more important than signalling to others when visibility, and thus gains from reputation, is minimal or uncertain. Correspondingly, our experiment is a one-shot game with no labour-leisure trade-offs. Further, all work is anonymous to ensure that there are no opportunities to obtain returns to reputation.

[4] To include reputation the utility function could be
$U(R,S;w) = we + \delta_1(gamma_S S(q,w) + \gamma_M M(q,w)) + \delta_2 \gamma_R R(q,w)$, where δ_1 is an indicator function equal to 1 if visibility is non-anonymous.

period. Were decisions visible in this judicial system, quantity would be the more verifiable of the two dimensions, but since decisions are not visible, judges optimise over quantity and quality according to their own preferences for performance and according to professional norms.

As in self-perception theory (Bem, 1972; Bodner and Prelec, 2003) we assume that an individual's characteristic, or type, θ, is revealed to the individual himself through his own actions, x. The individual then chooses his action to support the belief of the type he wishes to be – he updates his self-image distribution from $p(\theta)$ to $p(\theta|x)$. We depart from the model of Bodner and Prelec in that we only consider the "outcome utility", $u(x,\theta)$, where type, θ is encapsulated by $S(q,w)$. As in Prelec (2011) we consider one's type as given to the individual by the relevant social group. In this case, that group is the legal profession.[5]

Our separable term for self-image utility is also similar to "psychological costs of cheating" in Nagin et al. (2002). In their paper, the authors mention three models to describe opportunistic behaviour: rational cheater, conscience model and impulse control model. We investigate only the second – the others are not accommodated by the experimental design and are thus not possible confounds. Nonetheless, our model is not unlike that presented in Nagin et al. (2002), which includes a term for psychological costs of cheating, $\chi(c)$. Akin to our our self-image term, $\chi(c)$ is "the sole constraint on opportunism". Unlike our model, however, $\chi(c)$ does not include psychological gain from success. Finally, the Nagin and co-authors experiment considers the gains to various degrees of monitoring, thus looking implicitly at these psychological constraints. We, on the other hand, explicitly vary the salience of the psychological constraints and consider how this impacts the trade-offs workers make. Within this framework, we would expect self-image concerns to lead to changes in behaviour, even if visibility is anonymous. Specifically, an increase in δ would increase both q and w. The nature of the trade-off between increases in q and w we leave as an imperial question.

[5]Bodner and Prelec (2003) also include a diagnostic utility in their model: total utility = outcome utility + diagnostic utility = $u(x,\theta) + \sum_{\theta} p(\theta|x)V(\theta)$, where $p(\theta)$ is the uncertainty of person *we* being type θ, conditional on choosing x. Also related is Andreoni and Bernheim (2009)'s signalling model. In their model the agent cares about *others'* perception of his type rather than his own perception of his type. In this case, the agent acts fairly to signal his type to the other.

3 Sample and field context

We investigate these ideas among judicial professionals in Tajikistan. Tajikistan is a middle income former Soviet country in Central Asia, with a GDP per capita of USD$2300. Sharing borders with Afghanistan, China, Kyrgyz Republic, Pakistan and Uzbekistan, it is mountainous and transport between regions is difficult. According to the Polity IV index, Tajikistan is a relatively autocratic country. A civil war from 1992 to 1997 complicated transition to democracy and further damaged the economy. In the World Bank's *Doing Business* 2014 report, it ranks 143rd of 189 countries. Transparency International ranks it as 152nd of 175 countries and territories for perceived levels of public sector corruption, in company with the Democratic Republic of the Congo.[6] In a 2001 study on judicial systems in the transition region, Colman (2011) finds Tajikistan to fare the worst among the eight reviewed countries, which included the Kyrgyz Republic, Moldova, Mongolia, Russia and Ukraine.

Tajikistan has a civil law legal system. It is comprised of the Constitutional Court, Supreme Court, Supreme Economic Court and a number of regional, city and district courts. The Constitutional Court decides on the constitutionality of legislation and presides over all other courts in the country. Next in the hierarchy are the Supreme Court, which oversees 3 regional courts and 68 city and district courts, and the Supreme Economic Court, which oversees 4 regional economic courts.[7] Prerequisites for becoming a judge include 5 years of legal education (this is equivalent to a 5-year undergraduate degree with a major in law and legal studies in the United States), 3 years of experience and being at least 25 years old. Students must also pass a qualification exam and successfully complete a 1-year internship. Judgeships are appointed by an examination board. Judicial appointments in Tajikistan are for 10 years, with the possibility of renewal. Appointees can expect to earn the average public sector salary (that is, comparatively low) and receive additional fringe benefits such as a house and discounts on utilities (American Bar Association, 2008).

[6] 2014 Corruption Perceptions Index, http://www.transparency.org/cpi2014/results

[7] There also exist four military or garrison courts that solely deal with cases related to the military.

Our sample for this study includes 103 commercial and general jurisdiction court judges, 57 from in and around Dushanbe, the capital city, and 46 from rural areas. At the time this constituted approximately 40 to 50 per cent of the judges in the country that hear commercial cases. The average participant is a 39-year-old male with 5 years of legal education and 12 years' experience in the legal sector. This is reported in Table 1. We estimate that judges see around a dozen cases a month, including small procedural matters.

Judicial decisions in Tajikistan are not publicly available, there are no independent legal publications in Tajik and "quarterly" newsletters from high courts are published only sporadically (American Bar Association, 2008). Accordingly, there is little opportunity for judges to establish a quality-based reputation, except perhaps within multi-judge courts based on informal information exchange. Further, extrinsic incentives for quality are poorly defined. The courts operate under a tenure-based pay scale and, anecdotally, the qualifications that are considered for reappointment or promotion are not well defined. There is no system in place to track the number of decisions passed down by each judge or the quality, even within a system that suffers from a serious backlog of cases. Further, the Council of Justice has no requirement for number of decisions to be passed per year by a judge. There is no statute of limitations for deciding a case once it has been filed, so judges cannot simply let cases die. As one might expect given these institutional details, volume performance in the country is low and Tajik speed of justice is poor (Colman, 2011).

It is a highly politicised environment, where making quality decisions may be a lower priority than making politically prudent decisions. The system is also known for being corrupt (American Bar Association, 2008; Bertelsmann Stiftung, 2014). The EBRD and World Bank's Business Environment and Enterprise Performance Survey (BEEPS) 2012 finds that only 35 per cent of firms interviewed from a nationally representative sample agree with the statement that "the court system is fair, impartial and uncorrupted". Once judges are appointed they face uncertainty with regards how performance is measured and how to secure reappointment (American Bar Association, 2008). The reappointment process is not governed by any law, lacks any objective

measure, and thus lacks transparency.

Such uncertainty and low accountability can be highly corrosive to professionalism and quality (Hackenbrack and Nelson, 1996; Salterio and Koonce, 1997). For example, Campante, Chor, and Do (2009) shows that the risk of not getting reappointed can lead to strategic decision-making, where the main beneficiaries of decisions may be the government or the decision-maker himself. This strategic decision-making is apparent in any case involving the government, where the lack of independence in the judiciary makes the possibility of a foreign or domestic firm winning a case against the government is nearly impossible (European Bank for Reconstruction and Development, 2012).

This context is not unique to Tajikistan and begs a review of the question "What do judges maximise?" (Posner, 1993). In this setting, self-image may be key to motivating professional behaviour and determining the effectiveness of legal development policies. Since there are so many factors that can undermine professionalism in the field – lack of visibility, job uncertainty, low fixed-rate salaries and capture by the executive branch – using the lab to identify whether or not quality may be valued within the profession is an important first step for understanding the potential power of increased visibility for improving decisions. While the presence of shirking, strategic decision-making and corruption indicates a lack of professionalism, it does not preclude it. Some judges will perform well regardless of the circumstances. Nonetheless, professional (or individual) norms concerning quality, albeit comparatively weak, are critical for the success of any program of increasing visibility to improve the judicial system.

3.1 Experimental design

Following the convention of Izuma, Saito, and Sadato (2010) and Ariely, Bracha, and Meier (2009) we use anonymous peer observability (audience effects) to measure the impact of self-image on performance. The literature shows that people behave differently when they think or

know that they are being watched. An audience can increase generosity and cooperative behaviour (for example, Hoffman et al., 1996), even when the watcher is only suggested (Haley and Fessler, 2005; Rigdon et al., 2009). Psychologists suggest that this is due to the internally generated social rewards from behaving consistently with social norms (Gleitman, 1996; Izuma et al., 2010). Results extend beyond generosity and cooperative behaviour: providing an audience has been shown to decrease anti-social behaviour (**?**), influence purchasing decisions (Kimura et al., 2012) and to increase the likelihood of voting (Panagopoulos, 2014). Thus anonymous visibility may also incentivise effort among professionals, where there exists clear professional norms that put value on quality. Our treatment is therefore an invitation for participants to look into Gleitman's mirror (Gleitman, 1996).

We use lab experiments since the research question cannot be addressed using available field data. Since all judges faced the same tasks, the experiment provides information that we cannot observe in the real workplace setting, where all cases are different. The experiments took place in August 2013 in a classroom at the Judicial Training Centre in Dushanbe, Tajikistan. The experiment consisted of a real-effort task, a multiple price list activity (hereafer referred to as MPL) (Holt and Laury, 2002) to elicit preferences over ambiguity and a post-experiment survey. All activities were conducted using paper and pen. We reviewed informed consent and experiment instructions verbally and on paper before the start of the experiment. In all sessions, work was completely anonymous; judges were given a numeric ID card upon entering the classroom and used stickers to affix this ID to each worksheet they filled out. Performance on the real-effort task and the MPL exercise determined final earnings and judges were fully briefed on how payments would be determined before the experiment began.[8]

The experiment was a two-by-two between subjects design, summarised in Table 2. Each session had one of four treatments: no peer review (*control*, C), positive unknown probability of peer

[8] At the beginning of the experiment, subjects were instructed on the importance of their numeric ID card; in order to collect the correct payment at the end, subjects had to present their numeric ID to the experimenter. The experimenter could then match the ID with the points earned, which had been entered into an Excel file. Hence, any task form could not be associated with any individual, even at the moment of payment. In this way we maintained anonymity throughout the session.

review (*visibility*, T1), and each of these interacted with a tournament bonus incentive (*bonus*, T2, and *visibility+bonus*, T3). For the peer review treatment, instructions included a statement that there was a chance that subjects' work would be viewed by peers at the end of the session, but that their work would remain anonymous. This statement was also written on a blackboard and we called attention to this detail while reviewing instructions verbally. We also demonstrated how completed task forms would be chosen – by randomly drawing a number of forms from the pile of completed forms. Neither subjects nor the experimenter knew the probability that any given form would be selected for review. We thus refer to the treatment as uncertain peer review rather than risky peer review. For treatments without the possibility of peer review, this detail was absent.

All reviews occurred at the end of the experiment. The review consisted of publicly randomly selecting some completed and graded worksheets and making them available for subjects to go over on their own while we made the subject payments. Bonus earnings were not marked on the worksheets and remained the private information of those who won the bonus.[9]

Thus *visibility* is a manipulatation using what Haley and Fessler (2005) refer to as "subtle cues of observability"; we prime subjects to consider what others might think of their behaviour, while maintaining all subjects' anonymity. Since all tasks were completed before any review occurred, we measure the impact of the *possibility* of peer review, not the impact of the review itself. Our work goes further than the eye spots studies, however, in that our potential observer is an actual person. The potential observers, moreover, share a group identity with the subject.

For the bonus treatment instructions included a modification of how earnings would be determined. In the section "How will our earnings be determined" judges learned that the top three point earners of the session would get a 50 point bonus. We chose a tournament bonus because we wanted to incentivise volume and did not have a prior on a meaningful threshold level to set

[9]Graded worksheets had pluses marked next to each correct answer and minuses marked next to each incorrect answer. The top of the graded worksheet listed points earned from correct items, points earned from incorrect items, and total points earned.

ex ante. Further, a competitive bonus rather than a threshold bonus requires subjects to consider the performance they expect from their peers, thus creating a parallel with the self-reputation incentive where subjects also may reflect on what their peers are doing.[10] Note that the bonus was based on points earned (see below for a description of how points were earned) and not number of items completed correctly.

In all treatments subjects completed a real-effort task. The task entailed filling in blanks from excerpts of the Tajik commercial code, where we had removed key words or phrases. Excerpts were extracted from the Tajik commercial code and blanks covered a range of difficulty. A copy of the Tajik commercial code was provided. Thus the task mimicked key aspects of the judges' real work environment: it used relevant subject matter, there was no required minimum or maximum effort and judges could use references as desired. Simultaneously, we made sure that the task reflected standards in the literature. It was a mundane task and real effort was required to get points. It was simple enough that anyone could succeed, but impossible to complete in the time provided, thus allowing for variation in skill.

Subjects had 8 minutes to complete as many items on the worksheet as they could. Worksheets were designed to be impossible to complete all items correctly but such that anyone could have completed at least one item correctly. Subjects earned 5 points for any attempt at filling in the blank, and an additional 5 points if they did so correctly. This simplifies to 10 points for each correct answer and 5 points for each incorrect answer. Subjects earned 1 Tajik somoni (0.15 euros) for every 10 points. Finally, subjects could choose to answer some items correctly and some incorrectly.[11] Note that unlike the "rational cheater" model of motivation, the wrong answers in this experiment are not considered cheating as quality is fully observable, points are given for effort on both correct and incorrect answers and the agents themselves are the only beneficiaries of the effort (or lack thereof).

[10] The competitive aspect of the incentive may have led some subjects to self-select out of the tournament and thus provide lower than average effort.

[11] Participants completed an incentivised practice round so that they could familiarise themselves with the task. Data from the practice round are not included in the analysis.

After the real-effort task, while completed forms from the task were being scored, subjects completed the MPL activity and the survey. For the MPL activity, subjects made 11 choices. Each choice involved a trade-off between a safe payout and a payout that would be determined by a lottery. To determine the outcome for each lottery, we drew red and black balls from an urn. We capture ambiguity preferences by not specifying the contents of the urn. Subjects knew that there were red and black balls, but they did not know how many there were of each. Subjects were paid for each choice to minimise confusion. Unfortunately the subjects did not fully understand the MPL activity and we observe multiple switch points or no switch points for all but three subjects. Subjects did, however, understand that each choice was between a safe and uncertain lottery. Thus, we use as our measure of ambiguity preferences, r_i, the fraction of times the subject chose the lottery to determine their earnings in the MPL activity.

The post-experiment survey was not incentivised. It included items on demographics and asked subjects about strategies they may have used for completing the task. It also included a set of 12 items to measure grit (Duckworth et al., 2007).[12]

All payments were made at the end of the sessions. There was no feedback between rounds, so participants did not learn how many points (that is, how much money) they or their peers earned from round to round. Further, we collected all forms at the end of each round. Experiments were conducted in Tajik, with written instructions available in Russian.

A final note on the experimental design: we test what is essentially a blunt instrument. Subjects are told their peers present in the room may view their anonymised work. This anonymity is what ensures that we measure self-image and not external reputation building. But visibility is uncertain, as is who the peers are that will view any individual's work. Further, the review process itself is not articulated in advance; subjects are not told what details their peers will learn or if the review will be done as a group. The only definition is that the review is uncertain and anonymous.

[12]The grit score is the average response from a 12-item survey.

3.2 Empirical specifications

To look at the impact of the treatments on effort choices, we compare the performance on total number of items completed, number of items completed correctly, and number of items completed incorrectly between treatments. If subjects adhere to a professional norm that puts value on quality, triggering self-reputation should increase the number correct relative to other treatments. If self-image utility from quantity tends to be positive, we will also expect to see increases in incorrect items. At some quality threshold, subjects may experience diminishing marginal returns to self-image over quality or may face a capacity constraint. In that case, subjects may exert effort instead to maximise quantity, which would lead to increases in incorrect answers. Looking at judges by type, we predict that ambiguity-averse judges will react more strongly to the possibility of others seeing their work. Finally, as per classic theory on piece-rate incentives, we expect that the bonus on achieving the highest total number of points will act as an indirect incentive for subjects to maximise quantity over quality and thus expect large increases in incorrect answers for this treatment.

We also examine the question of how the different incentives impact the likelihood of a correct answer. This is a pertinent policy question. Users of the legal system rely on having a high probability of a just outcome, with reasoning behind decisions properly based in the legal code. In this sense, the probability of a correct answer is the most important outcome variable to consider. It may also be the more relevant choice variable when considering a quantity-quality trade-off; judges may have preferences over their average performance that do not change across treatments. Thus, while the number completed goes up or down, the per cent correct may remain constant. This would suggest that there is a well defined professional norm on average performance such that different incentive schemes can only act on quantity.

Finally, we look at the trade-off between quantity and quality. In the multi-tasking literature, agents are hypothesised to put more (or all) effort into the easily measurable task. While we can perfectly observe both quality and quantity in our experiment, quantity is easier to verify and

requires less effort to produce. We hypothesise that the trade-off is more salient under the bonus incentive, where there are disproportionate gains to incorrect answers.

We first look at the treatment effects using Mann-Whitney tests, and complement this with OLS regressions using bootstrapped standard errors. With the OLS analysis, we obtain estimates from using only treatment dummies on the right-hand side as well as estimates that include controls for subject and session characteristics. The regressions are as follows, where $Performance_i$ is either number correct, number incorrect, total items completed or the ratio of number correct to total items completed:[13]

$$Performance_i = \beta T_i + \gamma Z_i + \delta X_i + e_i. \tag{2}$$

The regression for the quantity-quality trade-off is:

$$Numberincorrect_i = \alpha(Numbercorrect_i) + \beta T_i + \gamma Z_i + \delta X_i + e_i. \tag{3}$$

T_i is a vector of treatment dummies, with the control as the excluded category. Z_i is a vector of judge characteristics including age, sex, education, seniority (years working in the legal sector), a dummy for being relatively ambiguity loving and a measure of grit.[14] Importantly, in sessions with more high-ranking judges, the response to potential peer review may be greater (monotonic shift) so we control for rank, education and age. X_i represents the day on which the subject participated to control for any information passed between subjects across days.[15]

Our visibility treatments embody two kinds of ambiguity: unknown positive probability of in-

[13] Other econometric specifications that account for non-normal error terms or count data yield the same results.

[14] While we included all 12 items from the grit survey on the post-experiment survey, we found non-response on a few items to be correlated with treatment. Since the grit score is validated using all 12 items, the grit score was missing for these respondents. To avoid missing responses that are not random, we do not utilise the full grit score. Instead, we use one item from the questionnaire –"I am a hard worker" – as it directly captures the idea of perseverance, it is strongly correlated with grit score (Pearson correlation of 0.66 and $p<0.001$) and non-response on this item was minimal and not correlated with treatment. Other items on the grit questionnaire are less straightforward to interpret in isolation and/or do not correlate as highly with the final grit score.

[15] We cannot control directly for session effects using session dummies because not all treatments were repeated across multiple sessions. Results do not change if we control for treatments that were repeated.

formation being posted, combined with not knowing who may see any given decision. Different peers' opinions will have different value and peers have different abilities to verify the quality. Thus, a possible confounding influence on performance differences between treatments is ambiguity aversion. Subjects who are more ambiguity averse may have a stronger response to the visibility treatment. Since ambiguity attitudes are well balanced across treatments, we do not expect any level effects.

We thus test the difference between total number completed, number correct and number incorrect across treatments. We also consider the policy question of how likely a correct answer will be under the different incentive schemes. Finally, we look at the trade-off of effort put on quantity versus quality across treatments.

4 Results

Results are summarised in Table 3. First, it is important to review whether the incentives functioned properly, since pay-for-performance is not part of the judicial system in Tajikistan. Recall that correct answers were twice as valuable as incorrect answers. Since it was impossible for subjects to complete the form with all correct answers in the time provided, a profit maximiser facing a capacity constraint may have chosen to provide a mixture of correct and incorrect answers; the incorrect answers satisfy preferences for quantity and are also "quick wins" for earning points. While quality is poor in general, with half of the sample scoring below 85 per cent correct in the control, there is considerable variation (individual values range from 0 to 100 per cent correct). Also, across treatments, subjects earned most of their points with correct answers. In all treatments we see subjects providing both correct and incorrect answers (the average number of items completed, items correct and items incorrect were 19.34, 10.36 and 8.98, respectively). The fact that we see correct answers, which have a higher effort cost, in all treatments appropriately reflects the relative power of the incentives. Further, subjects occasionally provided all correct but never chose to present all answers as incorrect. There were subjects who provided only correct answers in all treatments except for *bonus*, where all subjects provided a mixture of both correct and incorrect items. Thus, we conclude that despite the relative inexperience with experiments and with piece-rate incentives, subjects took the task and the incentive seriously.

We now move on to discuss the comparison of the treatments. Figure 1 displays the difference in total, correct and incorrect items completed across treatments. Each bar represents the average number of items completed in each session. This is broken down by average number correct (bottom portion of the bar) and average number incorrect (top portion of the bar). Quantities of items completed are both higher in the treatments than in the control and display greater variation, which could reflect heterogeneous responses to the various incentives, by type. But quantity completed does not go up as much in *visibility* and *visibility+bonus* as it does in *bonus*. This is due to a large increase in average number of items completed incorrectly in *bonus*. *Visibility* and

visibility+bonus also outperform *control* with respect to the average number of incorrect items, but it is most pronounced in *bonus*. The average number of correct items is higher in *visibility* and *visibility+bonus* than in *control*, but not higher in *bonus*. Finally, we fail to reject the hypothesis that average number correct differs between the two peer review treatments (Mann-Whitney test, p=0.43). Mann-Whitney results confirm that the differences mentioned are statistically significant, as shown in Table 4. Results are reiterated in the regression analysis, and also prove to be robust to adding subject and session characteristics. Regression results are reported in Table 5. We thus have our first result.

Result 1: ***(visibility improves productivity)*** *Self-image incentives improve productivity relative to the control, with increases in both correct and incorrect items. In contrast, a competitive bonus incentive motivates an increase in incorrect items only.*

Aside from the striking differences in the total items completed, *bonus* motivated a large decrease in the per centage of items completed correctly compared with the other three groups. Results appear in Table 4 and Table 6.[16] In *bonus* we see that subjects provided 1.6 incorrect answers to each correct answer, compared with a 0.44 ratio in *control*. The ratio is middling in *visibility* and *visibility+bonus* (0.78 and 0.60, respectively) and not significantly different than in *control*. Interestingly, the per cent correct does not differ between *visibility* and *visibility+bonus*, nor does it differ when comparing these two treatments with *control*. Anonymous visibility may not trigger higher quality compared with the control, but adding visibility to a bonus results in behaviour that is indistinguishable from visibility alone. This leads to our second result.

Result 2: ***(crowding out of extrinsic motivation)*** *Self-image concerns impact per cent correct only when combined with a relatively high powered competitive bonus. Self-image concerns thus enforce a quality standard that would otherwise be ignored in the presence of a bonus.*

It is worth mentioning that female judges and judges with more experience tend to have higher per centages correct. Looking back at Table 5 we can see that this is not due to answering more

[16] One subject chose to complete zero items and so their ratio of correct to total is undefined.

correctly, but instead is due to answering fewer incorrectly.

We now turn to investigating the quantity-quality trade-off across the treatments. This is partly informed by comparing performance on per cent correct. More interesting is the correlation between the number that subjects answer correctly and the number they answer incorrectly. The latter correlations are in Table 7, where we use interaction terms to look at how this correlation differs between treatments. We see that the number correct and the number incorrect are significantly negatively correlated on average. In all treatments people with more correct answers tend to also have fewer incorrect answers. This suggests a trade-off between quantity and quality. But the trade-off is not as we would expect from solving a straightforward profit maximisation problem. If subjects were simply maximising profit subject to a time constraint, and given the pay-offs, we would expect the coefficient on *Correct* in Table 7 to be equal to -2. This can be seen if we let profit be $10C+5I$, where C is the number correct and I is the number incorrect. Let $T = T_C C + T_I I$ be the time constraint, where T is total time available, T_C is time spent on each correct answer and T_I is time spent on each incorrect answer. Given this decision problem, we would expect subjects to spend twice as much time on correct answers as incorrect answers (that is, $1/2 = T_I T_C$). Solving the time constraint for I in terms of C then gives $I = T/T_C - T_C/T_I C$ or $I = T/T_C - 2C$.[17] But we do not see this in the data. Instead, we see from our regression of *Incorrect* on *Correct* that the coefficient is -0.3; on the margin, decreasing the number correct by 1 results in 0.3 additional incorrect answers. Further, since most subjects complete more correct than incorrect, it is clear that we cannot interpret this coefficient as the ratio of time spent on each correct answer to time spent on each incorrect answer. If we frame this in terms of choosing quantity and quality, where both quantity and quality are linear combinations of C and I, we obtain the same result. This leads to our fourth result.

Result 3: *Correct and incorrect items appear to be substitutes, but the trade-off cannot be described by standard profit maximisation.*

[17] We obtain the same result if we allow subjects to choose an effort level to achieve a certain probability of getting any given item correct, rather than choosing how many to complete correctly and how many to complete incorrectly.

From this we conclude that the optimisation problem subjects face in this experiment is likely more complex than basic profit maximisation. In particular, the profit maximisation problem ignores that subjects' optimisation problem may include self-image. Self-image utility can explain the apparent limit on willingness to earn points from incorrect answers. In a world where subjects have positive self-image utility from quality, for each additional item completed they decide the effort to put forth so as to achieve a quality that maximises the self-image payout. Meanwhile, a negative self-image value to appearing greedy leads subjects to replace each foregone correct with a disproportionately small number of incorrect items. In our experiment, the limit to profit maximisation is the same in the control and in the visibility treatments, but it is different in the bonus treatment. In the bonus treatment, subjects are more willing to gain more points from incorrect answers. We see from the results in columns 3 and 4 of Table 7 that in *bonus* this trade-off approaches 1. This means that more people are willing to have a higher incorrect-to-correct ratio in *bonus*. The range of ratio in *bonus* is (0.37 to 46). In the control it is (0 to 5). On average, this is a threefold increase in willingness to substitute foregone correct answers with incorrect answers. We also see this in the significant decrease of per cent correct in *bonus* compared with the other treatments. Note, however, that this is still lower than the expected ratio of 2 to 1, suggesting that even in this treatment subjects hold back – it is possible that they correctly estimate their peers' willingness to get points from incorrect answers and hem in their own opportunistic behaviour accordingly.

Result 4: *The private bonus treatment increases the willingness to trade off quality for quantity, but the trade-off still falls short of the pure profit maximising choice.*

This result holds even when we control for observables that will be correlated with ability to answer items correctly, such as education and years of experience. It also holds if we exclude outliers, who may be strongly driven by the competitive nature of the bonus and thus driving the result. The trade-off in *bonus* is robust to adding controls. The average correlations across treatments becomes insignificant when adding controls because the standard error on this variable more than doubles. Nonetheless, we can conclude that while the visibility treatments do not

impact the average quantity-quality trade-off, the bonus treatment does. This can be attributed to an increased willingness to fill in incorrect answers in the bonus treatment, presumably to gain quick and easy points and compete more for the prize.

Finally, with this sample we contribute to a small body of literature on judge performance in former soviet countries. Dimitrova-Grajzl et al. (2012, 2015) investigate productivity among judges in Slovenia and Bulgaria, respectively, using actual judge-level data on performance. In Slovenia, they find that volume is correlated with salary and that judges also increase volume if they think they will be up for promotion. These results echo our own findings – judges unambiguously respond to increased incentives by upping their volume. By using these larger datasets, the authors are also able to draw conclusions about heterogeneous response to incentives. For example, they find evidence of a quantity-quality trade-off in district courts in Slovenia, but not in lower courts. In contrast, they do not find evidence of a quantity-quality trade-off in Bulgaria. While we cannot look at heterogeneous treatment effects due to our sample size, but, unlike these papers we are able to directly observe causal implications of different incentives and directly observe how this impacts quantity and quality.

In summary, we find that social incentives do motivate more effort, but the quantity-quality trade-off does not appear to be affected by the presence of social rewards. Priming self-image concerns alone does increase quantity, but does not increase quality (measured as per cent correct). The (private) bonus treatment, on the other hand, overemphasises gains to incorrect answers that are easier to complete in a fixed time frame - here is the only place we see the quantity-quality trade-off suffer. The impact of the bonus on incorrect items completed is approximately four times the effect that visibility has on completing correct items. per cent correct is higher in all treatments compared with the *bonus*. This means that the probability that a court case will be handled well, instead of being handled quickly with errors, predictably decreases under high powered competitive bonus incentives. Adding self-image to a bonus incentive scheme mitigates the opportunism we see in the private bonus treatment and is not discernible from social rewards alone. Our results are consistent with a story where negative self-image utility from profit maximisation effectively

increases the cost of an additional incorrect answer, while positive self-image utility from being productive increases total items completed without sacrificing average quality. Finally, that visibility reduces the effect of the bonus suggests that social rewards can in fact crowd out extrinsic motivation along the socially undesirable dimension.

5 Conclusion

For this study, we explored how social rewards and money motivate judicial professionals to expend effort on quality and quantity. The existing literature suggests that the social rewards from self-image investment can motivate behaviour that deviates from selfishness and low effort equilibria (Bénabou and Tirole, 2002; Haley and Fessler, 2005). But this has been sparsely explored in more contextualised settings. We investigate this topic using a framed field experiment with a sample of judges in Tajikistan. Tajikistan is an ideal context to study these ideas. Creating incentives for quality and professionalism in a setting such as Tajikistan is difficult. Once judicial decisions have been passed down, they are not open for peers or the public to access. Further, income is tenure-based and promotion is not linked directly to performance. Thus, conscientious policy-makers must seek alternate ways to motivate workers. Leveraging and building judges' professional self-image may be one option for improving effort. In particular, priming professional identity can be an important element of promoting professionalism. The results from our study can inform researchers and policy-makers on the relative power of self-image incentives for motivating improved effort among professionals.

Both monetary and social incentives in our experiment increase effort put towards quantity. Under the social incentive, the additional effort does not come at the cost of quality. The bonus incentive, on the other hand, is very bad for quality. Extrapolating to the real world, such bonus incentives could actually decrease the chance that any given commercial law decision is of high quality, without a complementary social incentive. Also from a user perspective, visibility alone may be an appropriate policy option when judges display limited preferences for quality: the chances of a fair outcome are no worse and the volume of cases considered in a given time frame could be better, thus improving the speed of justice.

Social and bonus incentives together better motivate quality than a pure money market with a bonus. In fact, behaviour in the mixed treatment is indistinguishable from behaviour in social reward treatment with no bonus. Social incentives thus appear to crowd out motivation to earn

the bonus. Note that our results from professionals in a work environment correspond in part with Heyman and Ariely (2004)'s work on charitable acts, which finds that mixed markets more closely resemble money markets. This is reflected in the lack of difference in per cent correct in the control (pure money market) and the social rewards treatments (mixed market). But in contrast to the conclusion that money markets resemble mixed markets, we find that mixing the social incentive with the high-powered bonus more closely resembles the the visibility treatment (a mixed market) than the control, a pure money market. So while it is not advisable to implement the competitive bonus, per se, if considered it must be accompanied by additional incentives for professionalism.

Our results support the idea that self-image can influence performance and under certain conditions it requires a stimulus. Compared with the control, subjects respond to visibility with higher effort on both correct and incorrect answers. But while suggesting visibility increases productivity, it does not impact average quality. According to Ryan and Deci (2000), the intrinsically motivated will not change in the presence of the incentive, so we can conclude that the dominant type in the group is one who has internalised professional norms for average quality but not productivity. Meanwhile, a bonus incentive, if offered without additional social incentives, is absolutely corrosive for quality. If a social incentive is added to the bonus, the effect is neutralised and standard levels of quality obtain. This is consistent with a model that includes positive self-image value from effort and negative self-image value for appearing greedy, with a norm of quality that requires external stimulus in the presence of high stakes monetary rewards.

These results should be carefully considered in any policy aiming to increase visibility. Visibility policies are usually not anonymous, so there are additional reputational incentives to take into account. Concerns over reputation may create enough social pressure to motivate improvements in average quality that we do not see in the anonymous setting. And, on the one hand, reputation may make career concerns more salient, motivating competition over quality that is not present otherwise. On the other hand, if the self-image response to anonymous visibility reflects professional norms, as the theory suggests, it is not clear that reputational incentives will call for a

different quantity-quality trade-off than exists presently.

We note that increasing anonymous visibility in essentially a reduced form treatment. The message is sufficient to trigger self-image investment, but we cannot disentangle what elements of the message do this. With our model we assume that this is due to preferences over both quality and quantity, where quantity is emphasised because it is more easily verified by peers. But it may also be capturing career concerns, if the message additionally cues the belief that not only peers, but also authorities, will see the work.

Our results best apply to the work place for the marginal worker, who prefers to conform to quality norms but is too often uninspired to do so. In the real work place, increasing volume may require a different kind of effort and gains from rent-seeking may be a relatively large portion of income. But according to the theory, the social stimulus is a motivator that should work across contexts. Undoubtedly some judges persistently perform better than others, especially those for whom the professional identity is more often salient. For others, monetary opportunity costs may act as a constraint on norm conformance – the preference may be to perform according to professional norms, except when costs are high as judges may determine it is not worth the effort. But for the marginal worker, reminders of professional norms may hold the key to unlocking norm conformance.

References

G. A. Akerlof and R. E. Kranton (2000), "Economics and identity", *Quarterly Journal of Economics*, 715–753.

G. A. Akerlof and R. E. Kranton (2008), "Identity, supervision, and work groups", *The American Economic Review*, *98*(2), pp. 212–217.

American Bar Association (2008), *Judicial Reform Index for Tajikistan*, Washington, DC: American Bar Association.

J. Andreoni and B. D. Bernheim (2009), "Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects", *Econometrica*, *77*(5), 1607–1636.

D. Ariely, A. Bracha, and S. Meier (2009), "Doing good or doing well? Image motivation and monetary incentives in behaving prosocially", *The American Economic Review*, 544–555.

D. J. Bem (1972), "Self-perception theory", *Advances in Experimental Social Psychology*, *6*, 1–62.

R. Bénabou and J. Tirole (2002), "Self-Confidence and Personal Motivation", *The Quarterly Journal of Economics*, *117*(3), 871–915.

R. Bénabou and J. Tirole (2003), "Self-knowledge and self-regulation: An economic approach", *The Psychology of Economic Decisions*, *1*, 137–167.

R. Bénabou and J. Tirole (2011), "Identity, morals, and taboos: Beliefs as assets", *The Quarterly Journal of Economics*, *126*(2), 805–855.

Bertelsmann Stiftung (2014), *BTI 2014 - Tajikistan Country Report*, Gütersloh: Bertelsmann Stiftung.

R. Bodner and D. Prelec (2003), "Self-signaling and diagnostic utility in everyday decision making", *The Psychology of Economic Decisions*, *1*, 105–26.

S. Bowles and S. Polania-Reyes (2012), "Economic incentives and social preferences: substitutes or complements?", *Journal of Economic Literature*, 368–425.

F. R. Campante, D. Chor, and Q.-A. Do (2009), "Instability and the Incentives for Corruption", *Economics & Politics*, *21*(1), 42–92.

J. Carpenter and C. K. Myers (2010), "Why volunteer? Evidence on the role of altruism, image, and incentives", *Journal of Public Economics*, *94*(11), 911–920.

A. Cohn, E. Fehr, and M. A. Maréchal (2014), "Business culture and dishonesty in the banking industry", *Nature*, *516*, 86–89.

A. Colman (2011), "Court decisions in commercial matters: an EBRD assessment", *Business Law International*, *12*(2), 155–184.

D. Dickinson and M.-C. Villeval (2008), "Does monitoring decrease work effort?: The complementarity between agency and crowding-out theories", *Games and Economic Behavior*, *63*(1), 56–76.

V. Dimitrova-Grajzl, P. Grajzl, K. Zajc, and J. Sustersic (2012), "Judicial incentives and performance at lower courts: evidence from Slovenian judge-level data", *Review of Law & Economics*, *8*(1), 215–252.

V. P. Dimitrova-Grajzl, P. Grajzl, A. Slavov, and K. Zajc (2015), "Courts in a Transition Economy: Case Disposition and the Quantity-Quality Tradeoff in Bulgaria", Available at SSRN 2569351.

A. L. Duckworth, C. Peterson, M. D. Matthews, and D. R. Kelly (2007), "Grit: perseverance and passion for long-term goals", *Journal of Personality and Social Psychology*, *92*(6), 1087–1101.

European Bank for Reconstruction and Development (2012), *Commercial Laws of Tajikistan, April 2012, An Assessment by the EBRD*.

B. S. Frey (1993), "Does monitoring increase work effort? The rivalry with trust and loyalty", *Economic Inquiry*, *31*(4), 663–670.

B. S. Frey and R. Jegen (2000), "Motivation crowding theory: A survey of empirical evidence".

B. S. Frey and F. Oberholzer-Gee (1997), "The cost of price incentives: An empirical analysis of motivation crowding-out", *The American Economic Review*, 746–755.

H. Gleitman (1996), *Basic Psychology*, New York: Norton.

U. Gneezy, S. Meier, and P. Rey-Biel (2011), "When and why incentives (don't) work to modify behavior", *The Journal of Economic Perspectives*, 191–209.

U. Gneezy and A. Rustichini (2000), "A fine is a price", *The Journal of Legal Studies*, *29*(1), 1–17.

K. Hackenbrack and M. W. Nelson (1996), "Auditors' Incentives and Their Application of Financial Accounting Standards", *The Accounting Review*, *71*(1), 43–59.

K. J. Haley and D. M. Fessler (2005), "Nobody's watching?: Subtle cues affect generosity in an anonymous economic game", *Evolution and Human Behavior*, *26*(3), 245–256.

J. Heyman and D. Ariely (2004), "Effort for payment a tale of two markets", *Psychological Science*, *15*(11), 787–793.

K. Hoff and P. Pandey (2014), "Making up people - the effect of identity on performance in a modernizing society", *Journal of Development Economics*, *106*, 118 – 131.

E. Hoffman, K. McCabe, and V. L. Smith (1996), "Social distance and other-regarding behavior in dictator games", *The American Economic Review*, *86*(3), 653–660.

B. Holmstrom and P. Milgrom (1991), "Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design", *Journal of Law, Economics, & Organization*, 24–52.

C. A. Holt and S. K. Laury (2002), "Risk Aversion and Incentive Effects", *The American Economic Review*, *92*(5), 1644–1655.

K. Izuma (2012), "The social neuroscience of reputation", *Neuroscience Research*, *72*(4), 283–288.

K. Izuma, D. N. Saito, and N. Sadato (2010), "Processing of the incentive for social approval in the ventral striatum during charitable donation", *Journal of Cognitive Neuroscience*, *22*(4), 621–631.

A. Kimura, N. Mukawa, M. Yamamoto, Y. M. Masuda, Tomohiro and, S. Goto, T. Oka, and Y. Wada (2012), "The influence of reputational concerns on purchase intention of fair-trade foods among young japanese adults", *Food Quality and Preference*, *26*(2), 204–210.

D. S. Nagin, J. B. Rebitzer, S. Sanders, and L. J. Taylor (2002), "Monitoring, motivation, and management: The determinants of opportunistic behavior in a field experiment", *The American Economic Review*, *92*(4), pp. 850–873.

C. Panagopoulos (2014), "Watchful eyes: implicit observability cues and voting", *Evolution and Human Behavior*, *35*(4), 279–284.

R. A. Posner (1993), "What Do Judges and Justices Maximize? (The Same Thing Everybody Else Does)", *Supreme Court Economic Review*, *3*, 1–41.

D. Prelec (2011), "Decision analysis from a neo-calvinist point of view".

M. Rigdon, K. Ishii, M. Watabe, and S. Kitayama (2009), "Minimal social cues in the dictator game", *Journal of Economic Psychology*, *30*(3), 358–367.

R. M. Ryan and E. L. Deci (2000), "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being", *American Psychologist*, *55*(1), 68.

S. Salterio and L. Koonce (1997), "The persuasiveness of audit evidence: The case of accounting policy decisions", *Accounting, Organizations and Society*, *22*(6), 573–587.

J. C. Turner and R. S. Onorato (1999), "Social identity, personality, and the self-concept: A self-categorization perspective", *The Psychology of the Social Sself*, 11–46.

Table 1: Participant characteristics, by treatment

	N	Age	Sex	Education	Experience
Control	25	38.32	0.21	5.08	10.65
		(9.77)	(0.41)	(0.75)	(5.39)
Visibility	23	37.96	0.18	5.09	13.35
		(9.32)	(0.39)	(0.94)	(7.62)
Bonus	15	38.20	0.29	5.20	12.40
		(10.19)	(0.46)	(0.40)	(7.11)
Visibility + bonus	40	40.98	0.13	5.44	12.69
		(10.77)	(0.34)	(1.01)	(9.34)
p-value		0.29	0.52	0.07	0.39

Note: Standard deviation is reported in parentheses. Four judges with 8 and 9 years of education in the combined treatment. Results do not change if we control for education or drop them from the sample.

Table 2: Two-by-two design

	No chance of visibility	Chance of visibility
No bonus	*control*, C	*visibility*, T1
Bonus	*bonus*, T2	*visibility*+*bonus*, T3

Figure 1: Average performance across treatments

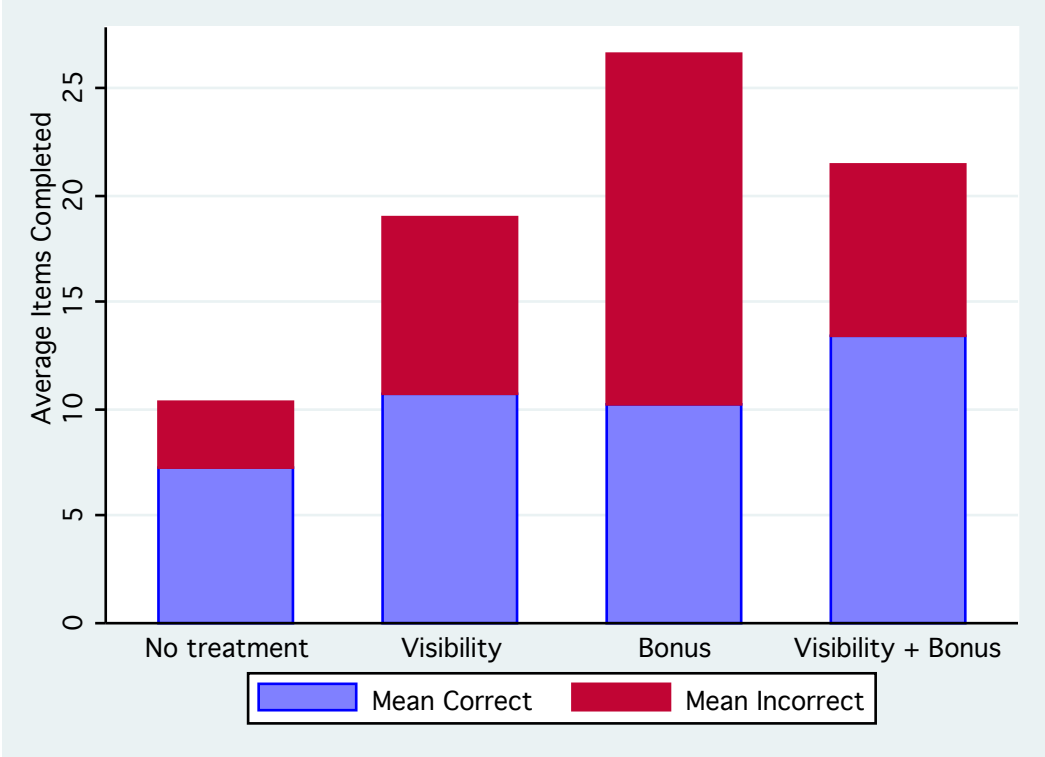

Table 3: Results summary, by treatment

	Mean items completed	Mean correct	Mean incorrect	% correct	% with C > we
Control	10.36	7.2	3.16	70.21%	72%
	(4.42)	(3.96)	(3.50)	(28.32)	(45.83)
Visibility	19	10.70	8.30	60.14%	65.21%
	(11.15)	(5.76)	(10.04)	(28.47)	(48.70)
Bonus	26.6	10.2	16.4	42.54%	47%
	(9.96)	(6.42)	(12.66)	(26.06)	(51.64)
Visibility + bonus	21.4	13.35	8.05	62.12%	73%
	(11.25)	(10.25)	(8.42)	(27.96)	(45.22)

Note: C, Standard deviation is reported in parentheses.

Table 4: Mann-Whitney results:
+ means column is more often higher than row,
- means that row is more often higher than column

		Visibility	Bonus	Visibility + bonus
Total items completed	Control	+***	+***	+***
	Visibility		+***	+
	Bonus			-*
Correct items	Control	+**	+	+***
	Visibility		-	+
	Bonus			+
Incorrect items	Control	+***	+***	+***
	Visibility		+***	-
	Bonus			-***
Per cent correct	Control	-	-***	-
	Visibility		-**	-
	Bonus			+***

Note: *** (**,*) indicates significance at the 1% (5%, 10%) level.

Table 5: Regression results, ordinary least squares

	(1) Total completed	(2) Incorrect	(3) Correct	(4) Total completed	(5) Incorrect	(6) Correct
Visibility	8.640***	5.144**	3.496**	8.583***	5.050**	3.533**
	(2.483)	(2.202)	(1.435)	(2.747)	(2.370)	(1.437)
Bonus	16.240***	13.240***	3.000	16.400***	14.074***	2.326
	(2.685)	(3.296)	(1.817)	(3.081)	(3.660)	(2.174)
Visibility + bonus	11.040***	4.890***	6.150***	10.549***	3.376**	7.173***
	(1.998)	(1.512)	(1.814)	(2.673)	(1.696)	(2.384)
Age				0.140	0.377**	-0.237*
				(0.174)	(0.152)	(0.126)
Female				-3.661	-7.809***	4.148
				(3.940)	(2.062)	(3.186)
Education				1.122	0.339	0.783
				(0.955)	(0.810)	(0.738)
Experience				-0.319	-0.469**	0.150
				(0.205)	(0.178)	(0.139)
Ambig. attitude				-4.328	-3.270	-1.058
				(3.220)	(2.759)	(2.330)
Work ethic				-0.252	-0.154	-0.098
				(1.082)	(0.832)	(0.790)
Constant	10.360***	3.160***	7.200***	7.276	-3.271	10.547
	(0.884)	(0.699)	(0.791)	(8.630)	(5.790)	(7.322)
R-square	0.237	0.179	0.093	0.298	0.322	0.189
N	103	103	103	92	92	92

Note: Robust standard errors are in parentheses. *** (**,*) indicates significance at the 1% (5%, 10%) level.

Table 6: Quantity-quality trade-off, per cent correct as the dependent variable

	(1)	(2)
Visibility	-0.101	-0.062
	(0.082)	(0.084)
Bonus	-0.324***	-0.368***
	(0.098)	(0.106)
Visibility + bonus	-0.074	-0.035
	(0.072)	(0.075)
Day of session	0.047	0.063
	(0.045)	(0.043)
Age		-0.012***
		(0.004)
Female		0.267***
		(0.066)
Education		0.037
		(0.032)
Experience		0.013***
		(0.004)
Ambig. attitude		0.097
		(0.083)
Work ethic		0.011
		(0.031)
Constant	0.608***	0.266
	(0.106)	(0.284)
R-square	0.097	0.290
N	103	92

Note: Robust standard errors are in parentheses. *** (**,*) indicates significance at the 1% (5%, 10%) level.

Table 7: Quantity-quality tradeoff, Number Incorrect as the dependent variable

	(1)	(2)	(3)	(4)
Correct	-0.326***	-0.284**	-0.265**	-0.041
	(0.115)	(0.138)	(0.126)	(0.263)
Visibility	6.283***	6.052**	4.772	4.504
	(2.274)	(2.509)	(3.833)	(5.127)
Bonus	14.217***	14.733***	23.959***	26.008***
	(3.111)	(3.453)	(5.981)	(6.419)
Visibility + bonus	6.893***	5.410***	6.113**	5.472
	(1.860)	(1.871)	(3.040)	(3.482)
Correct x visibility			0.122	0.052
			(0.242)	(0.417)
Correct x bonus			-0.973**	-1.174**
			(0.399)	(0.506)
Correct x (visibility+bonus)			0.031	-0.137
			(0.177)	(0.313)
Constant	5.505***	-0.280	5.071***	-4.206
	(1.080)	(5.663)	(1.288)	(5.676)
Controls	No	Yes	No	Yes
R-square	0.246	0.365	0.305	0.427
N	103	92	103	92

Note: Robust standard errors are in parentheses. *** (**,*) indicates significance at the 1% (5%, 10%) level. The controls included in columns (2) and (4) comprise *age*, *gender*, *education*, *years of experience in the legal sector* as well as measures of ambiguity preferences and work ethic derived from the Grit questionnaire.