Cala, Petr; Havranek, Tomas; Irsova, Zuzana; Matousek, Jindrich; Novak, Jiri

Working Paper

Financial Incentives and Performance: A Meta-Analysis of Economics Evidence

Financial Incentives and Performance: A Meta-Analysis of Economics Evidence*

Petr Cala[a], Tomas Havranek[a,b], Zuzana Irsova[a,c], Jindrich Matousek[a], and Jiri Novak[a]

[a]Charles University, Prague

[b]Centre for Economic Policy Research, London

[c]Anglo-American University, Prague

November 2, 2022

Abstract

Standard economics models require that financial incentives improve performance, while leading theories in psychology allow for the opposite. Experimental results are mixed, and so far have not been corrected for publication bias and model uncertainty. We collect 1,568 economics estimates together with 46 factors capturing the context in which the estimates were obtained. We use novel nonlinear techniques to correct for publication bias and employ Bayesian model averaging to account for model uncertainty. The corrected estimates are zero or tiny across contexts of field experiments, including differences in performance measurement, task definition, reward size and framing, motivation beyond money, subject pool, and estimation technique. Laboratory experiments produce statistically significant estimates on average after correction for publication bias, but even there the effect is weak. Experimental economics evidence is inconsistent with standard economics models.

Keywords: Incentives, experiments, meta-analysis, model uncertainty, publication bias

JEL Codes: C90, D91, M52

*Corresponding author: Zuzana Irsova, zuzana.irsova@ies-prague.org. Data and code are available in an online appendix at meta-analysis.cz/incentives.

1 Introduction

At least since 1971 psychologists have been pointing out that financial incentives can harm performance by crowding out the enjoyment we would otherwise earn while working on a task (Deci, 1971). An enjoyable task morphs into one that we do for the money, which crowds out intrinsic motivation. If extrinsic motivation provided by the financial incentive is not strong enough, money rewards result in reduced performance. While not universally accepted, the motivation crowding theory is the default incentive model in psychology and related fields. A widely cited meta-analysis by Weibel *et al.* (2010) reports that financial incentives indeed hurt performance in the case of interesting tasks. While economists have long been aware of the psychology theory and evidence (Camerer & Hogarth, 1999; Gneezy & Rustichini, 2000; Frey & Jegen, 2001; Gneezy *et al.*, 2011; Esteves-Sorenson & Broce, 2022), few models used in economics allow for motivation crowding. The following statement prominently and recently expressed on the website of a leading management consultancy reflects the prior of many economists:

> *Generous and specific financial incentives can help drive and sustain a rapid performance improvement.* (McKinsey, 2022)

We show that empirical evidence in economics emphatically rejects the prior. Our contribution is threefold. First, we correct the literature for publication bias, which can exaggerate the underlying effect multiplicatively (Bruns & Ioannidis, 2016; Ioannidis *et al.*, 2017; Brodeur *et al.*, 2020; Neisser, 2021; Stanley *et al.*, 2022).[1] Second, we allow for model uncertainty (Eicher *et al.*, 2011; Amini & Parmeter, 2012; Feldkircher & Zeugner, 2012; Moral-Benito, 2015; Steel, 2020), which is important given the heterogeneity of the literature. Third, we focus on experimental economics. Existing meta-analyses have focused on psychology exclusively or to a large extent: the meta-analysis with the largest proportion of economics evidence (24%, only 11 studies) is Weibel *et al.* (2010). The economics literature is thus largely unexplored, although several researchers have pointed out the vast differences in priors and methodological approaches be-

[1] Depending on the definition used, the term publication bias may or may not include p-hacking. When the definition excludes p-hacking, publication bias concerns the decision to publish the paper, while p-hacking concerns researchers' effort to obtain publishable estimates, for example by including different controls or focusing on different subsamples. The distinction is useful in simulations and some empirical applications: for example, using a unique dataset of submitted manuscripts, Brodeur *et al.* (2022) are able to unpack publication bias from p-hacking and conclude that p-hacking is what distorts the literature. In our sample, p-hacking and publication bias (narrowly defined) are observationally equivalent. For brevity and consistency with much of the meta-analysis literature, we thus use the broader definition of publication bias, which also includes p-hacking.

Figure 1: No consensus in the literature

Notes: The vertical axis shows the median partial correlation coefficient corresponding to the estimated effect of financial incentives on performance reported in individual studies. The horizontal axis shows the median year of the data used in the studies.

tween economics and psychology experiments when it comes to the effect of money on behavior (Camerer & Hogarth, 1999; Hertwig & Ortmann, 2001; Esteves-Sorenson & Broce, 2022).

Figure 1 presents a bird's-eye view of the experimental economics literature measuring the effect of financial incentives on performance. The median estimates from each study, recomputed to partial correlations for comparability, range commonly between 0 and 0.2, though some studies report correlations of -0.3 or 0.5. The estimates do not seem to be converging to a consensus value. After correcting for publication bias and allowing for model uncertainty, we obtain an estimate of zero for almost all experimental contexts. The only exception are laboratory experiments (including, coincidentally, some of the most recent studies in Figure 1), which yield a mean partial correlation of 0.07 with the 95% confidence interval (0.01, 0.14). The point estimate is on the boundary between a negligible and small effect according to the Doucouliagos (2011) guidelines for interpreting partial correlations. Laboratory experiments on this question, however, are generally quite rare in economics (only 23% of the estimates). The dominance of field experiments in economics, together with our correction for publication bias, may explain why a recent meta-analysis in psychology (Kim *et al.*, 2022) finds instead a positive effect of financial incentives on performance: their meta-analysis relies on laboratory experiments (83% of the estimates), which govern the psychology literature.

Two streams of research are closely related to our analysis. The first one concerns modern meta-analyses in experimental economics: Imai *et al.* (2021) presents a meticulous meta-analysis of the present bias, Brown *et al.* (2022) conduct a meta-analysis of loss aversion, and Matousek *et al.* (2022) focus on the individual discount rate. These studies highlight the importance of publication bias in experimental economics, along with systematic differences in the results related to the characteristics of the experiment.[2] The second stream of the literature concerns related meta-analyses in psychology. Most of the psychological research does not focus primarily on performance but intrinsic motivation: whether or not financial incentives lead to motivation crowding. The corresponding meta-analyses include Wiersma (1992); Cameron & Pierce (1994); Deci *et al.* (1999); Cameron (2001); Cerasoli *et al.* (2014)—their results are not clear-cut, but the Deci *et al.* (1999) study providing evidence for crowding out is the one most frequently cited in the literature. Regarding the cumulative effect of financial incentives on performance, meta-analyses include Jenkins *et al.* (1998); Condly *et al.* (2003); Weibel *et al.* (2010); Garbers & Konradt (2014); Kim *et al.* (2022). Again the results are mixed but, as we have noted, the Weibel *et al.* (2010) study (finding a negative effect on performance) is the one most prominently cited. None of the psychology meta-analyses correct the literature for publication bias.

Publication bias arises when some results, typically those that are intuitive and statistically significant, are preferentially selected for publication. Such selective reporting can work at the level of entire studies: for example, studies may end up unpublished, forever hidden in a file drawer, because of their insignificant results. More plausibly, however, selective reporting works as a form of voluntary self-censorship practiced by the authors themselves (Brodeur *et al.*, 2022). In the context of the incentive-performance literature, researchers can, for example, alter the measure of performance they report (Esteves-Sorenson & Broce, 2022) or choose a subset of the data until they get a desired outcome. Selective reporting does not equal cheating and can be completely unintentional. McCloskey & Ziliak (2019) draw a useful analogy between selective reporting in empirical research and the Lombard effect in psychoacoustics: speakers involuntarily increase their vocal effort in the presence of noise. In a similar way, researchers may increase their effort to find a plausible estimate when there is noise in the data. Consequently,

[2]Other recently published studies on meta-analysis and publication bias in economics more generally include Brodeur *et al.* (2016); Card *et al.* (2018); Christensen & Miguel (2018); Andrews & Kasy (2019); DellaVigna *et al.* (2019); Blanco-Perez & Brodeur (2020); Ugur *et al.* (2020); Xue *et al.* (2020); Stanley *et al.* (2021); DellaVigna & Linos (2022); Elliott *et al.* (2022); Iwasaki (2022).

publication bias is consistent with a correlation between reported estimates and their standard errors. In other words, studies with a large standard error will need a large point estimate to overcome noise and produce a statistically significant result.

Our initial identification assumptions are based on the Lombard effect: i) there is no correlation between estimates and standard errors in the absence of publication bias, and ii) publication bias is a linear function of the standard error. (We will relax both assumptions later.) Then a regression of estimates on their standard errors identifies both the extent of publication bias (the slope) and the mean estimate corrected for publication bias (the intercept). This "meta-regression", with appropriate weights and controls and study-level clustering, yields a robustly positive estimated slope and an estimated intercept in the vicinity of zero. The result is consistent with publication bias in favor of positive reported effects of financial incentives on performance: a plausible prior of many researchers in economics. (It is telling that only 57% of the studies in our sample mention the motivation crowding theory.) The result also implies that, according to the experimental economics literature, the underlying mean effect of financial incentives on performance is negligible.

The two assumptions mentioned above are commonly used in economics meta-analyses, but they are too strong for many contexts. Stanley & Doucouliagos (2014) and Andrews & Kasy (2019) show that publication bias is most likely a nonlinear function of the standard error. For this reason we employ a battery of recently developed nonlinear tests (Ioannidis *et al.*, 2017; Andrews & Kasy, 2019; Bom & Rachinger, 2019; Furukawa, 2020), which all corroborate our previous findings regarding publication bias and the mean underlying effect. The uncorrelation assumption is substantially more difficult to tackle. Havranek *et al.* (2022) show that, in economics, estimates can in principle be related to standard errors even in the absence of publication bias. For example, some method choices can systematically affect both quantities.

A straightforward solution is to use the inverse of the square root of the number of degrees of freedom as an instrument for the standard error. Such an instrument is correlated with the standard error by the definition of the latter and can be expected to be less related to various method choices. Unfortunately in the case of the incentive-performance literature the instrument is weak and the results uninformative. We thus use the p-uniform* technique by van Aert & van Assen (2021), which does not need the uncorrelation or the linearity assumption.

The technique uses the statistical principle that the distribution of p-values should be uniform at the underlying mean effect size, and once again we obtain a negligible mean effect after correction for publication bias. In addition, we use the tests of Gerber & Malhotra (2008) and Elliott *et al.* (2022), which also do not rely on the uncorrelation assumption and both corroborate the presence of publication bias.

The economics experiments measuring the effect of financial incentives on performance vary so much that a reader will ask how a mean estimate is informative regarding the field as a whole. Individual researchers focus on very different definitions of performance: school grades, blood donations, games, work outcomes, and others. The task itself can be appealing or unappealing, cognitive or manual. Outputs can be measured quantitatively or qualitatively. Reward size and framing differ across experiments, sometimes only individual people are paid, sometimes the rewards are group-specific. Some experiments are conducted in a lab, many are field studies. Subjects differ in terms of gender, occupation, age, and culture. Various econometric techniques are used to produce the main results. To allow for these many differences in the context in which the reported estimates were obtained, we employ Bayesian model averaging, which is the natural solution to model uncertainty in the Bayesian framework (Steel, 2020). To address collinearity in such an exercise we use the dilution prior (George, 2010). As a robustness check we use frequentist model averaging with Mallow's weights (Hansen, 2007) and orthogonalization of model space following the approach of Amini & Parmeter (2012).

The results of model averaging suggest that some method choices drive the results systematically. The composition of the subject pool matters, as does the framing of rewards, individual versus group rewards, and qualitative versus quantitative measurement of output. Financial incentives are even less efficient in improving grades and prosocial behavior than they are in improving performance at games and work. But these differences are surprisingly small. The implied correlations for various experimental contexts after correction for publication bias and accounting for model uncertainty are always statistically insignificant and negligible according to the classification of Doucouliagos (2011). The only exception, as we have mentioned, are laboratory experiments. But even here the implied effect is tiny. We conclude that, regarding the effect of financial incentives on performance, the experimental economics literature is inconsistent with most models commonly employed in economics.

Our results do not fit neatly in the mainstream psychology framework either. The motivation crowding theory assumes that the crowding out of intrinsic motivation happens only in the case of interesting tasks, exactly as reported by Weibel *et al.* (2010). When the task is fundamentally unappealing, intrinsic motivation is negligible, and there should be no crowding out. The problem is that the definition of an interesting task in most individual studies is subjective, and some people will enjoy tasks that other find unappealing. It is thus possible that intrinsic motivation exists even for tasks classified as uninteresting. Another potential explanation is that reward cues distract people from the task itself: a recent meta-analysis in psychology shows that this effect can be important (Rusz *et al.*, 2020): people focus on the rewards instead of the work. The distraction effect can be present for both interesting and uninteresting tasks and is more likely in field settings, where the experimenters do not always have full control over the connection between reward cues and the task itself. The distraction effect can thus be associated with our finding that lab experiments tend to yield more evidence for the effect of financial incentives on performance. Finally, it is possible that the experiments suffer from measurement error, which results in attenuation. Esteves-Sorenson & Broce (2022) survey 82 papers on related questions and highlight the variance in the different and sometimes inconsistent metrics used in these studies.[3]

2 Data

To build our dataset, first we search for studies that provide experimental evidence regarding the effect of financial incentives on performance. We use Google Scholar because of its powerful fulltext search; the details of our strategy, including the specific search query, are presented in Figure A1 in the Appendix. As we have noted, we only focus on economics journals and also only consider studies written in English. In addition, to be included in the meta-analysis each study must report standard errors or any other statistics from which standard errors can be computed (typically t-statistics or p-values). Standard errors are needed as weights in many meta-analysis techniques and also as regressors in meta-regression models of publication bias. Each included study has to report the number of degrees of freedom available for the estimation of the incentive-performance effect; information on the degrees of freedom is needed in order to

[3]These experiments also tend to have low power. Esteves-Sorenson & Broce (2022) report that across the 82 papers they review, the median number of subjects per per experimental condition was merely 15.

recompute the effects into a common standardized metric. Table A1 in the Appendix lists the 44 studies that fulfill all the selection criteria—we will call them primary studies.

An inclusion of unpublished working papers generally does not help alleviate publication bias. We exclude working papers because they are not peer-reviewed, are more prone to contain typos and other mistakes, and their classification into economics or psychology is sometimes unclear. The classification is clearer for journals, where we consider all journals listed in RePEc as primarily economics outlets, and all of the journals in which the included studies were published are also listed in the economics category in the Web of Science. As we have noted in the Introduction, our definition of publication bias is broad and also includes p-hacking and self-censoring on the side of the authors themselves. Indeed, Brodeur *et al.* (2022) show that editorial decisions are more likely to alleviate than strengthen publication bias. In a similar vein, Rusnak *et al.* (2013) show that the extent of publication bias among published studies does not exceed that among unpublished studies.

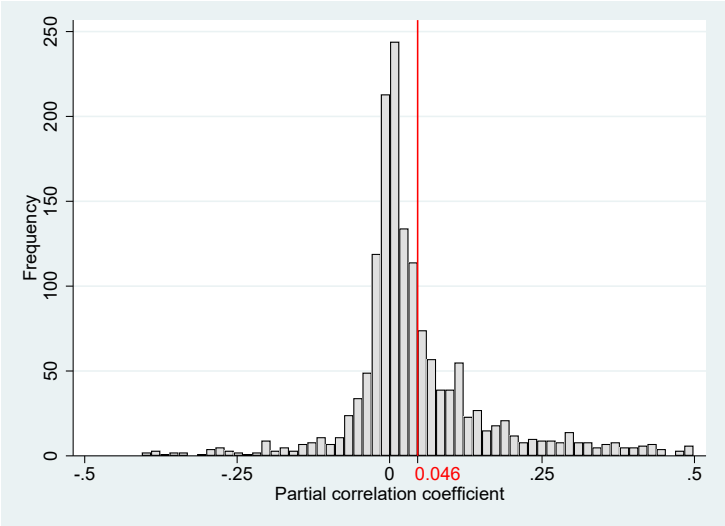
The basic statistics that we collect from the primary studies are the point estimate of the incentive-performance effect, the corresponding standard error, and the number of degrees of freedom used in the estimation. Because the measures of performance used in primary studies vary widely, the point estimates cannot be compared directly. We thus recompute them to a comparable metric, partial correlation coefficients (PCCs), according to the following formula:

$$PCC = \frac{t}{\sqrt{t^2 + df}}, \tag{1}$$

where t stands for the t-statistic of the reported coefficient and df indicates the number of degrees of freedom in the estimation. Using the computed partial correlation and the original t-statistic we then obtain the corresponding standard error of the standardized measure (Stanley & Doucouliagos, 2012).

Most primary studies report many different estimates of the incentive-performance effect. Typically these different estimates reflect different subsets of the subject pool, but sometimes there are also within-study differences in reward size, framing, estimation technique, and other aspects. We collect all estimates for which standard errors and degrees of freedom are reported. In total, we thus obtain 1,568 estimates, a large and rich dataset. We winsorize the effects at the 1% level to limit the influence of outliers. To account for the context in which the

Figure 2: Estimates around zero are most common in the literature



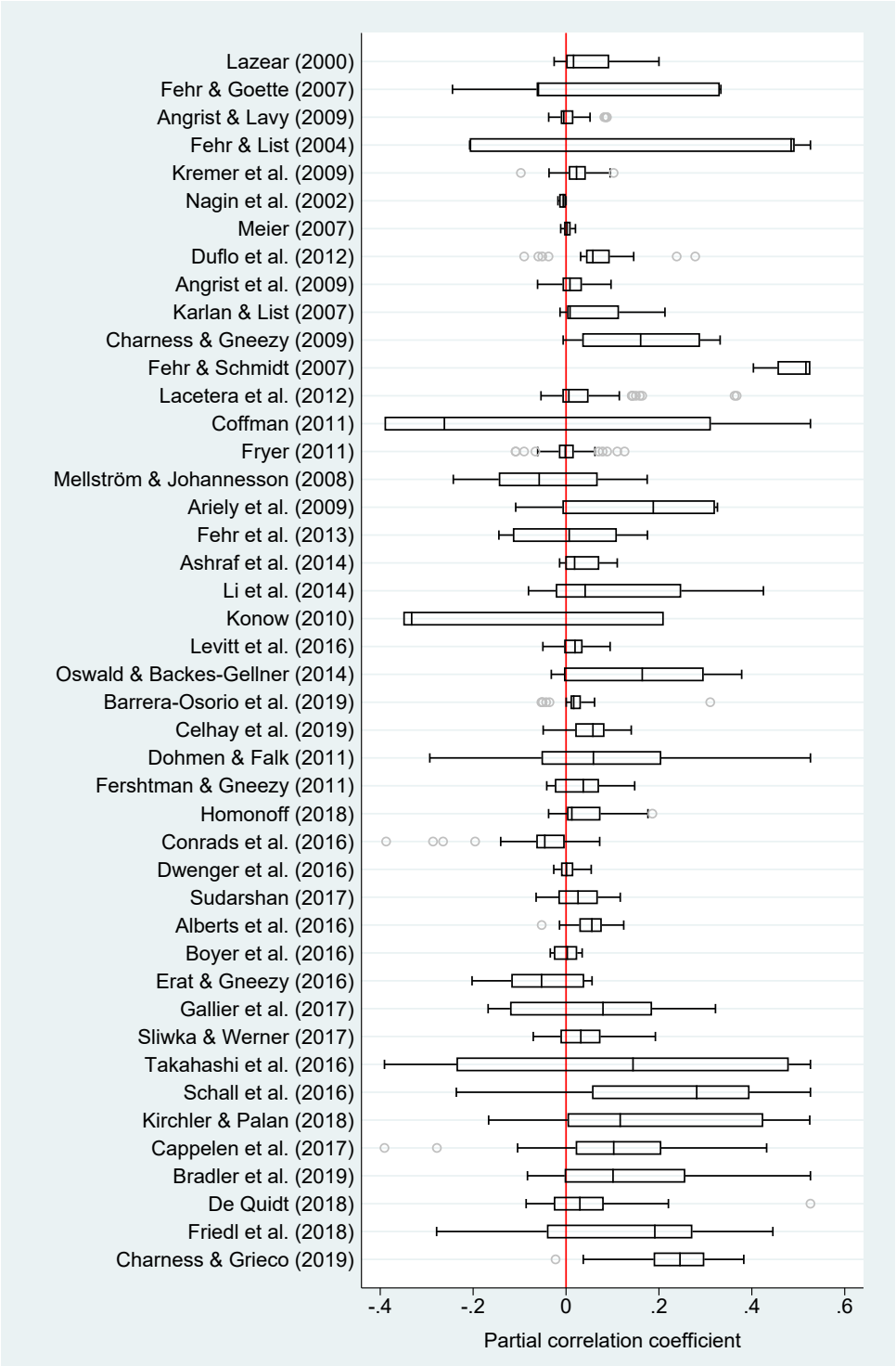
Notes: The figure depicts a histogram of the partial correlation coefficients corresponding to the estimated effects of financial incentives on performance reported in individual studies. The vertical line denotes the sample mean. Outliers are excluded from the figure for ease of exposition but included in all statistical tests.

estimates were obtained we collect 46 aspects of the data, experimental approach, and resulting publication. This means that we had to fill more than 70,000 data points by hand after reading the primary studies carefully. Three of the co-authors collected 1/3 of the data each; another co-author randomly checked 1/3 of the entire dataset. The discovered inconsistencies in coding were discussed among the co-authors and corrected for the entire dataset. The final dataset, together with the R code used in the meta-analysis, is available in an online appendix at `meta-analysis.cz/incentives`.

Figure 2 shows the distribution of the estimates in our dataset. Estimates close to zero are common, and the mean partial correlation is 0.046: a negligible effect according to the Doucouliagos (2011) guidelines for interpreting partial correlations.[4] The right-hand portion of the distribution is heavier than the left-hand portion, which might indicate publication bias in favor of positive estimates—but it may also simply indicate heterogeneity in the underlying effects. Few estimates exceed 0.33, a threshold denoting large estimates in the guidelines. Figure 3 shows the box plot of the estimates reported in individual studies. (Figure B1 in the Appendix shows the box plot for individual countries, where we observe no systematic

[4]Doucouliagos (2011) uses a large sample of economics meta-analyses to map partial correlations to elasticities. In his mapping, correlations below 0.07 are typically consistent with negligible elasticities even if statistically significant, correlations below 0.17 denote a small effect, and correlations above 0.33 denote a large effect.

Figure 3: Most studies report both positive and negative estimates



Notes: The figure depicts a box plot of the partial correlation coefficients corresponding to the estimated effects of financial incentives on performance reported in individual studies. The studies are sorted by the age of the data they use from oldest to youngest. The length of each box represents the interquartile range (P25-P75), and the dividing line inside the box is the median value. The whiskers represent the highest and lowest data points within 1.5 times the range between the upper and lower quartiles. The vertical line denotes zero effect. Outliers are excluded from the figure for ease of exposition but included in all statistical tests.

differences.) The studies are sorted by the age of the data they use: given the long and variable publication lags in economics, the year of data is more informative than the year of publication. Three stylized facts emerge from the figure. First, most studies report both positive and negative estimates of the incentive-performance effect. Second, the mean reported effect tends to be quite close to zero for most of the older studies. Third, the mean reported effects seem to be positive and non-negligible for about 10 of the most recent studies. These studies are typically conducted in a lab and measure performance in games (8 out of the 10 papers).

Table 1 shows summary statistics for selected subsets of the data. The first part of the table presents unweighted statistics, in which each estimate has the same weight. The second part shows statistics weighted by the inverse of the number of estimates reported per study, which means that here each study has the same weight. The main takeaway from the table is that estimates of the incentive-performance effect are small irrespective of context; different weights do not change the conclusion, and any systematic differences seem to be small. Figure 4 documents the lack of large systematic differences visually. The definition of the categories used in Table 1 and Figure 4 is available in Table 4 in the section on heterogeneity. Here we just briefly discuss the main differences in estimation contexts. A key difference is the definition of performance: only a small majority of studies focus on work outcomes, and the literature is dominated by performance measured in school grades, games, and prosocial behavior (for example, blood donations). While the mean effect is small for all the categories, it is smaller for grades and prosocial behavior than for game and work outcomes. Regarding the nature of the task, the effect seems to be larger for appealing than for unappealing tasks, but quite similar for cognitive and manual tasks and outcomes measured qualitatively and quantitatively.

Concerning the reward scheme, large rewards do not increase performance compared to small rewards. It does not seem to matter whether the framing of the experiment is positive (gain) or negative (loss), whether subjects get a show-up fee, and whether rewards are paid to individuals or to groups. The primary studies also differ in terms of the underlying motivation they provide beyond money. Some of the tasks are meaningless beyond the financial incentive (for example, counting dots on a screen), while other tasks involve aspects of altruism, reciprocity, and fairness. The mean effect is similar to the overall mean when money is the sole motivation (0.037 vs. 0.046). The effect seems to be larger for reciprocity, but here we only have 161 observations. Concerning the general design of the experiment, lab studies tend to report larger estimates than

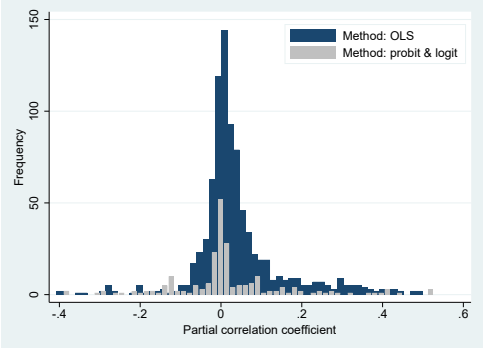
Table 1: Subsets of the literature do not differ much

		Unweighted			Weighted		
	No. of observations	Mean	95% conf. int.		Mean	95% conf. int.	
All estimates	1,568	0.046	0.040	0.053	0.063	0.055	0.072
Definition of performance effect							
Effect: grades	540	0.029	0.023	0.035	0.047	0.039	0.055
Effect: charity	444	0.035	0.028	0.042	0.032	0.023	0.041
Effect: game	437	0.073	0.053	0.092	0.084	0.061	0.107
Effect: work	147	0.067	0.039	0.095	0.081	0.054	0.109
Nature of the task							
Task: appealing	755	0.069	0.057	0.082	0.084	0.068	0.099
Task: unappealing	813	0.025	0.020	0.030	0.033	0.026	0.040
Task: cognitive	1,106	0.049	0.041	0.057	0.067	0.055	0.079
Task: manual	355	0.052	0.038	0.066	0.062	0.046	0.077
Performance: quantitative	1,101	0.043	0.034	0.052	0.061	0.050	0.072
Performance: qualitative	467	0.054	0.044	0.065	0.071	0.058	0.084
Reward scheme							
Reward size ≥ 0.5	863	0.038	0.032	0.045	0.057	0.045	0.065
Reward size < 0.5	705	0.056	0.043	0.069	0.069	0.053	0.084
Positive framing	1,303	0.048	0.041	0.056	0.069	0.058	0.079
Negative framing	237	0.040	0.032	0.049	0.039	0.030	0.048
All subjects paid	1,162	0.054	0.045	0.063	0.070	0.058	0.081
Individual reward	1,268	0.045	0.037	0.052	0.065	0.054	0.075
Group reward	300	0.054	0.043	0.065	0.057	0.041	0.073
Motivation beyond money							
Motivation: altruism	456	0.046	0.037	0.056	0.037	0.023	0.051
Motivation: reciprocity	161	0.102	0.078	0.126	0.090	0.065	0.115
Motivation: fairness	237	0.019	-0.003	0.042	0.021	-0.003	0.045
Motivation: money only	690	0.037	0.028	0.046	0.061	0.049	0.073
Study design							
Laboratory experiment	366	0.091	0.073	0.110	0.100	0.077	0.124
Field experiment	1,202	0.033	0.026	0.039	0.044	0.036	0.052
Crowding-out theory	765	0.051	0.041	0.060	0.056	0.044	0.068
Structural variation							
Subjects: students	957	0.038	0.029	0.047	0.042	0.031	0.053
Subjects: employees	113	0.065	0.039	0.091	0.064	0.037	0.091
Subjects: general	498	0.058	0.048	0.069	0.108	0.091	0.125
More than 50% males	440	0.055	0.040	0.069	0.060	0.043	0.076
Gender equity	780	0.041	0.032	0.050	0.057	0.045	0.068
Less than 50% males	348	0.049	0.035	0.063	0.084	0.061	0.107
Developed country	1,305	0.045	0.037	0.053	0.065	0.055	0.075
Developing country	253	0.055	0.042	0.069	0.060	0.045	0.076
Estimation technique							
Method: OLS	895	0.042	0.034	0.051	0.060	0.049	0.072
Method: logit	75	-0.007	-0.022	0.008	-0.021	-0.042	0.000
Method: probit	141	0.034	0.006	0.062	0.020	-0.009	0.048
Method: tobit	48	0.144	0.065	0.223	0.157	0.079	0.235
Method: fixed effects	61	0.026	0.007	0.046	0.026	0.007	0.046
Method: random effects	44	0.118	0.061	0.176	0.162	0.098	0.227
Method: DID	43	0.045	0.024	0.065	0.045	0.024	0.065
Method: other	261	0.057	0.046	0.069	0.065	0.049	0.081

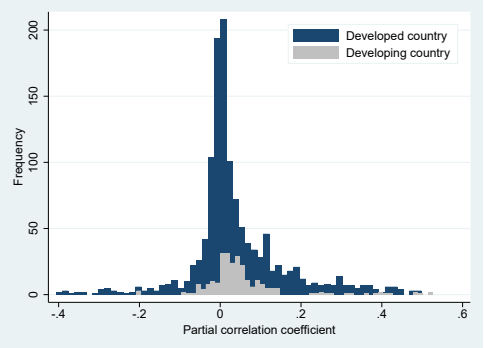
Notes: The table summarizes partial correlation coefficients corresponding to the estimated effects of financial incentives on performance reported in individual studies. The definition of the variables is available in Table 4. We use the IMF definition to classify countries as developed or developing. Weighted = estimates are weighted by the inverse of the number of estimates reported per study so that each study has the same weight in the resulting mean. OLS = ordinary least squares, DID = difference-in-differences.

Figure 4: Few prima facie patterns in the data

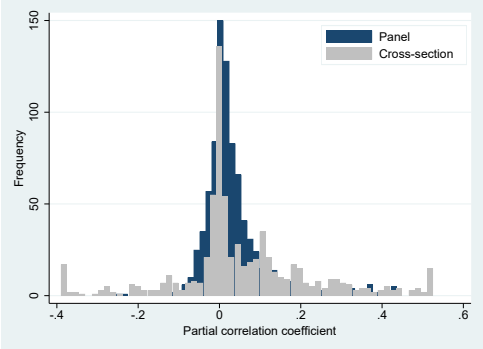
(a) Methods used



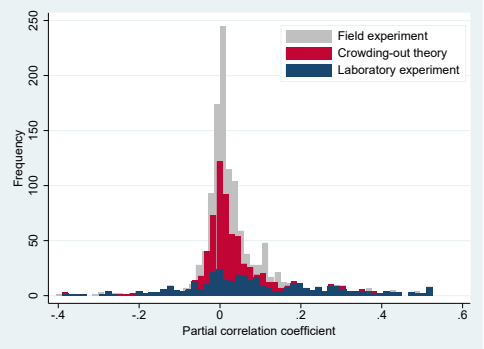
(b) Country classification



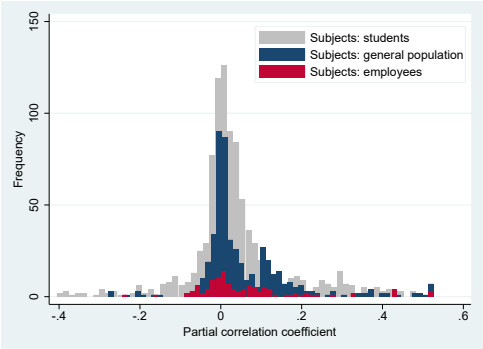
(c) Data dimension



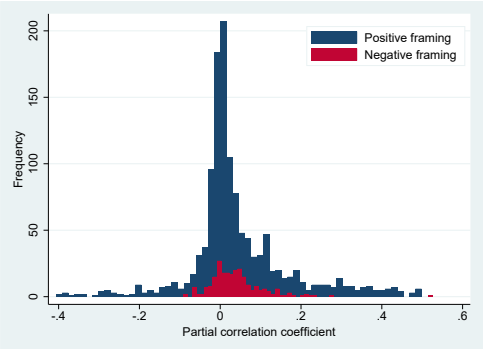
(d) Study design



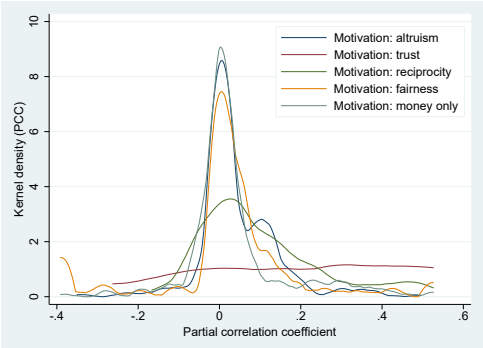
(e) Subject selection



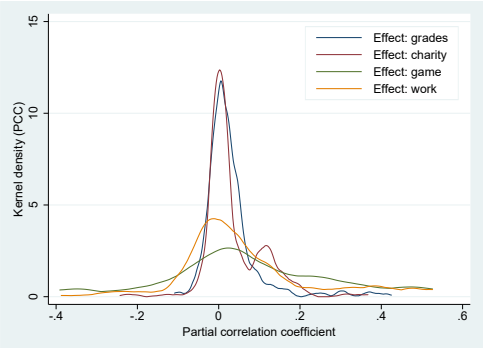
(f) Task framing



(g) Motivation beyond money



(h) Definition of performance



Notes: The figure depicts histograms of partial correlation coefficients corresponding to the estimated effects of financial incentives on performance reported in individual studies. The definition of the variables is available in Table 4.

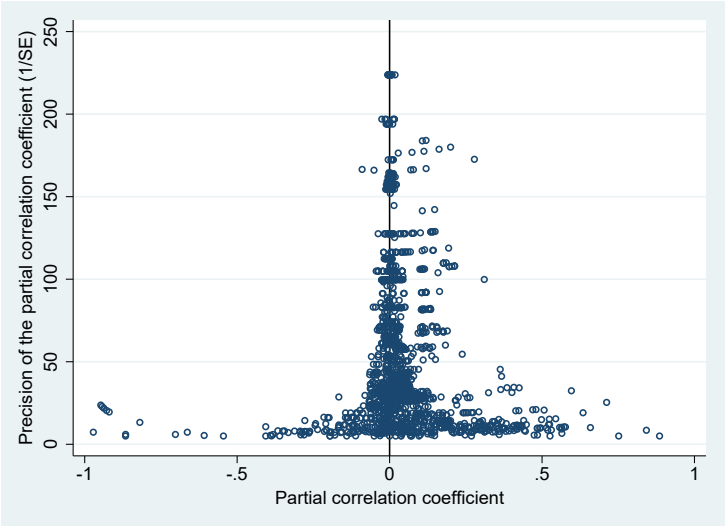
field studies. Studies that explicitly mention the motivation crowing theory report estimates similar to the overall mean (0.051 vs. 0.046). The composition of the subject pool also does not seem to matter much, and the same applies for the estimation technique—here some subsets display means above 0.1, but for these subsets we have very few observations. Table 1, however, ignores publication bias, which can distort the reported findings substantively (Ioannidis *et al.*, 2017).

3 Publication Bias

As Camerer & Hogarth (1999, p. 7) put it, "the predicted effect of financial incentives on human behavior is a sharp theoretical dividing line between economics and other social sciences, particularly psychology." Economists often take it for granted that people respond to financial incentives by working harder and producing more. It is perhaps a case in point that the motivation crowding theory, mentioned prominently in just about every psychology experiment we have seen on the topic, has been noted by only 25 of the 44 economics studies we collect for this meta-analysis. If researchers expect that positive, statistically significant results are natural, they can treat negative or insignificant results with suspicion. They may choose not to write papers based on such results, not to publish such papers, or to (intentionally or not) adjust their methodology or dataset in order to produce the intuitive outcome. The resulting distortion of the research record is called publication bias. As documented by the many references we provide in the Introduction, publication bias is widespread across economics and related disciplines. The bias is natural and inevitable, this is no crisis: it is the task for those who take stock of the literature to correct for the distortion.

A basic visual tool used for the detection of publication bias is the so-called funnel plot. It is a scatter plot of point estimates on the horizontal axis against the estimates' precision (the inverse of the standard error) on the vertical axis. In the absence of publication bias, small-sample effects, and systematic heterogeneity, the most precise estimates should be close to the mean underlying effect. With decreasing precision, estimates get more dispersed around the mean; consequently, the scatter plot will attain the shape of an inverted funnel. If some negative estimates are discarded (unpublished, unrecorded, or re-estimated), the funnel plot will no longer be symmetrical around the mean. The symmetry of the funnel plot thus serves

Figure 5: The funnel plot is consistent with modest publication bias



Notes: The figure shows partial correlation coefficients corresponding to the estimated effects of financial incentives on performance reported in individual studies. In the absence of publication bias (and systematic heterogeneity and potential small-sample biases) the funnel should be symmetrical around the most precise estimates.

as a basic test of publication bias. Figure 5 shows that, in the case of the incentive-performance literature, the scatter plot indeed resembles the theoretically predicted inverted funnel, and that the funnel is asymmetrical: the right-hand part is heavier, though the asymmetry is not particularly strong. We can also see from the funnel that the most precise estimates are close to zero, but that there is also substantial heterogeneity.

In Panel A of Table 2 we test the asymmetry of the funnel plot by regressing estimates on their standard errors (Egger *et al.*, 1997; Stanley, 2005). If publication bias is a linear function of the standard error and if there is no correlation between estimates and standard errors in the absence of publication bias, then the slope coefficient in the "meta-regression" identifies the degree of publication bias and the constant determines the mean incentive-performance effect corrected for the bias. The linearity assumption is motivated by the Lombard effect mentioned in the Introduction: with increasing noise (that is, the standard error) researchers increase their effort (to produce larger estimates) so that they obtain a statistically significant result. Because statistical significance, measured by the t-statistic, is given by the ratio of the estimates to its standard error, there is hope that selection effort will increase proportionally with the standard error in order to achieve the same t-statistic. The uncorrelation assumption is motivated by the fact that the ratio of estimates and standard errors is assumed to have

Table 2: Most techniques suggest significant publication bias, corrected effect around 0.02

Panel A: Linear techniques					
	OLS	FE	RE	Study	Precision
Publication bias	0.319**	0.879***	0.627***	0.203	0.879***
(Standard error)	(0.131)	(0.037)	(0.125)	(0.134)	(0.172)
Effect beyond bias	0.0320***	0.014***	0.020***	0.035***	0.014***
(Constant)	(0.004)	(0.001)	(0.003)	(0.004)	(0.003)
Observations	1,568	1,568	1,568	1,568	1,568
Panel B: Nonlinear techniques					
	WAAP	Top10	Stem	AK	Kink
Publication bias				P = 0.351	0.879***
				(0.030)	(0.153)
Effect beyond bias	0.024***	0.019***	0.021***	0.022*	0.013***
	(0.003)	(0.004)	(0.007)	(0.013)	(0.002)
Observations	1,568	1,568	1,568	1,568	1,568
Panel C: Endogeneity-robust techniques					
				IV	p-uniform*
Publication bias				0.194	L = 2.17
				(2.696)	(*p = 0.14*)
				{-5.1, 5.5}	
Effect beyond bias				0.037	0.021***
				(0.121)	(0.001)
First-stage robust *F-stat*				0.35	
Observations				1,568	1,568

Notes: Panel A: Results of regression $PCC_{is} = PCC_0 + \gamma SE(PCC_{is}) + \epsilon_{is}$, where PCC_{is} denotes the partial correlation coefficient of the i-th estimate from the s-th study and $SE(PCC_{is})$ denotes its standard error. The standard errors of the regression parameters are clustered at the study level and shown in parentheses. OLS = ordinary least squares, FE = study-level fixed effects, RE = study-level random effects, Study = weighted by the inverse of the number of estimates reported per study, Precision = weighted by the inverse of the estimate's standard error. Panel B: WAAP = weighted average of adequately powered estimates (Ioannidis *et al.*, 2017). Top10 = the method due to Stanley *et al.* (2010). Stem = the stem-based method due to Furukawa (2020). Kink model = the endogenous kink method due to Bom & Rachinger (2019). AK = the selection model due to Andrews & Kasy (2019) where P denotes the probability that estimates insignificant at the 5% level are published relative to the probability that significant estimates are published (the latter normalized at 1). Panel C: IV = the inverse of the square root of the number of observations is used as an instrument for the standard error. In curly brackets we show the two-step weak-instrument-robust 95% confidence interval based on Andrews (2018) and Sun (2018). P-uniform* = the method by van Aert & van Assen (2021) where L denotes test statistic of p-uniform*'s publication bias test, and the corresponding p-value is reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% level.

a symmetrical distribution, which means that estimates and standard errors are statistically independent quantities: a property implied by most empirical techniques. In economics practice, however, both assumptions can easily be violated, and we will address these violations later.

The first column in Panel A of Table 2 is a simple OLS regression with standard errors clustered at the study level. In the second column we add study-level fixed effects to account for unobserved study-level heterogeneity. In the third column we use random effects instead of fixed effects. Random effects are frequently used in meta-analysis, especially outside economics, but the exogeneity assumption can easily be violated because publication bias can differ systematically across studies, and thus the standard error in the regression can be correlated with the random effects term. In the fourth column we use weights equal to the inverse of the number of estimates reported per study; this way we give each study the same weight. In the last column we use classical meta-analysis weights based on inverse variance—here more precise estimates get more weight, and the specification explicitly addresses the heteroskedasticity inherent in regressing estimates on a measure of their variance. Four out of the five linear techniques find significant publication bias, and all agree that the mean corrected effect is around 0.02 (between 0.014 and 0.035), compared to the uncorrected mean of 0.046 discussed in the previous section.

The linearity assumption is unlikely to hold in general, as shown by Stanley & Doucouliagos (2014) and Andrews & Kasy (2019). In practice, thresholds for t-statistics (such as 1.96) are important for researchers. If the standard error increases but the t-statistic is safely above 1.96, the researcher has no incentive for more intensive specification search, and therefore here the connection between publication bias and the standard error disappears. The linearity assumption can be expected to hold only in the immediate vicinity of 1.96 or other important thresholds. In Panel B of Table 2 we use 5 techniques that allow for a generally nonlinear relationship between the standard error and publication bias. (In other words, these techniques are not based on a linear regression.) The first technique we use is the weighted average of adequately powered estimates (Ioannidis *et al.*, 2017). It is an inverse-variance weighted average of all the estimates with power at least 80%, and Stanley *et al.* (2017) show that the estimator works well in simulations. The second technique, "top10", is a simple average of the 10% of the most precise estimates (Stanley *et al.*, 2010). The third technique, the stem-based method due to Furukawa (2020), extends the previous one by endogenously determining what

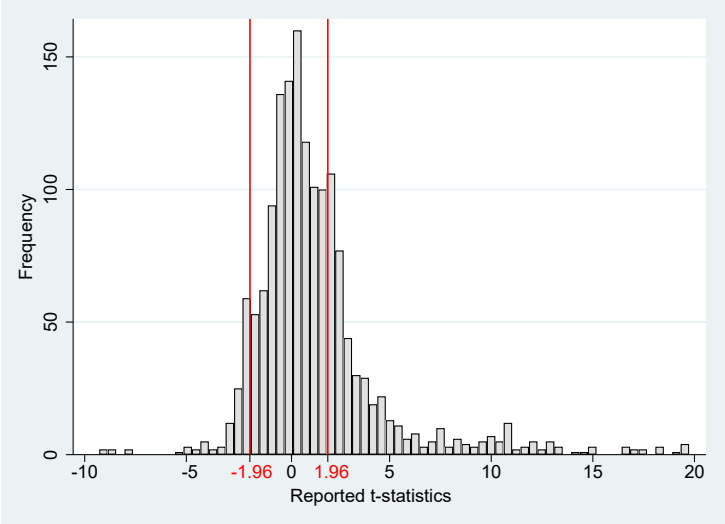
proportion of the most precise estimates to use. The proportion is determined by exploiting the trade-off between bias and variance: it is inefficient to discard estimates (variance increases), but imprecise estimates are more likely to be selectively reported (publication bias increases). The technique minimizes the sum of bias and variance.

The fourth technique in Panel B of Table 2 is the selected model by Andrews & Kasy (2019). This technique has arguably the most rigorous foundations, and has been shown to perform relatively well both in simulations (Hong & Reed, 2021) and in comparisons of meta-analyses and pre-registered replications (Kvarven *et al.*, 2019). The technique assumes that publication probability is constant for estimates with the same degree of statistical significance: for example, those with two stars for significance at the 5% level. The probability of publication changes when an important t-statistic threshold is crossed. Andrews & Kasy (2019) estimate the probability that each estimate is published and then re-weight the estimates by the inverse of the probability in order to recover the unbiased distribution of estimates. Finally, the fifth nonlinear model, the endogenous kink technique (Bom & Rachinger, 2019), is based on the linear meta-regression but adds a constant segment for highly statistically significant estimates, when it probably does not matter for publication bias if the standard error changes. Taken together, the nonlinear models provide a robust evidence that the corrected incentive-performance effect is around 0.02. The last two models, which also yield tests of publication bias, show strong bias. For example, the Andrews & Kasy (2019) model implies that positive estimates significant at the 5% level are almost three times more likely to be published than statistically insignificant estimates.

Nevertheless, all the models mentioned so far assume, explicitly or implicitly, that any correlation between estimates and standard errors is due to publication bias. Put more generally, the meta-regression in Panel A of Table 2 suffers from endogeneity. The endogeneity can have at least three sources. First, measurement error, because the standard error is itself an estimate. Second, reverse causality, because some researchers may, intentionally or not, manipulate the standard error in order to get statistically significant estimates (for example, by changes in clustering). Third, unobserved heterogeneity, because some method choices may systematically influence both estimates and standard errors. One solution to these problems is to use the inverse of the square root of the number of degrees of freedom as an instrument for the standard error

(Havranek *et al.*, 2022). The instrument is correlated with the standard error by definition, but does not suffer from the three sources of endogeneity described above. Unfortunately, in our case the instrument is weak, the first-stage robust F-statistic is only 0.35, and the weak-instrument-robust confidence interval (Andrews, 2018; Sun, 2018) is consequently wide. Thus we use the p-uniform* technique recently developed in psychology (van Aert & van Assen, 2021), which is a nonlinear model based on the statistical principle that p-values should be uniformly distributed at the mean underlying effect size. Once again we obtain a mean effect of 0.02, though the estimate for publication bias is marginally insignificant (p-value = 0.14).

Figure 6: The distribution of t-statistics peaks at 0, is consistent with selective reporting



Notes: The figure represents the distribution of t-statistics corresponding to the effect of financial incentives on performance reported in the literature. Vertical lines represents critical values associated with statistical significance at the 5% level.

Two other models of publication bias do not rely on the uncorrelation assumption, but they only test for the bias and do not yield an estimate of the corrected mean effect. Because the models use the reported t-statistics (or p-values), the results cannot be affected by the normalization to partial correlation coefficients that we choose to ensure compatibility in the case of all the previous techniques. The first additional technique is the so-called caliper test due to Gerber & Malhotra (2008); the second is the p-hacking tests due to Elliott *et al.* (2022): the test of non-increasingness and the test of monotonicity and bounds. The caliper test focuses on an important threshold of the t-statistic (typically 1.96, which denotes statistical significance

Table 3: Tests based on the distribution of t-statistics and p-values

Panel A: Caliper tests due to Gerber & Malhotra (2008)		
	Threshold 1.96	Threshold −1.96
Caliper width 0.05	0.370^{***}	-0.366^{***}
	(0.038)	(0.053)
n_1/n_2	29 / 4	2 / 16
Caliper width 0.1	0.352^{***}	-0.329^{***}
	(0.033)	(0.045)
n_1/n_2	41 / 7	5 / 23
Caliper width 0.2	0.303^{***}	-0.310^{***}
	(0.023)	(0.032)
n_1/n_2	79 / 20	10 / 44

Panel B: P-hacking tests due to Elliott *et al.* (2022)		
	Test for non-increasingness	Test for monotonicity and bounds
p-value	0.09	0.05
Observations ($p \leq 0.15$)	788	788
Total observations	1,568	1,568

Notes: Panel A shows the results of two sets of caliper tests around t-statistic thresholds of 1.96 and −1.96. Caliper width 0.05 means $t \in < 1.91; 2.01 >$ & $t \in < -2.01; -1.91 >$. A test statistic of 0.37, for example, means that 87% estimates are above the threshold and 13% estimates are below the threshold, far from the 50% expected in the absence of selective reporting. Standard errors, clustered at the study level, are included in parentheses. n_1/n_2 = number of observations above and below the threshold, respectively. Panel B reports tests developed by Elliott *et al.* (2022), which also feature cluster-robust variance estimators. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

at the 5% level) and compares the number of reported t-statistics just below and just above the threshold. In the absence of publication bias and with a sufficiently narrow caliper, there should be no difference. Figure 6 shows the distribution of reported t-statistics in the case of the incentive-performance literature. The threshold associated with 5% significance features jumps in the distribution for both positive and negative estimates, even though the jumps are smaller than the one associated with zero.

Panel A of Table 3 shows the results of the caliper test. In three calipers of different width around the 1.96 and −1.96 thresholds, t-statistics above the caliper (in absolute value) are much more common, which is consistent with selective reporting in favor of results that are just statistically significant at the 5% level; the result is also in line with the earlier findings of the Andrews & Kasy (2019) selection model. Figure 6 shows that the threshold of zero is even more important: estimates that are just positive are more likely to be reported than those that are

just negative. The jump is so clear in the figure that it is not necessary to report a formal caliper test, which corroborates the conclusion. Panel B shows the results of p-hacking tests due to Elliott *et al.* (2022). The main advantage of these rigorous tests is that they do not need us to specify a threshold of the t-statistic: they test publication bias using the general distribution of all p-values. As noted by Havranek *et al.* (2022), these tests need a lot of observations to work well, and in our case of a moderately large dataset they find evidence for publication bias at the 10% level. The finding of publication bias seems to be robust across different methods—but some of the evidence may be contaminated by the differences in the data and methods used to identify the incentive-performance effect.

4 Heterogeneity

So far we have not taken explicitly into account the fact that different estimates of the incentive-performance effect are obtained in different context. Several of the tests of publication selection bias allow for systematic heterogeneity: for example, the p-uniform* model, the instrumental variable meta-regression, and, as far as between-study heterogeneity is concerned, the fixed effects meta-regression. But none of these techniques allow for a full-fledged treatment of heterogeneity. That is what we provide in this section, and our goals are threefold. First, to see whether the finding of publication bias is robust to an explicit control for heterogeneity. Second, to find out which characteristics of study design systematically affect the reported results. Third, to estimate the effect of financial incentives on performance for different contexts after correction for publication bias and other potential biases.

We collect 46 variables that reflect the differences in data, estimation, and publication characteristics within and across primary studies. While the list of variables associated with heterogeneity is potentially unlimited, we believe that these 46 factors capture the differences most commonly discussed in the literature on the incentive-performance nexus. The variables are explained in detail in Table 4, and here we provide but a brief overview. The first group of variables concerns the definition of performance. The experiment can focus on school grades, charity (prosocial behavior such as blood donations or charitable givings), games, or work outcomes. The effect can be measured in terms of the time taken to finish the task or alternatively in terms of evaluating the outputs.

The experiments also differ in the way how appealing the task is, whether it is cognitive or manual, and whether performance is measured quantitatively or qualitatively. Researchers use incentives of various size, but because experiments are conducted in different countries, incentive size is not directly comparable. We thus divide the mean reward size in the experiment by the median expenditure in the corresponding country. The studies in our sample also differ in the framing they employ: typically the incentive is framed as a reward, but sometimes researchers explicitly punish participants for bad performance. In most cases all participants receive some money, such as a show-up fee, irrespective of their performance. But some studies intentionally do not offer show-up fees in order to increase the likelihood that participants apply because they like the experimental task (such as tasting cookies, Esteves-Sorenson & Broce, 2022), and that they consequently self-select for a task that can be classified as appealing for the participants. The rewards themselves are typically individual, but we also include a few studies that consider rewards for group performance.

Table 4: Description and summary statistics of regression variables

Variable	Description	Mean	SD
PCC	The partial correlation coefficient corresponding to the effect of financial incentives on performance reported in individual studies.	0.046	0.136
Standard error	The standard error of the partial correlation coefficient.	0.045	0.044
Definition of performance effect			
Effect: grades	= 1 if the estimated effect captures study performance (typically grade point average).	0.344	0.475
Effect: charity	= 1 if the estimated effect captures prosocial behavior (e.g., charitable givings, blood donations).	0.283	0.451
Effect: game	= 1 if the estimated effect captures the outcome of a game.	0.279	0.449
Effect: work	= 1 if the estimated effect captures employees' performance at work (reference category).	0.094	0.292
Effect: positive	= 1 if the proxy for performance is such that a positive reported estimate means better performance (e.g., quantity).	0.869	0.338
Effect: negative	= 1 if the proxy for performance is such that a negative reported estimate means better performance (e.g., time) and thus has to be multiplied by −1 for consistency in our meta-analysis (reference category).	0.131	0.338
Nature of the task			
Task: appealing	= 1 if the performed task is appealing to the subjects; defined following the authors of the primary studies and, when in doubt, following the standards used in psychology (Weibel *et al.*, 2010).	0.482	0.500
Task: unappealing	= 1 if the performed task is not appealing to the subjects (reference category).	0.518	0.500
Task: cognitive	= 1 if the task involved cognitive work; defined following the authors of the primary studies and, when in doubt, based on the standards used in psychology (Condly *et al.*, 2003).	0.705	0.456

Continued on next page

Table 4: Description and summary statistics of regression variables (continued)

Variable	Description	Mean	SD
Task: manual	= 1 if the task involved manual work (reference category).	0.226	0.419
Performance: quantitative	= 1 if the measure of performance is quantitative.	0.702	0.457
Performance: qualitative	= 1 if the measure of performance is qualitative (reference category).	0.298	0.457
Reward scheme			
Reward size	The logarithm of the average payoff from the experiment divided by the logarithm of the median monthly expenditure in the corresponding country (World Bank, data for the year when the experiment was conducted).	0.599	0.292
Positive framing	= 1 if the study rewards its subjects for good performance instead of punishing them for bad performance.	0.831	0.375
Negative framing	= 1 if the study punishes its subjects for bad performance instead of rewarding them for good performance (reference category).	0.151	0.358
All subjects paid	= 1 if all subjects involved in the experiment received any financial payment, = 0 if only some received it.	0.741	0.438
Individual reward	= 1 if, as a reward for the subject's good performance, the subject individually receives a payment.	0.809	0.393
Group reward	= 1 if, as a reward for the subject's good performance, the subject's group receives a payment (reference category).	0.191	0.393
Motivation beyond money			
Motivation: altruism	= 1 if the context of the experiment, the reason why the subjects should show any effort in the absence of monetary incentives, is altruism.	0.291	0.454
Motivation: reciprocity	= 1 if the context of the experiment, the reason why the subjects should show any effort in the absence of monetary incentives, is reciprocity.	0.103	0.304
Motivation: fairness	= 1 if the context of the experiment, the reason why the subjects should show any effort in the absence of monetary incentives, is fairness.	0.151	0.358
Motivation: money only	= 1 if money is the sole context of the experiment (reference category).	0.455	0.482
Study design			
Laboratory experiment	= 1 if the experiment took place in a lab.	0.233	0.423
Field experiment	= 1 if the experiment took place in a field (reference category).	0.767	0.423
Crowding-out theory	= 1 if the study mentions the motivation crowding theory.	0.488	0.500
Structural variation			
Subjects: students	= 1 if the subjects are students only.	0.610	0.488
Subjects: employees	= 1 if the subjects are employees only.	0.072	0.259
Subjects: general	= 1 if the subjects are both students and employees (reference category).	0.318	0.466
Gender: males	The ratio of male to female subjects (= 1 if all male, 0 = if all female).	0.530	0.232
Subjects' age	The logarithm of the average age of the subjects.	2.934	0.320
Data year	The logarithm of the average year of the experiment's time span.	7.606	0.002
Developed country	= 1 if the corresponding country is developed at the time of the experiment (classification based on the World Bank).	0.835	0.369
Developing country	= 1 if the corresponding country is developing at the time of the experiment (reference category).	0.165	0.369
Estimation technique			
Method: OLS	= 1 if the authors use ordinary least squares.	0.571	0.495

Continued on next page

Table 4: Description and summary statistics of regression variables (continued)

Variable	Description	Mean	SD
Method: logit	= 1 if the authors use logit regression.	0.048	0.213
Method: probit	= 1 if the authors use probit regression.	0.090	0.286
Method: tobit	= 1 if the authors use tobit regression.	0.031	0.172
Method: fixed effects	= 1 if the authors use fixed-effects estimation.	0.039	0.193
Method: random effects	= 1 if the authors use random-effects estimation.	0.028	0.165
Method: DID	= 1 if the authors use difference-in-differences estimation.	0.027	0.163
Method: other	= 1 if the authors use a other methods (reference category).	0.037	0.189
Cross-section	= 1 if the data is cross-section instead of panel.	0.446	0.497
Panel	= 1 if the study uses panel data (reference category).	0.554	0.497
Publication characteristics			
Journal impact	The Journal Citation Reports impact factor of the outlet in which the study is published (collected in January 2021).	5.490	3.235
Study citations	The logarithm of the mean number of Google Scholar citations received per year since the study first appeared in Google Scholar (collected in January 2021).	4.839	1.780

Notes: SD = standard deviation.

The setup of many experiments is complex, and the classification into work, charity, and other categories described above does not sufficiently capture the different approaches in the literature. While we cannot hope to capture all the differences, we additionally include a category that reflects the general context of the experiment beyond the main monetary incentive. Often there is no additional context, and money is the only motivation the participants can reasonably have to fulfill the experimental task (such as when they compute dots on the screen). In other cases there are elements of other sources of motivation as well: altruism (a participant can help other participants, but she knows the action will not change her own reward), reciprocity (a form of cooperation for mutual benefit is present in the experiment), and fairness (the experiment features a design related to inequality among participants). An important difference between primary studies, of course, is whether the experiment is conducted in the lab or in the field. About three quarters of the studies in our sample are field experiments. It is also worth noting that only about half of the studies mention the motivation crowding theory, which is ubiquitous in psychology.

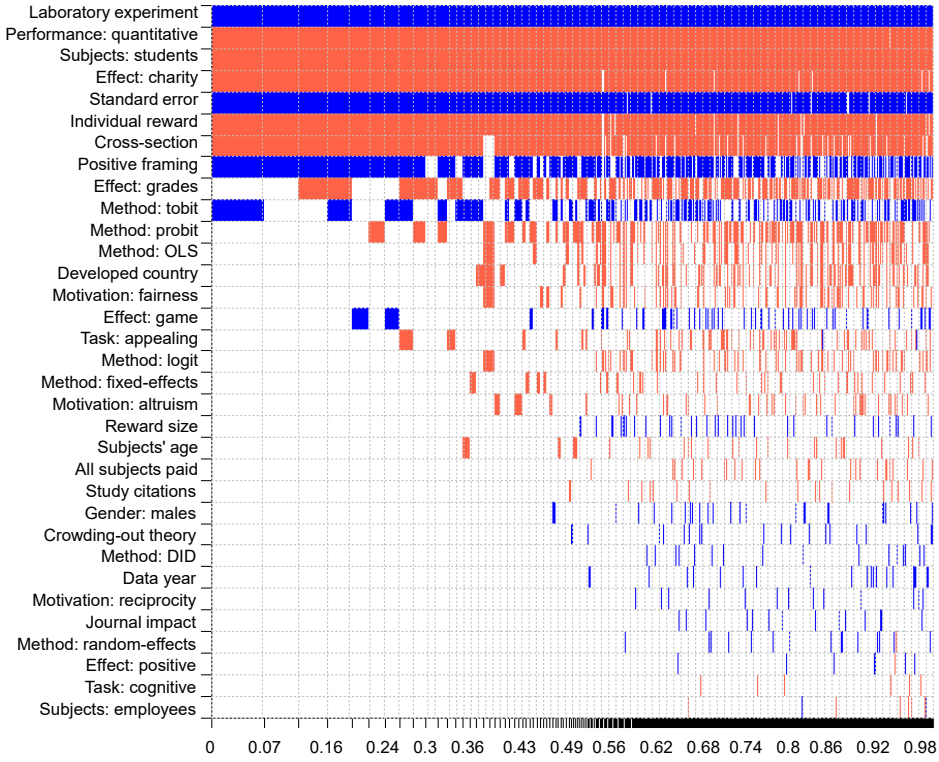
We take into account the differences among participants. Most studies rely on students, but in about 1/3 of the experiments the subject pool has a more general composition. We account for gender differences among participants, their age, and the year and country when and where the experiment was conducted. Only about 16% of the experiments were conducted in developing countries. The studies differ in the estimation technique they employ, though most

commonly OLS is sufficient with experimental data that provide arguably exogenous variation. About half of the studies have multiple observations per participant and employ panel data techniques. Finally, we also control for the publication characteristics of individual studies: the impact factor of the outlet and the number of per-year citations. These variables may reflect aspects of quality not captured by the data and methods variables mentioned above.

In general, there are two ways how to explicitly incorporate heterogeneity in a meta-analysis. First, one can apply a battery of publication bias tests, such as those discussed in the previous section, separately for each category listed in Table 4. In our case the results of such an exercise are similar to the simple summary statistics reported previously in Table 1, but the corrected means for individual categories are even closer to zero according to most bias-correction techniques, and we thus do not report these results. Second, one can add the variables to a specific bias-correction technique, which is the approach we choose in this meta-analysis. After eliminating the reference categories of dummy variables, we are left with 33 factors that can be used in the analysis of heterogeneity. For the bias-correction technique we select the linear meta-regression, regression of estimates on their standard errors. While the technique is based on strong assumptions, in the previous section we show that in the case of the incentive-performance literature it yields results very similar to far more complex techniques. The simplicity and tractability of the linear meta-regression allows us to address two key problems: model uncertainty and collinearity.

Model uncertainty arises inevitably in meta-analysis because it is unclear ex ante which of the many factors capturing study design are systematically important in affecting the estimates reported in primary studies. If the model includes all the factors, it will yield inefficient estimates for those that are systematically important. The problem is discussed in much detail by Steel (2020), who also explains that the natural response to model uncertainty is Bayesian model averaging. Bayesian model averaging considers many models that include different combinations of the explanatory variables (in our case, 2^{33} models) and weights them according to data fit and parsimony. In our application we use agnostic priors recommended by Eicher *et al.* (2011): each model has the same prior probability, and the prior that each regression coefficient is zero has the same weight as one observation in the data. Because there are more than 8 billion models to consider, we use a Markov chain Monte Carlo algorithm (Zeugner & Feldkircher, 2015) to

Figure 7: Model inclusion in Bayesian model averaging



Note: The response variable is the partial correlation coefficient corresponding to the effect of financial incentives on performance reported in individual studies. The columns denote individual regression models; variables are sorted by posterior inclusion probability in descending order. The horizontal axis denotes cumulative posterior model probabilities. The estimation is based on the agnostic unit information g-prior recommended by Eicher *et al.* (2011) and the dilution model prior suggested by George (2010), which penalizes collinearity. Blue color (darker in grayscale) = the variable has a positive estimated effect. Red color (lighter in grayscale) = the variable has a negative estimated effect. No color = the variable is excluded from the model. Table 4 presents a detailed description of the variables. The numerical results are reported in Table 5.

walk only through the most important part of the model mass. While the correlations between the variables we collect are not substantial (see Figure B2 in the Appendix), we additionally use the dilution prior due to George (2010), which penalizes collinearity by assigning less weight to models with a small determinant of the correlation matrix. As a robustness check we employ frequentist model averaging with Mallow's weights (Hansen, 2007) and orthogonalization of model space according to Amini & Parmeter (2012).

Figure 7 illustrates the results of Bayesian model averaging. Each column denotes a regression model, and the width of the column captures posterior model probability (depicted, in cumulative terms, on the horizontal axis). The color of each cell denotes the sign of the estimated regression coefficient: blue means positive, red means negative, and white means zero—in the latter models the corresponding variable is not included. The most important

Table 5: Why do reported effects of financial incentives vary?

Response variable: Partial correlation coefficient	Bayesian model averaging (baseline)			Frequentist model averaging (robustness check)		
	P. mean	P. SD	PIP	Coef.	SE	p-value
Constant	-0.337	NA	1.000	18.83	27.09	0.487
Standard error (pub. bias)	0.439	0.119	0.987	0.518	0.132	0.000
Definition of performance effect						
Effect: grades	-0.017	0.020	0.504	-0.048	0.015	0.002
Effect: charity	-0.052	0.014	0.988	-0.060	0.014	0.000
Effect: game	0.003	0.009	0.125	0.017	0.015	0.259
Effect: positive	0.000	0.001	0.011	-0.003	0.013	0.816
Nature of the task						
Task: appealing	-0.003	0.010	0.123	-0.036	0.014	0.011
Task: cognitive	0.000	0.001	0.010	0.000	0.010	0.962
Performance: quantitative	-0.059	0.012	0.998	-0.043	0.014	0.002
Reward scheme						
Reward size	0.002	0.010	0.062	0.011	0.023	0.628
Positive framing	0.038	0.024	0.776	0.036	0.019	0.068
All subjects paid	-0.001	0.005	0.038	-0.052	0.014	0.000
Individual reward	-0.048	0.014	0.978	-0.085	0.019	0.000
Motivation beyond money						
Motivation: altruism	-0.001	0.006	0.076	-0.034	0.015	0.023
Motivation: reciprocity	0.000	0.003	0.017	-0.003	0.016	0.814
Motivation: fairness	-0.004	0.011	0.138	-0.042	0.015	0.005
Study design						
Laboratory experiment	0.081	0.013	0.999	0.100	0.020	0.000
Crowding-out theory	0.000	0.002	0.026	0.008	0.010	0.445
Structural variation						
Subjects: students	-0.065	0.014	0.998	-0.055	0.015	0.000
Subjects: employees	0.000	0.002	0.009	0.004	0.017	0.798
Gender: males	0.001	0.005	0.035	0.011	0.016	0.500
Subjects' age	-0.001	0.007	0.056	0.023	0.022	0.307
Data year	0.068	0.550	0.022	-2.449	3.559	0.491
Developed country	-0.005	0.012	0.169	-0.036	0.015	0.016
Estimation technique						
Method: OLS	-0.005	0.012	0.170	-0.030	0.013	0.022
Method: logit	-0.007	0.020	0.120	-0.059	0.022	0.009
Method: probit	-0.015	0.024	0.340	-0.048	0.018	0.008
Method: tobit	0.027	0.033	0.446	0.034	0.024	0.167
Method: fixed-effects	-0.003	0.012	0.076	0.027	0.028	0.338
Method: random-effects	0.000	0.004	0.014	0.008	0.022	0.719
Method: DID	0.001	0.006	0.024	0.060	0.032	0.067
Cross section	-0.059	0.021	0.935	-0.046	0.017	0.007
Publication characteristics						
Journal impact	0.000	0.000	0.017	0.001	0.002	0.424
Study citations	0.000	0.001	0.036	-0.004	0.004	0.385
Studies	44			44		
Observations	1,568			1,568		

Notes: The response variable is the partial correlation coefficient corresponding to the effect of financial incentives on performance reported in individual studies. SE = standard error, P. mean = posterior mean, P. SD = posterior standard deviation, PIP = posterior inclusion probability. The posterior mean in Bayesian model averaging (and the "coefficient" in frequentist model averaging) denotes the marginal effect of a study characteristic on the partial correlation coefficient. In Bayesian model averaging we use the agnostic unit information g-prior recommended by Eicher *et al.* (2011) and the dilution model prior suggested by George (2010), which penalizes collinearity. Frequentist model averaging applies Mallow's weights (Hansen, 2007) using orthogonalization of covariate space suggested Amini & Parmeter (2012) to reduce the number of estimated models. For a detailed description of the variables see Table 4.

models, those which fit the data best given their complexity, are shown on the left-hand side. The very best model includes 9 variables out of 33. The sum of posterior model probabilities for all the models in which the corresponding variable is included gives rise to posterior inclusion probability for each variable, which is shown in Table 5 along with other numerical results. Only a handful of variables have posterior inclusion probabilities above 0.5, which means that most of the variables are not useful for the explanation of the differences in the reported incentive-performance effects. Table 5 also reports the results of frequentist model averaging, and all the variables with posterior inclusion probabilities above 0.5 in Bayesian model averaging are also statistically significant at least at the 10% level in frequentist model averaging.

The first important finding of Table 5 concerns publication bias: the correlation between estimates and standard errors is robustly positive even when we explicitly control for various aspects of study design. In fact, the standard error belongs among the variables most effective in explaining the variation in reported incentive-performance effects: the corresponding posterior inclusion probability is 0.99 in Bayesian model averaging, and the p-value is below 0.001 in frequentist model averaging. The result further strengthens the evidence on publication bias presented in the previous section. Next, definition of performance matters: the effect of incentives is smaller for grades and prosocial behavior than for work and game outcomes. The effect is also smaller for quantitative than for qualitative measurement of performance, for negative than for positive framing, for individual than for group rewards, for field than for lab experiments, for students than other subjects, and for cross-section than panel approaches. Crucially, our results suggest that, on average, reward size does not matter for the effect of incentives. Also for the year of data the resulting partial derivative is zero, suggesting that, ceteris paribus, newer studies do not bring larger estimates.

While the variables mentioned above are statistically important in influencing the estimates reported in the literature, the economic effects are small, as shown in Table 6. Even drastic shifts in the variables are associated with relatively modest changes in the partial correlation coefficients, with two exceptions: the standard error (a proxy for publication bias) and a dummy variable for lab experiments. Switching from field to lab experiments can, on average, change the effect from zero to one that can be considered "small" according to the Doucouliagos (2011) guidelines for the interpretation of partial correlation coefficients.

Table 6: Economic significance of key variables

	One-std.-dev. change		Maximum change	
	Effect on PCC	% of mean	Effect on PCC	% of mean
Standard error (pub. bias)	0.019	41%	0.085	184%
Effect: grades	-0.008	-17%	-0.017	-37%
Effect: charity	-0.023	-51%	-0.052	-112%
Performance: quantitative	-0.027	-58%	-0.059	-127%
Positive framing	0.014	31%	0.038	82%
Cross-sectional data	-0.029	-63%	-0.059	-127%
Laboratory experiment	0.034	74%	0.081	175%
Individual reward	-0.019	-41%	-0.048	-104%
Subjects: students	-0.032	-68%	-0.065	-140%

Notes: The table presents the marginal influence of selected variables on the partial correlation coefficient (PCC) corresponding to the effect of financial incentives on performance. The column "one-std.-dev. change" shows how the PCC changes when we increase the value of the variable by one standard deviation. The column "maximum change" represents the change in the PCC when the variable is increased from its minimum to its maximum. The percentage values indicate the magnitude of the implied effect in relation to the sample mean (0.046). For a detailed explanation of the variables, see Table 4.

Table 7: Estimates implied for different contexts

	PCC	95% conf. int.	
Mean best practice	0.010	-0.054	0.075
Effect: grades	0.013	-0.054	0.080
Effect: charity	-0.022	-0.090	0.046
Effect: game	0.033	-0.034	0.099
Effect: work	0.030	-0.038	0.098
Performance: quantitative	-0.007	-0.068	0.054
Performance: qualitative	0.052	-0.013	0.117
Positive framing	0.017	-0.050	0.083
Negative framing	-0.021	-0.088	0.045
Laboratory experiment	0.072	0.004	0.141
Field experiment	-0.009	-0.080	0.063
Subjects: students	-0.055	-0.121	0.012
Subjects: employees	0.010	-0.056	0.077
Individual reward	0.001	-0.067	0.069
Group reward	0.049	-0.019	0.117
Mean based on Lazear (2000)	0.019	-0.044	0.083
Mean based on Angrist & Lavy (2009)	-0.022	-0.055	0.011
Mean based on Takahashi *et al.* (2016)	0.029	-0.075	0.134

Notes: The table presents the partial correlation coefficient (PCC) corresponding to the effect of financial incentives on performance for different contexts implied by the results of Bayesian model averaging and i) our definition of best-practice approach, ii) the approach by Lazear (2000), iii) the approach by Angrist & Lavy (2009), and iv) the approach by Takahashi *et al.* (2016). That is, the table attempts to answer the question what the mean PCC would look like if the literature was approximately corrected for publication bias and all studies in the literature used the same strategy as the one we prefer or the ones employed by Lazear (2000), Angrist & Lavy (2009), and Takahashi *et al.* (2016). Approximate 95% confidence intervals constructed using frequentist model averaging are reported in the last two columns.

As the bottom line of our analysis, in Table 7 we compute the implied incentive-performance effect for different contexts. For the computation we use the results of Bayesian model averaging and construct fitted values of partial correlation conditional on the following values of explanatory variables: zero for the standard error (to correct for publication bias), zero for cross-sectional data (to prefer panel data approaches), sample maximum for the year of the data (to prefer experiments conducted recently), zero for motivation by altruism, reciprocity, and fairness (to filter out potential biases introduced by additional non-monetary motivation), and zero for students alone used as subjects (to prefer more representative subject pools). For all other variables we use sample means, reflecting our agnostic priors. The overall mean incentive-performance effect based on our definition of "best practice" described above is 0.01. Because our definition is subjective, we also use, as robustness checks, the practices used by prominent studies in the literature: Lazear (2000), Angrist & Lavy (2009), and Takahashi *et al.* (2016). The largest of the implied estimates for the overall mean is 0.029. Regarding the implied estimates for individual estimation contexts, we obtain small and statistically insignificant effects in all cases, again with the borderline exception of lab experiments (0.07).

5 Conclusion

We present a meta-analysis of the experimental economics literature measuring the effect of financial incentives on performance. Ours is the first meta-analysis on the topic that corrects the estimates for publication bias. We focus on economics evidence because no previous meta-analysis has concentrated on economics, economics experiments are generally more homogeneous than psychology experiments (Hertwig & Ortmann, 2001), and economists tend to focus on overall performance instead of intrinsic motivation. Economists are also likely to have a prior for the effectiveness of financial incentives, an outcome underlying most conventional models in economics and related fields.

It is therefore all the more remarkable that we find very little evidence for financial incentives to improve performance. Incentives seem to be ineffective not only on average, but also for various individual contexts: work performance, prosocial behavior, games, grades; quantitative and qualitative measurement; interesting and uninteresting tasks; individual and group rewards; various compositions of the subject pool. The size of the reward does not matter in the meta-

analysis, which contrasts the classical result of Gneezy & Rustichini (2000): "Pay enough or don't pay at all." It is also worth noting that most experiments focus on a short time span and are consequently unable to capture potential long-term detrimental effects on intrinsic motivation when financial incentives are removed (Gneezy *et al.*, 2011). A long-term force working in the opposite direction is habit formation (Havranek *et al.*, 2017), which can sustain increased performance even after the removal of incentives.

Our findings are not fully consistent with the motivation crowding theory dominating the psychology literature because the effect of incentives seems to be similarly negligible for both interesting and uninteresting tasks. Another plausible explanation is the distraction effect (Rusz *et al.*, 2020), which makes people concentrate on reward cues instead of the task itself, especially in field settings. It is also possible that economics experiments have, on average, been unable to identify the underlying positive effects of incentives because of measurement error and limited power (Esteves-Sorenson & Broce, 2022). In any case we find it preliminary to proclaim with any certainty that research has showed financial incentives to improve performance. Yet such statements are common in practical applications of economics evidence, as we discuss in the Introduction. Consider, for example, the following statement by the Chartered Institute of Personnel and Development, the largest professional association in human resources, in its recent summary of the corresponding literature:

> *In the past three decades, a large number of high-quality studies and meta-analyses . . . have shown that financial incentives are indeed strongly and positively related to individual performance.* (CIPD, 2022, p. 5)

Three qualifications of our rather depressing results are in order. First, we collect multiple estimates from individual studies, and the estimates are unlikely to be independent within studies. We use two strategies to alleviate this problem: i) study-level clustering of standard errors and ii) weighting by the inverse of the number of estimates reported per study. The weighting scheme ensures that each study has the same prior impact on the results. Second, the individual studies use very different definitions of performance. Therefore we cannot compare the estimates directly, we need to recompute them to a common metric. The only common metric that allows us to recompute all the estimates in our sample is the partial correlation coefficient. The partial correlation coefficient is a measure of statistical, not economic importance. We

partly remedy this problem by using the guidelines for interpreting partial correlation due to Doucouliagos (2011). To construct the guidelines, Doucouliagos (2011) uses a large sample of economics meta-analyses to map partial correlations to elasticities. Our main results (Table 7), however, suggest statistically insignificant effects around zero in any case. In addition, some of the publication bias tests we use rely on the distribution of t-statistics or p-values, and thus are not affected by the transformation. Third, the preferred, instrumental variable solution to the potential violation of the uncorrelation assumption in meta-analysis (Havranek *et al.*, 2022) does not work in the incentive-performance literature because here the instrument is weak. As an alternative, we use the p-uniform* technique recently developed in psychology (van Aert & van Assen, 2021). The technique also finds a very small effect of financial incentives on performance after correction for publication bias.

References

van Aert, R. C. & M. van Assen (2021): "Correcting for publication bias in a meta-analysis with the p-uniform* method." *Working paper*, Tilburg University & Utrecht University.

Alberts, G., Z. Gurguc, P. Koutroumpis, R. Martin, M. Muûls, & T. Napp (2016): "Competition and norms: A self-defeating combination?" *Energy Policy* **96**: pp. 504–523.

Amini, S. M. & C. F. Parmeter (2012): "Comparison of model averaging techniques: Assessing growth determinants." *Journal of Applied Econometrics* **27(5)**: pp. 870–876.

Andrews, I. (2018): "Valid Two-Step Identification-Robust Confidence Sets for GMM." *The Review of Economics and Statistics* **100(2)**: pp. 337–348.

Andrews, I. & M. Kasy (2019): "Identification of and correction for publication bias." *American Economic Review* **109(8)**: pp. 2766–94.

Angrist, J., D. Lang, & P. Oreopoulos (2009): "Incentives and services for college achievement: Evidence from a randomized trial." *American Economic Journal: Applied Economics* **1(1)**: pp. 136–63.

Angrist, J. & V. Lavy (2009): "The effects of high stakes high school achievement awards: Evidence from a randomized trial." *American Economic Review* **99(4)**: pp. 1384–1414.

Ariely, D., A. Bracha, & S. Meier (2009): "Doing good or doing well? Image motivation and monetary incentives in behaving prosocially." *American Economic Review* **99(1)**: pp. 545–55.

Ashraf, N., O. Bandiera, & B. K. Jack (2014): "No margin, no mission? A field experiment on incentives for public service delivery." *Journal of Public Economics* **120**: pp. 1–17.

Barrera-Osorio, F., L. L. Linden, & J. E. Saavedra (2019): "Medium-and long-term educational consequences of alternative conditional cash transfer designs: Experimental evidence from Colombia." *American Economic Journal: Applied Economics* **11(3)**: pp. 54–91.

Blanco-Perez, C. & A. Brodeur (2020): "Publication Bias and Editorial Statement on Negative Findings." *Economic Journal* **130(629)**: pp. 1226–1247.

Bom, P. R. D. & H. Rachinger (2019): "A kinked meta-regression model for publication bias correction." *Research Synthesis Methods* **10(4)**: pp. 497–514.

Boyer, P. C., N. Dwenger, & J. Rincke (2016): "Do norms on contribution behavior affect intrinsic motivation? field-experimental evidence from germany." *Journal of Public Economics* **144**: pp. 140–153.

Bradler, C., S. Neckermann, & A. J. Warnke (2019): "Incentivizing Creativity: A Large-Scale Experiment with Performance Bonuses and Gifts." *Journal of Labor Economics* **37(3)**: pp. 793–851.

Brodeur, A., S. Carrell, D. Figlio, & L. Lusher (2022): "Unpacking p-hacking and publication bias." *Working paper*, University of Ottawa.

Brodeur, A., N. Cook, & A. Heyes (2020): "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* **110(11)**: pp. 3634–3660.

Brodeur, A., M. Le, M. Sangnier, & Y. Zylberberg (2016): "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics* **8(1)**: pp. 1–32.

Brown, A. L., T. Imai, F. Vieider, & C. Camerer (2022): "Meta-Analysis of Empirical Estimates of Loss-Aversion." *Journal of Economic Literature* **(forthcoming)**.

Bruns, S. B. & J. P. A. Ioannidis (2016): "p-Curve and p-Hacking in Observational Research." *PloS ONE* **11(2)**: p. e0149144.

Camerer, C. F. & R. M. Hogarth (1999): "The effects of financial incentives in experiments: A review and capital-labor-production framework." *Journal of Risk and Uncertainty* **19(1)**: pp. 7–42.

Cameron, J. (2001): "Negative effects of reward on intrinsic motivation-A limited phenomenon: Comment on Deci, Koestner, and Ryan (2001)." *Review of Educational Research* **71(1)**: pp. 29–42.

Cameron, J. & W. D. Pierce (1994): "Reinforcement, reward, and intrinsic motivation: A meta-analysis." *Review of Educational Research* **64(3)**: pp. 363–423.

Cappelen, A. W., T. Halvorsen, E. Ø. Sørensen, & B. Tungodden (2017): "Face-saving or fair-minded: What motivates moral behavior?" *Journal of the European Economic Association* **15(3)**: pp. 540–557.

Card, D., J. Kluve, & A. Weber (2018): "What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations." *Journal of the European Economic Association* **16(3)**: pp. 894–931.

Celhay, P. A., P. J. Gertler, P. Giovagnoli, & C. Vermeersch (2019): "Long-run effects of temporary incentives on medical care productivity." *American Economic Journal: Applied Economics* **11(3)**: pp. 92–127.

Cerasoli, C. P., J. M. Nicklin, & M. T. Ford (2014): "Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis." *Psychological Bulletin* **140(4)**: pp. 980–1008.

Charness, G. & U. Gneezy (2009): "Incentives to exercise." *Econometrica* **77(3)**: pp. 909–931.

Charness, G. & D. Grieco (2019): "Creativity and incentives. Journal of the European Economic Association." *Journal of the European Economic Association* **17(2)**: pp. 454–496.

Christensen, G. & E. Miguel (2018): "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature* **56(3)**: pp. 920–980.

CIPD (2022): "Financial Incentives: An Evidence Review." *Scientific summary*, Chartered Institute of Personnel and Development.

Coffman, L. C. (2011): "Intermediation reduces punishment (and reward)." *American Economic Journal: Microeconomics* **3(4)**: pp. 77–106.

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

Conrads, J., B. Irlenbusch, T. Reggiani, R. M. Rilke, & D. Sliwka (2016): "How to hire helpers? Evidence from a field experiment." *Experimental Economics* **19(3)**: pp. 577–594.

De Quidt, J. (2018): "Your loss is my gain: a recruitment experiment with framed incentives." *Journal of the European Economic Association* **16(2)**: pp. 522–559.

Deci, E. L. (1971): "Effects of externally mediated rewards on intrinsic motivation." *Journal of Personality and Social Psychology* **18(1)**: pp. 105–115.

Deci, E. L., R. Koestner, & R. M. Ryan (1999): "A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation." *Psychological Bulletin* **125(6)**: pp. 627–668.

DellaVigna, S. & E. Linos (2022): "RCTs to Scale: Comprehensive Evidence From Two Nudge Units." *Econometrica* **90(1)**: pp. 81–116.

DellaVigna, S., D. Pope, & E. Vivalt (2019): "Predict science to improve science." *Science* **366(6464)**: pp. 428–429.

Dohmen, T. & A. Falk (2011): "Performance pay and multidimensional sorting: Productivity, preferences, and gender." *American Economic Review* **101(2)**: pp. 556–90.

Doucouliagos, H. (2011): "How large is large? Preliminary and relative guidelines for interpreting partial correlations in economics." *Working Papers 5/2011*, Deakin University.

Duflo, E., R. Hanna, & S. P. Ryan (2012): "Incentives work: Getting teachers to come to school." *American Economic Review* **102(4)**: pp. 1241–78.

Dwenger, N., H. Kleven, I. Rasul, & J. Rincke (2016): "Extrinsic and intrinsic motivations for tax compliance: Evidence from a field experiment in Germany." *American Economic Journal: Economic Policy* **8(3)**: pp. 203–32.

Egger, M., G. D. Smith, M. Schneider, & C. Minder (1997): "Bias in meta-analysis detected by a simple, graphical test." *BMJ* **315(7109)**: pp. 629–634.

Eicher, T. S., C. Papageorgiou, & A. E. Raftery (2011): "Default priors and predictive performance in Bayesian model averaging, with application to growth determinants." *Journal of Applied Econometrics* **26(1)**: pp. 30–55.

Elliott, G., N. Kudrin, & K. Wuthrich (2022): "Detecting p-hacking." *Econometrica* **90(2)**: pp. 887–906.

Erat, S. & U. Gneezy (2016): "Incentives for creativ-

ity." *Experimental Economics* **19(2)**: pp. 269–280.

Esteves-Sorenson, C. & R. Broce (2022): "Do Monetary Incentives UnderminePerformance on Intrinsically EnjoyableTasks? A Field Test." *Review of Economics and Statistics* **(forthcoming)**.

Fehr, E. & L. Goette (2007): "Do workers work more if wages are high? Evidence from a randomized field experiment." *American Economic Review* **97(1)**: pp. 298–317.

Fehr, E., H. Herz, & T. Wilkening (2013): "The lure of authority: Motivation and incentive effects of power." *American Economic Review* **103(4)**: pp. 1325–59.

Fehr, E. & J. A. List (2004): "The hidden costs and returns of incentives-trust and trustworthiness among CEOs." *Journal of the European Economic Association* **2(5)**: pp. 743–771.

Fehr, E. & K. M. Schmidt (2007): "Adding a stick to the carrot? The interaction of bonuses and fines." *American Economic Review* **97(2)**: pp. 177–181.

Feldkircher, M. & S. Zeugner (2012): "The impact of data revisions on the robustness of growth determinants—a note on determinants of economic growth: Will data tell?" *Journal of Applied Econometrics* **27(4)**: pp. 686–694.

Fershtman, C. & U. Gneezy (2011): "The tradeoff between performance and quitting in high power tournaments." *Journal of the European Economic Association* **9(2)**: pp. 318–336.

Frey, B. S. & R. Jegen (2001): "Motivation Crowding Theory." *Journal of Economic Surveys* **15(5)**: pp. 589–611.

Friedl, A., L. Neyse, & U. Schmidt (2018): "Payment scheme changes and effort adjustment: the role of 2D: 4D digit ratio." *Journal of Behavioral and Experimental Economics* **72**: pp. 86–94.

Fryer, R. G. (2011): "Financial incentives and student achievement: Evidence from randomized trials." *The Quarterly Journal of Economics* **126(4)**: pp. 1755–1798.

Furukawa, C. (2020): "Publication bias under aggregation frictions: Theory, evidence, and a new correction method." *Working paper*, MIT.

Gallier, C., C. Reif, & D. Römer (2017): "Repeated pro-social behavior in the presence of economic interventions." *Journal of Behavioral and Experimental Economics* **69**: pp. 18–28.

Garbers, Y. & U. Konradt (2014): "The effect of financial incentives on performance: A quantitative review of individual and team-based financial incentives." *Journal of Occupational and Organizational Psychology* **87**: pp. 102–137.

George, E. I. (2010): "Dilution priors: Compensating for model space redundancy." In "IMS Collections Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown," volume 6, p. 158–165. Institute of Mathematical Statistics.

Gerber, A. & N. Malhotra (2008): "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals." *Quarterly Journal of Political Science* **3(3)**: pp. 313–326.

Gneezy, U., S. Meier, & P. Rey-Biel (2011): "When and Why Incentives (Don't) Work to Modify Behavior." *Journal of Economic Perspectives* **25(4)**: pp. 191–210.

Gneezy, U. & A. Rustichini (2000): "Pay enough or don't pay at all." *The Quarterly Journal of Economics* **115(3)**: pp. 791–810.

Hansen, B. (2007): "Least Squares Model Averaging." *Econometrica* **75(4)**: pp. 1175–1189.

Havranek, T., Z. Irsova, L. Laslopova, & O. Zeynalova (2022): "Skilled and Unskilled Labor Are Less Substitutable than Commonly Thought." *The Review of Economics and Statistics* **(forthcoming)**.

Havranek, T., M. Rusnak, & A. Sokolova (2017): "Habit formation in consumption: A meta-analysis." *European Economic Review* **95**: pp. 142–167.

Havranek, T., T. D. Stanley, H. Doucouliagos, P. Bom, J. Geyer-Klingeberg, I. Iwasaki, W. R. Reed, K. Rost, & R. C. M. van Aert (2020): "Reporting Guidelines for Meta-Analysis in Economics." *Journal of Economic Surveys* **34(3)**: pp. 469–475.

Hertwig, R. & A. Ortmann (2001): "Experimental practices in economics: A methodological challenge for psychologists?" *Behavioral and Brain Sciences* **24**: pp. 383–451.

Homonoff, T. A. (2018): "Can small incentives have large effects? The impact of taxes versus bonuses on disposable bag use." *American Economic Journal: Economic Policy* **10(4)**: pp. 177–210.

Hong, S. & W. R. Reed (2021): "Using Monte Carlo experiments to select meta-analytic estimators." *Research Synthesis Methods* **12(2)**: pp. 192–215.

Imai, T., T. A. Rutter, & C. F. Camerer (2021): "Meta-Analysis of Present-Bias Estimation Using Convex Time Budgets." *The Economic Journal* **131(636)**: pp. 1788–1814.

Ioannidis, J. P., T. D. Stanley, & H. Doucouliagos (2017): "The Power of Bias in Economics Research." *Economic Journal* **127(605)**: pp. F236–F265.

Iwasaki, I. (2022): "The finance-growth nexus in Latin America and the Caribbean: A meta-analytic perspective." *World Development* **149(C)**.

Jenkins, G. D., A. Mitra, N. Gupta, & J. D. Shaw (1998): "Are financial incentives related to performance? A meta-analytic review of empirical research." *Journal of Applied Psychology* **83(5)**: pp. 777–787.

Karlan, D. & J. A. List (2007): "Does price matter in

charitable giving? Evidence from a large-scale natural field experiment." *American Economic Review* **97(5)**: pp. 1774–1793.

Kim, J. H., B. Gerhart, & M. Fang (2022): "Do Financial Incentives Help or Harm Performance in Interesting Tasks?" *Journal of Applied Psychology* **107(1)**: pp. 153–167.

Kirchler, M. & S. Palan (2018): "Immaterial and monetary gifts in economic transactions: Evidence from the field." *Experimental economics* **21(1)**: pp. 205–230.

Konow, J. (2010): "Mixed feelings: Theories of and evidence on giving." *Journal of Public Economics* **94(3-4)**: pp. 279–297.

Kremer, M., E. Miguel, & R. Thornton (2009): "Incentives to learn." *The Review of Economics and Statistics* **91(3)**: pp. 437–456.

Kvarven, A., E. Stroemland, & M. Johannesson (2019): "Identification of and Correction for Publication Bias: Comment." *MetaArXiv dh87m*, Center for Open Science.

Lacetera, N., M. Macis, & R. Slonim (2012): "Will there be blood? Incentives and displacement effects in pro-social behavior." *American Economic Journal: Economic Policy* **4(1)**: pp. 186–223.

Lazear, E. P. (2000): "Performance pay and productivity." *American Economic Review* **90(5)**: pp. 1346–1361.

Levitt, S. D., J. A. List, S. Neckermann, & S. Sadoff (2016): "The behavioralist goes to school: Leveraging behavioral economics to improve educational performance." *American Economic Journal: Economic Policy* **8(4)**: pp. 183–219.

Li, Tao, L. H., L. Zhang, & S. Rozelle (2014): "Encouraging classroom peer interactions: Evidence from Chinese migrant schools." *Journal of Public Economics* **111**: pp. 29–45.

Matousek, J., T. Havranek, & Z. Irsova (2022): "Individual Discount Rates: A Meta-Analysis of Experimental Evidence." *Experimental Economics* **25(1)**: pp. 318–358.

McCloskey, D. N. & S. T. Ziliak (2019): "What Quantitative Methods Should We Teach to Graduate Students? A Comment on Swann's Is Precise Econometrics an Illusion?" *The Journal of Economic Education* **50(4)**: pp. 356–361.

McKinsey (2022): "The powerful role financial incentives can play in a transformation." https://www.mckinsey.com/capabilities/transformation/our-insights/the-powerful-role-financial-incentives-can-play-in-a-transformation, published on January 19, 2022.

Meier, S. (2007): "Do subsidies increase charitable giving in the long run? Matching donations in a field experiment." *Journal of the European Economic Association* **5(6)**: pp. 1203–1222.

Mellstrom, C. & M. Johannesson (2008): "Crowding out in blood donation: was Titmuss right?" *Journal of the European Economic Association* **6(4)**: pp. 845–863.

Moral-Benito, E. (2015): "Model Averaging In Economics: An Overview." *Journal of Economic Surveys* **29(1)**: pp. 46–75.

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

Neisser, C. (2021): "The Elasticity of Taxable Income: A Meta-Regression Analysis." *Economic Journal* **131(640)**: pp. 3365–3391.

Oswald, Y. & U. Backes-Gellner (2014): "Learning for a bonus: How financial incentives interact with preferences." *Journal of Public Economics* **118**: pp. 52–61.

Rusnak, M., T. Havranek, & R. Horvath (2013): "How to solve the price puzzle? A meta-analysis." *Journal of Money, Credit and Banking* **45(1)**: pp. 37–70.

Rusz, D., M. L. Pelley, M. Kompier, L. Mait, & E. Bijleveld (2020): "Reward-driven distraction: A meta-analysis." *Psychological Bulletin* **146(10)**: pp. 872–899.

Schall, D. L., M. Wolf, & A. Mohnen (2016): "Do effects of theoretical training and rewards for energy-efficient behavior persist over time and interact? A natural field experiment on eco-driving in a company fleet." *Energy Policy* **97**: pp. 291–300.

Sliwka, D. & P. Werner (2017): "Wage increases and the dynamics of reciprocity." *Journal of Labor Economics* **35(2)**: pp. 299–344.

Stanley, T. D. (2005): "Beyond Publication Bias." *Journal of Economic Surveys* **19(3)**: pp. 309–345.

Stanley, T. D. & H. Doucouliagos (2012): *Meta-regression analysis in economics and business.* New York, USA: Routledge.

Stanley, T. D. & H. Doucouliagos (2014): "Meta-regression approximations to reduce publication selection bias." *Research Synthesis Methods* **5(1)**: pp. 60–78.

Stanley, T. D., H. Doucouliagos, & J. P. Ioannidis (2017): "Finding the power to reduce publication bias." *Statistics in medicine* **36(10)**: pp. 1580–1598.

Stanley, T. D., H. Doucouliagos, & J. P. A. Ioannidis (2022): "Retrospective median power, false positive meta-analysis and large-scale replication." *Research Synthesis Methods* **13(1)**: pp. 88–108.

Stanley, T. D., H. Doucouliagos, J. P. A. Ioannidis, & E. C. Carter (2021): "Detecting publication selection bias through excess statistical significance." *Research Synthesis Methods* **12(6)**: pp. 776–795.

Stanley, T. D., S. B. Jarrell, & H. Doucouliagos (2010): "Could it be better to discard 90% of the data? A statistical paradox." *The American Statistician* **64(1)**: pp. 70–77.

Steel, M. F. J. (2020): "Model Averaging and its Use in Economics." *Journal of Economic Literature* **58(3)**: pp. 644–719.

Sudarshan, A. (2017): "Nudges in the marketplace: The response of household electricity consumption to information and monetary incentives." *Journal of Economic Behavior & Organization* **134(C)**: pp. 320–335.

Sun, L. (2018): "Implementing valid two-step identification-robust confidence sets for linear instrumental-variables models." *Stata Journal* **18(4)**: pp. 803–825.

Takahashi, H., J. Shen, & K. Ogawa (2016): "An experimental examination of compensation schemes and level of effort in differentiated tasks." *Journal of Behavioral and Experimental Economics* **61**: pp. 12–19.

Ugur, M., S. Awaworyi Churchill, & H. Luong (2020): "What do we know about R&D spillovers and productivity? Meta-analysis evidence on heterogeneity and statistical power." *Research Policy* **49**: p. 103866.

Weibel, A., K. Rost, & M. Osterloh (2010): "Pay for Performance in the Public Sector—Benefits and (Hidden) Costs." *Journal of Public Administration Research and Theory* **20(2)**: pp. 387–412.

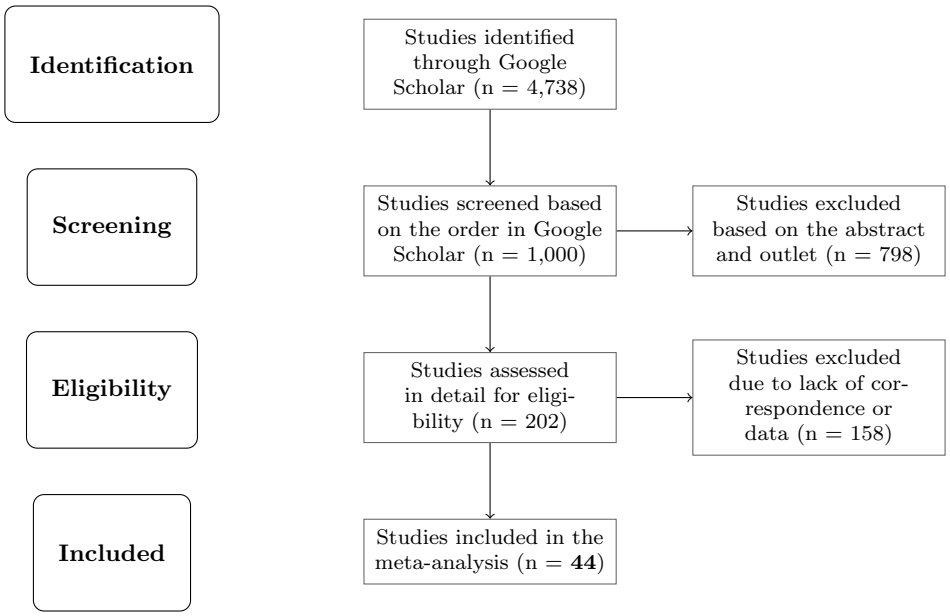
Wiersma, U. J. (1992): "The effects of extrinsic rewards in intrinsic motivation: A meta-analysis." *Journal of Occcupational and Organizational Psychology* **65(2)**: pp. 101–114.

Xue, X., W. R. Reed, & A. Menclova (2020): "Social capital and health: a meta-analysis." *Journal of Health Economics* **72(C)**: p. 102317.

Zeugner, S. & M. Feldkircher (2015): "Bayesian model averaging employing fixed and flexible priors: The BMS package for R." *Journal of Statistical Software* **68(4)**: pp. 1–37.

A Details of Literature Search

Figure A1: PRISMA flow diagram



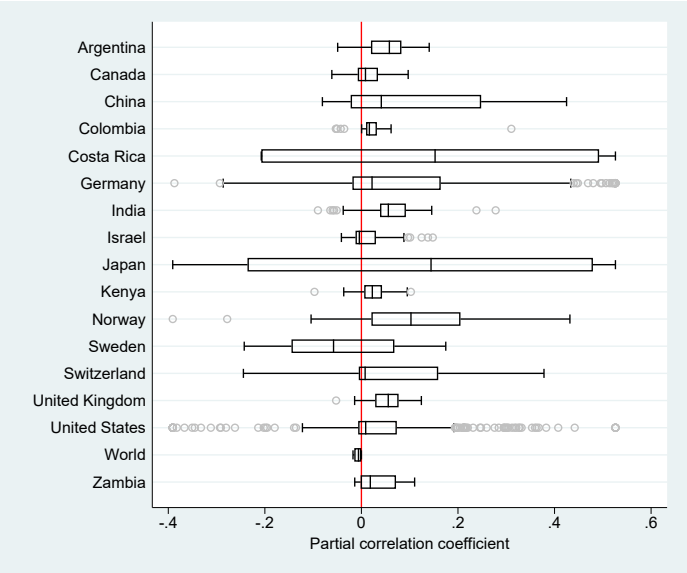
Notes: We use the following query in Google Scholar: `(''financial reward'' OR ''financial incentive'' OR ''money'' OR ''monetary'') AND (''performance'' OR ''motivation'' OR ''effort'') AND ''experiment''`. Note that Google Scholar provides fulltext search, not only the search of the title, abstract and keywords; consequently, our query is broad and inclusive. In the screening stage we only consider studies published in the top 50 economics journals according to the discounted recursive impact factor in RePEc. The search was terminated on January 31, 2021. The list of the 44 studies included in the meta-analysis is available in Table A1; the collected dataset is available in the online appendix at `meta-analysis.cz/incentives`. PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses. More details on PRISMA and reporting standards of meta-analysis in general are provided by Havranek *et al.* (2020).

Table A1: Studies included in the meta-analysis

Alberts *et al.* (2016)	Dohmen & Falk (2011)	Konow (2010)
Angrist & Lavy (2009)	Duflo *et al.* (2012)	Kremer *et al.* (2009)
Angrist *et al.* (2009)	Dwenger *et al.* (2016)	Lacetera *et al.* (2012)
Ariely *et al.* (2009)	Erat & Gneezy (2016)	Lazear (2000)
Ashraf *et al.* (2014)	Fehr & List (2004)	Levitt *et al.* (2016)
Barrera-Osorio *et al.* (2019)	Fehr & Goette (2007)	Li *et al.* (2014)
Boyer *et al.* (2016)	Fehr & Schmidt (2007)	Meier (2007)
Bradler *et al.* (2019)	Fehr *et al.* (2013)	Mellstrom & Johannesson (2008)
Cappelen *et al.* (2017)	Fershtman & Gneezy (2011)	Nagin *et al.* (2002)
Celhay *et al.* (2019)	Friedl *et al.* (2018)	Oswald & Backes-Gellner (2014)
Charness & Gneezy (2009)	Fryer (2011)	Schall *et al.* (2016)
Charness & Grieco (2019)	Gallier *et al.* (2017)	Sliwka & Werner (2017)
Coffman (2011)	Homonoff (2018)	Sudarshan (2017)
Conrads *et al.* (2016)	Karlan & List (2007)	Takahashi *et al.* (2016)
De Quidt (2018)	Kirchler & Palan (2018)	

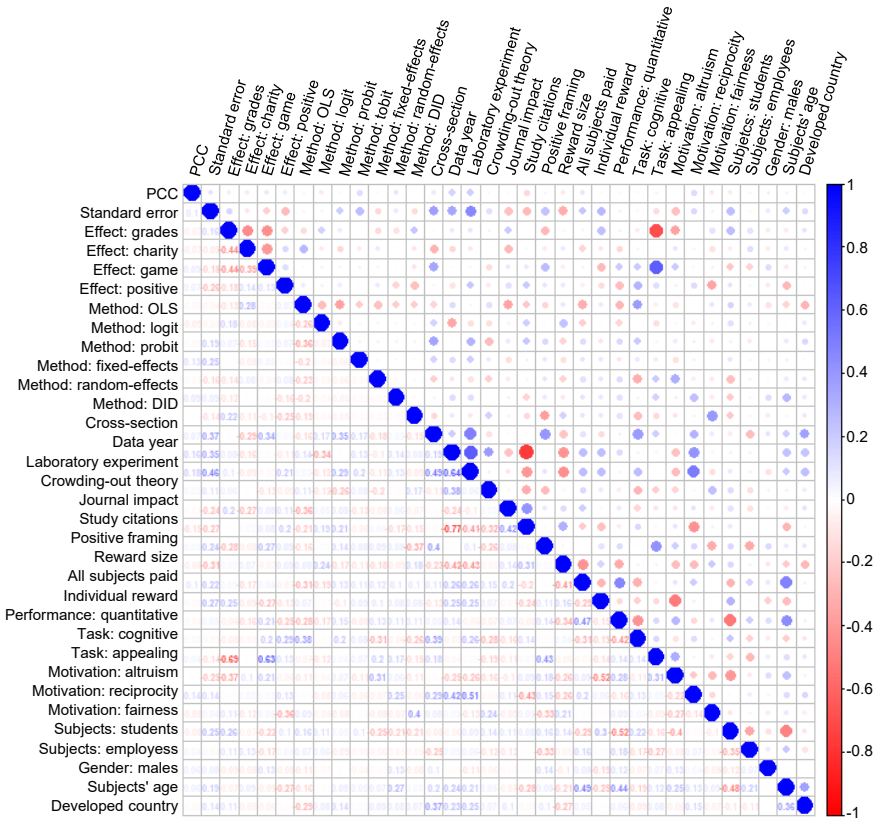
B Additional Details on the Dataset

Figure B1: No systematic differences in results across countries



Notes: See notes to Figure 3.

Figure B2: Correlation matrix for regression variables



Note: The figure shows correlation coefficients for variables described in Table 4. Blue color (dark in grayscale) indicates positive correlation, while red color (light in grayscale) indicates negative correlation.