CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Do money rewards motivate people? A meta-analysis

Bachelor's thesis

Author: Petr Čala

Study program: Economics and Finance

Supervisor: doc. PhDr. Zuzana Havránková, Ph.D.

Year of defense: 2021

Declaration of Authorship

The author hereby declares that he or she compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis in whole or in part and agrees with the thesis being used for study and scientific purposes.

Prague, April 21, 2021

Petr Cala

Abstract

Do financial incentives motivate people to work better? A plethora of research papers in psychology have long tried to answer this question, together with more recent papers from behavioral economics. We take a stock of emerging research in economics and conduct a quantitative analysis from a strictly economic point of view. We collect a total of 1568 estimates from 44 different studies and codify over 30 variables to capture the underlying nature of the effect money has on motivation and performance. A range of statistical tests suggests the overall effect to be virtually zero, which we confirm using a specific design check. We then employ Bayesian and frequentist model averaging to identify the most prominent determinants of the effect. Among these, publication bias pushes this effect upwards the most, along with laboratory setting and positive framing in the task. Six variables then pull the effect in the opposite direction - school setting, charitable giving, cross-sectional data, self-reward, quantitative performance, and students subgroup.

JEL Classification D64, D90, D91, Z10, Z13,
Keywords financial, rewards, motivation, performance, effort

Title Do money rewards motivate people? A meta-analysis

Author's e-mail 61505008@fsv.cuni.cz
Supervisor's e-mail zuzana.havrankova@fsv.cuni.cz

Abstrakt

Jsou lidé motivováni peněžními odměnami? Zodpovězení této otázky se již dlouho věnuje řada výzkumných prací z psychologie, společně s nedávnými pracemi z oboru behaviorální ekonomie. V této práci se zaměřujeme na nově publikované studie z oblasti ekonomie a sestavíme kvantitativní analýzu z čistě ekonomického hlediska. Celkově nasbíráme 1568 odhadů ze 44 různých studií a následně zakódujeme více než 30 proměnných s cílem zachytit podstatu chování efektu peněz na motivaci a výkon. Řada statistických testů nám naznačuje, že efekt je velmi blízko nule, což je dále podpořeno testem pro specifický design. Dále využijeme bayesovského a frekventistického průměrování modelů pro zachycení nejdůležitějších vlivů na zmíněný efekt. Z těchto pozorujeme nejvyšší vliv směrem vzhůru u publikačního zkreslení, dále pak u laboratorních experimentů a pozitivních odměn. Šest proměnných pak má opačný efekt - těmito jsou školní a charitativní experimenty, průřezová data, odměna pro sebe sama, kvantitativní výkon a podskupina studentů.

Klasifikace JEL	D64, D90, D91, Z10, Z13,
Klíčová slova	finanční, odměny, motivace, výkon, snaha

Název práce	Jsou lidé motivováni penežními odměnami? Meta analýza
E-mail autora	61505008@fsv.cuni.cz
E-mail vedoucího práce	zuzana.havrankova@fsv.cuni.cz

Acknowledgments

I am incredibly grateful to the doc. PhDr. Zuzana Havránková, Ph.D., for leading me through the treacherous waters of meta-analytic methodology, providing me with valuable and swift feedback, and letting me quickly understand new techniques through the code of meta-analyses conducted by her and her colleagues. In a similar light, I would like to thank the Professor of Economics at Harvard University Isaiah Andrews and two Associate Professors, Maxmilian Kasy at Oxford University and Chishio Furukawa at MIT, for indirectly providing me with code for their new statistical methods. Lastly, I would like to show appreciation to my good friend Daniel Bartušek, who offered me invaluable advice along with a number of clever suggestions throughout my work on this thesis.

Typeset in FSV LaTeX template with great thanks to prof. Zuzana Havrankova and prof. Tomas Havranek of Institute of Economic Studies, Faculty of Social Sciences, Charles University.

Bibliographic Record
Cala, Petr: *Do money rewards motivate people? A meta-analysis.* Bachelor's thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2021, pages 71. Advisor: doc. PhDr. Zuzana Havránková, Ph.D.

Contents

List of Tables

List of Figures

Acronyms

BMA	Bayesian Model Averaging
BPE	Best Practice Estimate
CET	Cognitive Evaluation Theory
EK	Endogenous Kink
FAT	Funnel Asymmetry Test
FE	Fixed-Effect
FMA	Frequentist Model Averaging
GIT	General Interest Theory
GPA	Grade Point Average
IV	Instrumental Variable
MSE	Mean Squared Error
OLS	Ordinary Least Squares
PCC	Partial Correlation Coefficient
PIP	Posterior Inclusion Probability
PMP	Posterior Model Probability
PET	Precision Effect Test
RE	Random-Effect
SE	Standard Error
VIF	Variance Inflation Factor
WAAP	Weighted Average of Adequately Powered
WLS	Weighted Least Squares

Chapter 1

Introduction

If one were to ask an economist how to get people to work better, he would probably be suggested to try increasing people's pay. After all, it is common knowledge to economists that people respond to incentives (Mankiw 2014). On the flip side, there might be a different response when posing the same question to a psychologist. As suggested by Deci et al. (1999), a psychologist may object that an emphasis on the reward for performing a task may diminish the enjoyment one associates with said task, possibly resulting in decreased motivation and performance during the task. Deci (1971) observed this phenomenon and studied it empirically, following which it became commonly known as the "crowding-out intrinsic motivation" theory. Since then, decades of research have further followed in similar spirit, providing countless experiments and several meta-analyses, which helped put these findings into a quantitative perspective. Here is a short peek into what could be considered the most prominent ideas from this field. Jenkins et al. (1998) associate money reward with higher performance only in the production quantity rather than its quality. Cameron & Pierce (1994) and Deci et al. (1999) claim that such rewards could go a long way into people increasing their performance during tasks where they display little to no interest. Bridging the gap between psychology and economics, Camerer & Hogarth (1999) look at tasks involving judgment and find a positive monetary incentive effect. Bonner et al. (2000) then observe that money positively influences performance only in less cognitive tasks. Going against the theory of crowding-out intrinsic motivation, Cerasoli et al. (2014) show that extrinsic and intrinsic incentives can play a simultaneous role in predicting one's performance. Regarding the economic aspect of this subject, we would like to mention the highly influential study of Gneezy & Rustichini (2000), who claim that one has to be paid enough if there should appear an increase in performance. Another valuable discussion about the topic's economic side is presented, for example, in Ariely et al. (2009), who discover that excessive rewards may have detrimental effect on performance.

The effect of rewards-motivation is quite well defined and observed in psychology, thanks to the numerous already existing quantitative studies. Nevertheless, to our awareness, no such synthetic study exists that would look at how the effect behaves across literature from a strictly economic perspective. Hence, this thesis aims to synthesize the decades of research on this topic in a quantitative meta-analysis and evaluate the effect of rewards-motivation utilizing literature purely from the field of economics. To briefly clarify, a meta-analysis is defined as a method by which one aggregates individual studies together statistically, allowing them to observe the underlying relationships and causalities, along with potential systematic misbehavior (Hunter et al. 1982). Our take on meta-analysis lies in employing several linear and non-linear methods to uncover potential publication bias in the literature, along with model averaging methods. Overall, we find a minimal overall impact of rewards on motivation, together with a noticeable publication bias. As far as literature heterogeneity is concerned, similar results highlight the importance of individual driving factors behind the effect. Most of these results are also in line with the 'crowding-out of intrinsic motivation' theory (Deci 1971) and suggest the effect's behavior to be similar across different fields of studies.

The rest of the thesis is structured as follows: Chapter 2 explores our topic's background, along with the most prominent already existing research and our contribution to it. Chapter 3 describes the methodology and the collection of our data set. Chapter 4 goes over various statistical tests for publication bias and presents their results. Chapter 5 uses model averaging to explain the influence of literature heterogeneity on our findings. Chapter 6 estimates the best-practice effect in the literature. Chapter 7 then concludes the thesis. We include the data set along with the code in the online appendix.

Chapter 2

Estimating the effect of rewards on motivation

As far as the theoretical background on the interaction between rewards and intrinsic motivation is concerned, numerous studies have established a strong foundation of beliefs and results. Consequently, we decided to focus in this chapter only on a handful of the theories most important to our research, along with the already existing meta-analyses conducted on a similar topic. In discussing the results of previous studies, we hope to summarize to the reader the most important ideas and findings from the topic, along with the contribution our work hopes to bring.

2.1 Theoretical overview

Chapter 1 mentioned the role of the 'crowding-out intrinsic motivation' theory (Deci 1971) on people's motivation when presented with incentives. Looking at how this motivation changes when we vary the type of rewards (e.g., verbal versus tangible), it is possible to refer to another theory that tries to describe this problem. The theory is commonly known as the Cognitive Evaluation Theory (CET), and Deci & Ryan (1985) provide its detailed explanation. In short, the authors argue that intrinsic motivation is tightly connected to people's self-perception of their competence and determination. In other words, the driving force (i.e., motivation) when performing a task is directly influenced by how competent they feel while doing so. Deci & Ryan (1985) then differentiate between two main groups of rewards that can affect this self-perception. The first group, which leads to an increase of one's perceived competence (verbal rewards, for example), should enhance an individual's intrinsic motivation. The latter group, which decreases one's perceived competence when presented (tangible rewards, for example), should have the opposite effect and undermine intrinsic motivation. Ryan (1982) then adds to this claim stating that this is true in both directions as long as an individual's competence is perceived together with one's self-determination to perform the task. We discuss the results of empirical testing for the validity of this theory further below.

We can explore the effect of rewards on motivation even further by looking at the General Interest Theory (GIT), discussed by Eisenberger et al. (1999). Compared to the CET, this theory takes both the positive and negative effects of the rewards into consideration. Furthermore, according to the authors, a subject's intrinsic motivation should increase or decrease depending on the task's relevance regarding the subject's satisfaction, needs, and desires. The rewards are then observed instead as a means of altering the subject's self-determination and, in consequence, their motivation for doing the task.

When conducting our meta-analysis, we will not be testing for the validity of these theories, as a number of such tests already exist, and we will discuss these further below. Instead, we present the theoretical overview here to gain consciousness of the possible effects various factors might have on our results.

2.2 Previous research

The meta-analysis we present in this article is certainly not the only one carried out on the topic of rewards-motivation. According to our knowledge, there exist 11 different meta-analyses already published on this topic until this point, namely Rummel & Feinberg (1988); Wiersma (1992); Cameron & Pierce (1994); Tang & Hall (1995); Jenkins et al. (1998); Deci et al. (1999); Eisenberger et al. (1999); Cameron (2001); Deci et al. (2001); Cerasoli et al. (2014); Van Iddekinge et al. (2018). Further in the text, we shall refer to these as the 'primary studies.' Eight of them, specifically all except Jenkins et al. (1998); Cerasoli et al. (2014), and Van Iddekinge et al. (2018), focus mainly on the undermining effect, along with the respective theories, which we briefly described above. The other three studies then look at the relation of rewards, motivation, and cognitive ability to performance.

In their methodology, the use of Cohen's d as an indicator for the undermining effect appears to be the most prevalent. Hedges & Olkin (2014) describe the calculation of Cohen's d as obtaining the difference between the means of the treatment group and the control group and then dividing the result by the pooled within-group standard deviations while adjusting for sample size. The result's size then indicates either the enhancement effect in case of a positive sign or the undermining effect in the opposite case. Only two of the primary studies do not make use of this measure. Cerasoli et al. (2014) test for their main effect using Pearson correlation (denoted by 'ρ'). More specifically, they take inspiration from Hunter & Schmidt (2004) and compute the 'corrected population correlation', which involves assuming population-level estimates of the effect. As mentioned above, this correlation mainly captures the relationship between motivation and performance. The second study that employs a methodology different from Cohen's d is Van Iddekinge et al. (2018). The authors choose their methods following Hunter & Schmidt (2004) as well, computing the corrected population correlation (in this case between ability, motivation and performance). Furthermore, they compute the relative weight statistics in a regression model (Johnson 2000), which allows them to look at effect sizes, rather than statistical significance.

As for the primary studies' actual findings, we would first like to summarize their testing results for the undermining effect. Rummel & Feinberg (1988) present the first complete meta-analysis of the effect, but find no signs of rewards undermining motivation (Cohen's d = 0.329). Wiersma (1992) then claims the opposite, reporting d = -0.5 for a group, which had a free-time on a reward-contingent task. Subsequent studies then start to differentiate between the effect of tangible and verbal incentives on motivation. Deci et al. (1999) measure a positive effect of verbal rewards on intrinsic motivation (Cohen's d = 0.3) and put this positive relationship into direct contrast with tangible rewards, with which the same study suggests the opposite is the case (d = -0.4 for task-contingent rewards). Similar numbers are presented by Cameron & Pierce (1994) and Tang & Hall (1995), who find that verbal praise increases intrinsic motivation (d = 0.38 & d = 0.34) and that tangible, task-contingent rewards undermine this motivation (d = -0.21 & d = -0.51). Deci et al. (2001) then strengthen this claim by providing a result of d = -0.39, again with the task-contingent rewards. The last of the primary studies, which focus on the undermining effect, namely Eisenberger et al. (1999); Cameron (2001); Deci et al. (2001), more or less present results, which are in line with the other six analyses. Their claims differ mainly in technical details of the topic, so for further detail, we refer the reader to the original works. Nevertheless, we can say that a possible undermining effect might appear in our work due to the reward scheme we choose to employ, and which we further discuss in Chapter 5.

If we should briefly summarize the results of the rest of the primary studies, Jenkins et al. (1998) report an effect of financial incentives on performance, which is not statistically different from zero. Cerasoli et al. (2014) then take a step away from the other meta-analyses and observe the combined effect of both intrinsic and extrinsic motivation on performance and find that intrinsic motivation is a solid predictor of performance regardless of whether rewards are present or not ($\rho = .21$-$.45$). Their results suggest that the intrinsic motivation and rewards do not have to work in the opposite direction, which directly contrasts with one part of the undermining theory. Nevertheless, Van Iddekinge et al. (2018) bring uncertainty to the previous suggestion by showing that the ability-motivation interaction may not fully explain or predict performance.

2.3 Our contribution

Having gone only as far as the shore of the theory behind rewards-motivation, we now redirect our focus to the present thesis. Similarly to the works of Cerasoli et al. (2014), and Van Iddekinge et al. (2018), we leave the topic of undermining intrinsic motivation and instead try to answer a question yet unexplored, namely 'how does the effect of rewards-motivation behave when we look purely at the field of economics and what could drive this effect.' Thus, where our approach will differ significantly from the primary studies is in the sample of studies on which we will collect our data. Being the first to impose such a restriction, we will be filtering out only those studies published in economic journals. This filter should allow us to observe how the effect behaves from an economic point of view, and further, we may compare this behavior to the existing results of the primary meta-analyses. This approach makes our study the first of its kind to our knowledge, and by it, we hope to bring a new outlook on the already well-discovered topic.

Another insight we are looking to bring to this field lies in our use of the methodology. First, our approach to quantifying the effect is going to differ considerably from the primary meta analyses. As we plan to focus more on the change in performance itself, rather than its connection to motivation or rewards, a measure allowing for a measurable comparison would be the most appropriate. Thus we plan to employ a measure not yet used in this field of research, namely the partial correlation coefficient.

Second, we observed that out of the eleven of these studies, only Cameron

& Pierce (1994) explicitly mention and look for publication bias in their works. Several others are aware of its presence, but none of these studies go into further detail. We plan to search for this bias extensively using a plethora of statistical tests and hope to explain how it could affect the effect's behavior; this relationship the primary studies do not consider.

Lastly, a significant contribution we hope to present in terms of methodology is our search for heterogeneity in the literature. Because we will be working with a completely new, unexplored set of data, we intend to quantitatively define the influence of numerous factors on the underlying effect using model averaging. With these steps taken, our thesis should serve the existing results as a robustness check, employed in a wholly new and different data environment.

Chapter 3

Assembling the data set

3.1 Literature search

To assemble the list of studies to collect, we started with constructing a query using various combinations of terms tightly connected to the relationship between financial incentives and motivation (e.g. 'financial rewards,' 'monetary incentives,' 'performance,' 'motivation,' 'reward'). The final form design aimed to return the most relevant studies exclusively from the field of economics. With this query, we conducted a thorough search using the Google Scholar database and looked through the top 30 economic journals according to the IDEAS/RePEc aggregate rankings.[1]

The Google Scholar search using the query yielded 202 studies in total, which we categorized and saved. From here, we established the criteria for filtering out irrelevant or unusable studies. Firstly, the final data set includes only those papers, which capture an experiment or a study observing the relationship between an incentive and its effect on subjects' measurable performance. Furthermore, studies not reporting an effect with its standard errors were also discarded, given the nature of the methods later used in this thesis. The result yielded a total of 44 relevant studies, which we then coded. Their list, along with the finalized query, can be found in Appendix A.

[1] The search, along with the final list of journals to go through, was completed in July 2020, taking into account the volatile nature of the IDEAS/RePEc ranking.

3.2 Data collection

To decide which variables are ideal for our purpose, we thoroughly read each study and established a coding scheme. The complete list of the variables and detailed reasoning behind the choice of the more technical ones appears in Chapter 5. In this section, we shall only briefly comment on the structure of some of the usual variables. We split the description of our variable choice like this to better justify our thought process and procedures regarding other meta-analyses conducted on a similar topic, which we do not discuss until further on. For now, we will go over but a few of the more than 60 study characterizations collected.

The original effect, collected along with its standard error, had to capture the relationship between monetary incentives and a measurable kind of output (e.g., a change in physical/mental performance, pro-social behavior, students' Grade Point Average (GPA), among others). Based on the type of these outputs, we then categorized the effect. The 'effect variable' captures this categorization and allows us to focus on different kinds of performance while simultaneously using the partial correlation coefficient to observe the effect as a whole.

Besides these two measures, the rest of the usual variables includes statistical information, the number of observations, sample size, name of the dependent variable, or whether the data was panel or cross-sectional. Furthermore, we denoted the time horizon over which the experiment took place along with the range of years, the journal impact according to RePEc, and the number of citations for each study. Lastly, we included a section detailing additional information about the paper, which could not be quantified. This setup allowed us to observe the influence of various factors in the data set.

With these specifications, 1655 estimates were collected together with their respective variables, yielding more than 120,000 data points in total.

3.3 Initial analysis

We then cleaned the data set, observed the summary statistics for each of the variables, and removed the redundant ones. We purposely included the variable 'grp_reward' in the data set, which is equal to 1 if the observation captures a treatment group. The initial purpose of this variable was to control for discrepancies between the control and treatment groups. However, we decided not to use this approach in the end, and the final data set contains only

1568 observations, which correspond to only the groups that received a reward. The remaining 87 observations, corresponding to the control groups, were thus discarded. We did this mainly to unify the reward effect across observations, which allows for a more straightforward computation.

It is important to note that some of the effects capture a positive influence on performance, while others capture this influence in the opposite direction. In the first case, the higher the number, the better the performance, such as when measuring the number of clicks a subject makes in a given time frame. In the other case, a higher number indicates worse performance, such as when measuring the time taken to finish a task. We decided to remedy this problem by using a dummy equal to one if this relationship is positive. Using such a dummy allowed us to unify the effect's direction, meaning that an increase in the effect size will always indicate better performance/outcome. Consequently, it became straightforward to compare the various effects.

Contrary to the primary meta-analyses, our approach turns its focus in a new direction and focuses more on the outcome itself, so the use of Cohen's d or Pearson correlation does not seem appropriate. Given the diverse nature and size of the collected estimates, we instead need a measure that would allow us to unify and compare the varying effects. Partial Correlation Coefficient (PCC) is presumably the most fitting choice, being a standard in numerous meta-analyses (Doucouliagos & Laroche 2003; Zhou et al. 2013; Zigraiova & Havránek 2016). In short, it is a measure capturing the strength of the relationship between two variables using t-values and degrees of freedom (Stanley & Doucouliagos 2012). Choosing this procedure allows us to partially mitigate the differences in scales of the effect and its nature while highlighting the size of the relationship between our two variables, rewards and performance.

To calculate the partial correlation coefficient, we use the following formula:

$$PCC = \frac{t}{\sqrt{t^2 + df}}, \tag{3.1}$$

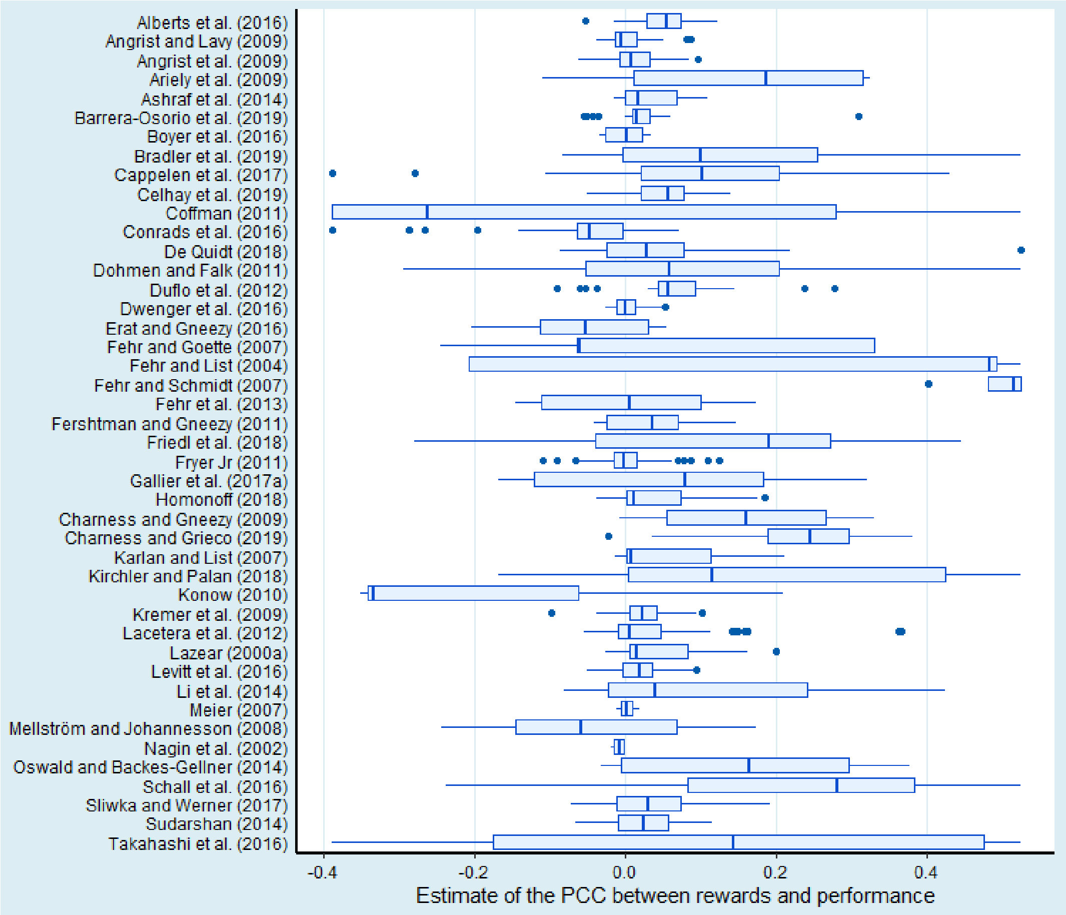
where t stands for the t-statistic of the reported coefficient and df indicates the number of degrees of freedom in the estimation. One can observe the implication of the earlier mentioned criterion for study inclusion in the data set, namely the requirement for the inclusion of both the effect and its standard error. This rule allows us to calculate the t-statistic for *all* collected observations, further establishing the PCC as the optimal choice for measuring the effect.

To obtain the corresponding standard errors of the PCC, we carry out the following calculation:

$$SE_{PCC} = \sqrt{\frac{(1 - PCC^2)}{df}}. \tag{3.2}$$

Considering further procedures, we now calculate standard error's precision as simply $1/SE_{PCC}$. To deal with potentially misleading outliers in the data, we also winsorize the estimates at the 1% level. We decide to choose this level after calculating the trade-off between an artificial intervention (i.e., winsorization) and the results' stability. Given these calculations, we can quickly look at how the PCC behaves across our data set.

Figure 3.1: Partial correlation coefficient across individual studies



Note: This figure shows a box plot of the partial correlation coefficient estimates across individual studies on the data we obtained after winsorization. PCC = Partial Correlation Coefficient.

To further evaluate the behavior of the underlying effect in our data, we calculate the mean of said effect and the corresponding confidence intervals across various subsets of data. We present these statistics in table 3.1.

The baseline effect shows a mean of 0.046 and suggests a minimal, almost negligible incentives-motivation effect. The rest of the table tells a very similar story, coming closer to 0.05 when the sample size is large and deviating either up or down when this size is decreasing. Several variables then display particularly fascinating results, although we can not draw decisive conclusions about those defined by a small number of observations (such as 'Trust' or the 'Methodology' variables). We can instead take a look at the outlying results of the variables with a decent number of observations, say more than one hundred, which corresponds to a little over 5% of the sample. Going through these hierarchically, we would first like to highlight the increased effect size during game-based and work-based tasks (0.073 & 0.067). The 'Game' variable specifically then retains this above average size even through the weighing procedure (0.085), proposing that the subjects show more effort during the typical controlled experiment with a game as the task. This fining appears to be backed up by the unusually large coefficient tied to the 'Lab study' variable (0.091), which equals almost three times its field counterpart even across weighted specifications (0.100). Furthermore, we can also notice the effect observed during the appealing tasks, which is more than 2.5x larger than the one observed during the non-appealing tasks (0.069 & 0.025) even when weighted by the number of estimates (0.063 & 0.014). This difference is also highly statistically significant, suggesting that the task nature might substantially determine subjects' motivation. The last of these coefficients that we would like to point out is the 'Reciprocity' coefficient (0.100), which again remains extraordinarily high even after the weighing procedure (0.110). This unusual increase proposes that social influence may play a large role in determining the subjects' motivation during the experiment. However, we can not claim that with certainty at this point.

We must keep in mind that these are just the initial findings and should provide but a quick insight into the effect's behavior in our data set. Only once we have already established more robustly how the effect behaves, we compare these results to the ones obtained by other studies. For this comparison, along with a detailed discussion about the individual variables, we refer the reader to Chapter 5.

Table 3.1: Mean statistics across various subsets of data

	Unweighted			Weighted			
	Mean	95% conf. int.		Mean	95% conf. int.		No. of observations
All estimates	0.046	0.039	0.054	0.040	0.032	0.048	1568
Effect characteristics							
GPA of students	0.029	0.023	0.035	0.012	0.006	0.018	540
Charity	0.035	0.028	0.042	0.053	0.046	0.059	444
Game	0.073	0.049	0.097	0.085	0.060	0.110	437
Work	0.067	0.039	0.095	0.039	0.011	0.067	147
Positive effect	0.050	0.043	0.057	0.041	0.035	0.048	1362
Negative effect	0.023	-0.015	0.062	0.004	-0.035	0.042	206
Methodology							
OLS	0.042	0.032	0.053	0.026	0.016	0.037	895
Logit	-0.007	-0.021	0.007	0.003	-0.012	0.017	75
Probit	0.034	0.002	0.066	0.018	-0.014	0.051	141
Tobit	0.140	0.046	0.241	0.042	-0.055	0.140	48
Fixed-effects	0.026	0.007	0.046	0.026	0.007	0.046	61
Random-effects	0.120	0.061	0.176	0.050	-0.007	0.108	44
Diff-in-diff	0.045	0.025	0.064	0.045	0.025	0.064	43
Other method	0.088	0.052	0.124	0.042	0.005	0.078	58
Study specifications							
Cross-sectional data	0.057	0.042	0.072	0.052	0.037	0.068	700
Panel data	0.038	0.031	0.044	0.019	0.012	0.025	868
Lab study	0.091	0.072	0.110	0.100	0.082	0.120	366
Field study	0.033	0.025	0.041	0.034	0.026	0.042	1202
Crowding-out	0.051	0.041	0.060	0.033	0.023	0.042	765
Reward scheme							
Positive framing	0.048	0.039	0.058	0.040	0.031	0.049	1303
Negative framing	0.033	0.022	0.044	0.031	0.020	0.041	189
Reward scaled ≥ 0.2	0.074	0.062	0.086	0.063	0.050	0.075	644
Reward scaled < 0.2	0.027	0.017	0.037	0.013	0.003	0.023	924
All paid	0.054	0.044	0.064	0.049	0.039	0.059	1162
Reward own	0.045	0.035	0.054	0.025	0.016	0.035	1268
Reward else	0.054	0.043	0.065	0.058	0.047	0.069	300
Task nature							
Quan. performance	0.043	0.033	0.053	0.043	0.033	0.053	1101
Qual. performance	0.054	0.044	0.065	0.033	0.022	0.043	467
Cognitive task	0.049	0.039	0.059	0.046	0.036	0.056	1106
Manual task	0.052	0.037	0.066	0.038	0.023	0.052	355
Appealing task	0.069	0.054	0.085	0.063	0.048	0.078	755
Non-appealing task	0.025	0.020	0.030	0.014	0.009	0.019	813
Motivation							
Altruism	0.046	0.037	0.056	0.055	0.046	0.065	456
Trust	0.210	0.092	0.327	0.082	-0.035	0.200	24
Reciprocity	0.100	0.079	0.126	0.110	0.091	0.139	161
Fairness	0.020	-0.013	0.052	0.024	-0.008	0.0569	237
Monetary	0.037	0.028	0.046	0.014	0.005	0.023	690
Subject and country characteristics							
Students	0.038	0.027	0.049	0.025	0.014	0.036	957
Employees	0.065	0.039	0.091	0.062	0.036	0.088	113
Mix	0.058	0.047	0.069	0.053	0.043	0.064	498
Gender > 0.5	0.055	0.038	0.071	0.040	0.023	0.056	440
Gender < 0.5	0.049	0.033	0.064	0.033	0.018	0.048	348
Developed country	0.045	0.036	0.054	0.039	0.031	0.048	1305
Developing country	0.055	0.042	0.069	0.048	0.035	0.062	253

Note: This table presents basic summary statistics of the partial correlation coefficient calculated on various subsets of the data. Unweighted = We use the original data set. Weighted = We weigh the estimates by the inverse number of estimates reported by each study. GPA = Grade Point Average, OLS = Ordinary Least Squares, diff-in-diff = Difference in Differences. For a detailed explanation of the variables, see table 5.1.

Chapter 4

Publication bias

To uncover the underlying patterns behind the behavior of the effect, we are going to focus in this chapter on detecting publication bias in our literature sample. In short, this bias represents a particular preference of researchers for statistical significance, as discussed by Ferriar (1792); Sterling (1959); Easterbrook et al. (1991); De Long & Lang (1992); Thornton & Lee (2000); Rothstein et al. (2005); Stanley (2005); Ioannidis & Trikalinos (2007); Stanley & Doucouliagos (2012). The simple count of the existing studies, out of which we enumerated but a few, should highlight to the reader why we consider it essential to search for this effect in our thesis. To explain a bit further, the publication bias captures the fact that the statistically less significant effects tend to appear in the literature less often than their counterparts. This preference arises mainly because more significant results are easier to publish, leading to a general overestimation of the reported effect. Another common name for this phenomenon is the 'file drawer problem,' as the less significant findings tend to be 'left in the drawer' and remain unpublished (Stanley 2005; Rothstein et al. 2005).

The latter of these two authors also suggests in the cited book that this bias is becoming more present in recent literature because of the higher use of systematic and quantitative methods. In that spirit, reviewing one's work, which is arguably tightly connected to writing a research paper, may be the cause that could lead the authors to prefer specific results. However, this behavior does not necessarily imply cheating but should rather be viewed as a byproduct of the research itself. When an outlier appears in the data, it is feasible for the author to disregard it on an individual level. What could be problematic is when this behavior occurs systematically, possibly leading to publication bias. To look at the publication bias in a more critical light, we would like to mention the work of Aguinis et al. (2011), who pointed out that some methods which test for publication bias may be misleading due to the small amount of information taken into account.

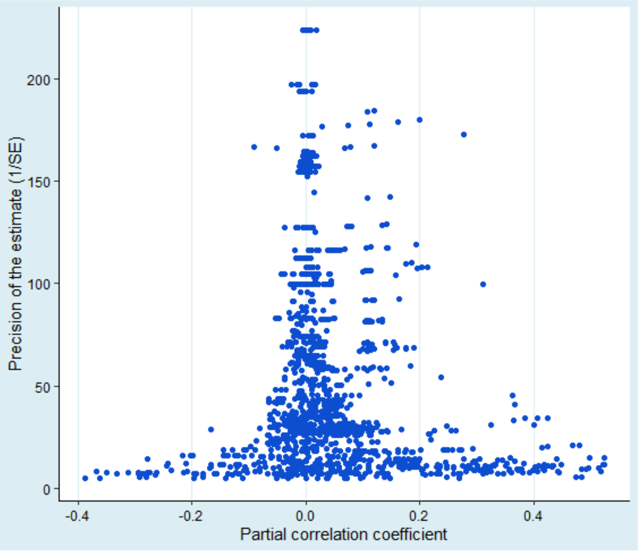
Having this theoretical background in mind, we can quickly look at what other researchers have found when conducting meta-analyses on a topic similar to ours. Out of the ten meta-analyses mentioned in Chapter 2, only Cameron & Pierce (1994); Cerasoli et al. (2014) explicitly mention the possible effects of publication bias on their work. Cerasoli et al. (2014) use the so-called 'File drawer analysis' to correct the bias, which implies their findings are not directly comparable to ours (Rosenthal & Rubin 1988). On the other hand, Cameron & Pierce (1994) use a very similar metric, providing (as far as we are aware) the only existing comparable measure of publication bias to our findings. We observed that they report little to no evidence of publication bias in their work. Considering the year when their study was published, it appears very interesting to look for this bias in the topic of rewards-motivation using a more recent and newer set of literature.

In our search for publication bias, we take inspiration from and closely follow the methodology of Zigraiova & Havránek (2016); Gechert et al. (2021), and Havránek et al. (2021), among others. Their approach to meta-analyses motivated us to employ various standard and relatively new procedures alike, all of which we further describe in the rest of this section.

Arguably one of the most popular and common methods meta-analytic methods for detecting a publication bias (with more than 33,000 citations on Google Scholar) is the funnel plot (Egger et al. 1997). Estimates of the effect (in our case, partial correlation coefficient) are plotted against the inverse of the standard error (precision) to create a simple visual representation of the effect's behavior. The higher the precision, the closer the estimates should be to the underlying effect. The more imprecise the estimates are, the more scattered they shall appear, creating an inverted funnel, hence the practice's name. In case of a publication bias, such a plot may appear asymmetrical on one side due to the omission of some estimates, which can happen when the authors leave out estimates that are in contrast with popular beliefs. In other cases, the plot will be hollow at certain parts due to the omission of insignificant effects. A combination of these two is also possible, all due to the authors' effort to make the study more attractive and easy to publish.

The funnel plot we obtained is shown in figure 4.1, suggesting no evident presence of publication bias. Apart from not appearing hollow even across different levels of precision, the graph seems more or less symmetrically centered around a value close to 0, which would suggest our results are in line with the findings of Cameron & Pierce (1994). However, one can hardly overlook the number of positive, highly precise estimates on the right side of the funnel graph. Considering this to be a trend rather than a fluke, we took a second look at the nature of the studies reporting these estimates, but found no visible irregularities in their structure. Thus, it is possible that either some hidden properties or situations might lead to an unexpected surge in performance or that a preference for statistically significant results influences these estimates' behavior.

Figure 4.1: Funnel plot (Egger et al., 1997)



Note: The figure displays a funnel plot as described by Egger et al. (1997). Such plot should be symmetrical in case of no publication bias. Winsorized outliers were hidden for better clarity of the effect but remained in the calculations.

4.1 Linear tests for publications bias

To search for potential publication bias more rigorously, we employ a number of both linear and non-linear statistical tests following the meta-analysis guidelines reported by Stanley et al. (2013) and further discussed by Havránek et al. (2020).

Shown to be excellent when it comes to publication bias detection (Stanley 2008; Moreno et al. 2009), we first conduct the Funnel Asymmetry Test (FAT)-

Precision Effect Test (PET). These aim to observe for potential correlation between the estimates and their standard errors by means of a simple regression. In theory, as mentioned by Stanley (2005; 2008), the estimates should be uncorrelated with their standard errors, for the opposite would suggest preference of some results over others, such as those with higher statistical significance. The following equation is thus estimated to test for this:

$$PCC_{ij} = \beta_0 + \beta_1 * (SE_{PCC})_{ij} + u_{ij}, \tag{4.1}$$

where PCC_{ij} denotes the i-th partial correlation coefficient with its standard error $(SE_{PCC})_{ij}$, observed in the j-th study. The intercept β_0 shows the 'true underlying value' of the effect, corrected for publication bias, β_1 captures the size of this bias, and u_{ij} represents disturbance. In the tables that follow and capture the calculation results, we shall refer to β_0 as the 'Effect beyond bias' and β_1 as the 'Publication bias.'

We employ several methods for estimating equation 4.1, and if not stated otherwise, we cluster the standard errors at the study level and assume exogeneity in the model. Apart from the usual OLS estimation, we account for potential heteroscedasticity in the sample by weighing the equation by the inverse of SE_{PCC}, as suggested by Ioannidis et al. (2017). In another estimation, we use the inverse of the number of estimations collected from each study, thus weighing each study the same and accounting for their different sizes. In the last two of these estimations from table 4.1, we use Fixed-Effect (FE) and Random-Effect (RE) models, respectively.

Table 4.1: Linear tests for publication bias

	OLS	**FE**	**BE**	**Study**	**Precision**
SE	0.319**	0.879***	0.627***	0.203	0.879***
Publication bias	(0.131)	(0.037)	(0.125)	(0.134)	(0.172)
Constant	0.032***	0.014***	0.020***	0.035***	0.014***
Effect beyond bias	(0.004)	(0.001)	(0.003)	(0.004)	(0.003)
Studies	44	44	44	44	44
Observations	1568	1568	1568	1568	1568

Note: The table displays the results obtained from estimating equation 4.1. OLS = Ordinary Least Squares. FE = Fixed Effects. BE = Between Effects. Precision = We weigh the estimates by the inverse of their standard error. Study = We weigh the estimates by the inverse of the number of observations reported per study. Standard errors, clustered at the study level, are included in parentheses. ***p<0.01, **p<0.05, *p<0.1

It is visible at first glance that the FAT-PET tests confirm the minuscule size of the underlying effect, which is significant at the 1% level across all five estimations. Another unique property is that the publication bias drops when weighing by the number of estimates per study, albeit to a non-significant value. This drop suggests that certain studies may be driving the publication bias. We can observe this, for example, in figure 4.1, where clustering around specific values may indicate the presence of this bias.

Looking further at table 4.1, four out of the five methods suggest a significant presence of publication bias, which again might be driven by the large number of positive, highly precise estimates reported by some studies. We could accredit this level of precision to the large number of observations used in each estimation. However, one should note that we collect a little over 35 observations per study on average, which is relatively more than meta-analyses usually report. Consequently, we might expect a few studies to drive up the effect considerably.

4.2 Non-linear tests for publications bias

While the previous tests provide a good baseline in search for publication bias, they assume a linear relationship between the PCC and its standard error. This assumption might lead to imprecise estimation if such a relationship is non-linear or includes kinks, for example. One should also be aware that the FAT-PET method tends to underestimate the 'true underlying effect' when it is different from zero (Stanley & Doucouliagos 2014; Bom & Rachinger 2019). These imperfections, among others, were the reason we decided to search for publication bias further using tests, which assume non-linearity. We present the results of these tests in table 4.2.

The first method we decided to use is the Weighted Average of Adequately Powered (WAAP) proposed by Ioannidis et al. (2017). They suggest using unrestricted Weighted Least Squares (WLS) only on estimates of those studies, which are adequately powered. The method tests this condition by comparing the calculated standard errors to a power threshold defined using statistical significance and adequate power. As explained further, WAAP is well suitable for estimating the publication bias size, as it does not require specification of numerous implicit properties regarding the bias. In our data set, we find a total of 331 estimates, which satisfy the assumption of adequate power.

Stanley et al. (2010) suggest a very straightforward approach in testing for publication bias, specifically discarding 90% of the data and leaving only the

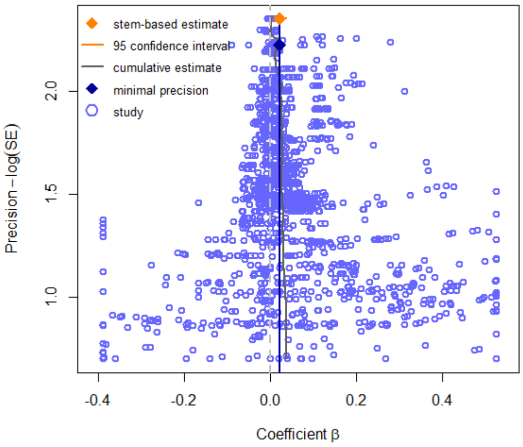


Figure 4.2: Stem-based method

Note: This figure shows a non-linear estimation of the underlying effect, according to Furukawa (2019). The orange diamond represents the stem-based estimate of the partial correlation coefficient, with the orange line corresponding to the 95% confidence interval. The dark gray line corresponds to estimates throughout various levels, and the dark blue diamond indicates the minimal precision above from which the model calculates the stem. Lastly, the purple circles correspond to individual estimates of the partial correlation coefficient.

remaining 10% with the highest precision in the sample (naming the method *Top10*). They consider that studies may be published based on the statistical significance of their reported effect. They further argue that most of the sample could be, because of this property, not representative of the effect. Thus, it should be preferable to leave out most of the data and look only at the most precise part. In our data set, that makes for a total of 157 observations.

Furukawa (2019) takes a similar approach to the Top10 method and again suggests using only the most precise estimates out of the sample - the 'stem,' so to say. One can identify these observations by using the equation

$$\underset{n}{min}MSE(\hat{b_0^n}|\sigma_0) = Var(\hat{b_0^n},\sigma_0) + Bias^2(\hat{b_0^n},b_0)\ subject\ to\ Var(b_i|\hat{b_0^n},P) = \sigma_0^2, \tag{4.2}$$

which is in detail described in Furukawa (2019). In short, it seeks to minimize the Mean Squared Error (MSE), which equals the sum of variance and bias given information about the mean estimate and the number of studies, among other variables. As variance increases with the number of studies included, bias decreases, and vice versa. Those observations with the lowest MSE are then used in the estimation, as mentioned above. Figure 4.2 shows the results of implementing this method on our sample.

We further test for publication bias using the Selection model proposed by Andrews & Kasy (2019). They suggest correcting the publication bias utilizing the so-called 'conditional publication probability,' which represents the probability of a study being published given the results found during the said study. They show that one can nonparametrically calculate this probability, allowing them to observe the behavior of the underlying effect in the data sample.

Another one of the non-linear tests we use is the Hierarchical Bayes model, which we constructed following the procedure of Allenby & Rossi (2006). The model uses Bayesian statistics and variation within studies to identify the weights of individual observations pooled at the study level. For further detail, we refer the reader to the original paper.

The last of the models we employ when testing for a non-linear relationship in the data is called the Endogenous Kink (EK) meta-regression model (Bom & Rachinger 2019). This approach identifies a kink at a specific cutoff value of the standard error, below which it would be highly improbable to find any publication bias. Having obtained this kink, Bom & Rachinger (2019) then propose fitting a piecewise linear regression of the collected estimates on their respective standard errors in to identify the underlying effect. The estimation results of all the non-linear methods can be found in table 4.2.

The non-linear tests further support the results we obtained from the FAT-PET tests. Four of theses non-linear methods suggest a minimal yet highly significant effect beyond bias, which would confirm the patterns in the behavior we observed up to this point. The Selection model procedure estimates pure zero when utilizing t-distribution at the 5% significance level, and the slightly higher estimate given by the Hierarchical Bayes model is an insignificant one. Finally, we can observe that the two models which report publication bias (Hierarchical Bayes & Endogenous kink) propose very similar results to the models from the previous section. As the other four models do not look for publication bias, we do not report this statistic for these models.

Table 4.2: Non-linear tests for publication bias

	Effect beyond bias		
WAAP	0.024*** (0.003)	0.000 (0.003)	**Selection model**
Top10	0.019*** (0.004)	0.049 (0.068)	**Hierarchical Bayes**
Stem-based method	0.021*** (0.007)	0.012*** (0.002)	**Endogenous kink**
	Publication bias		
Hierarchical Bayes	0.684 (0.677)	0.887*** (0.152)	**Endogenous kink**

Note: The table reports estimates of the effect beyond bias using six non-linear methods and estimates of the publication bias obtained using two of these methods. WAAP = Weighted Average of the Adequately Powered. Top10 = Top10 Method. Standard errors, clustered at the study level, are included in parentheses. ***p<0.01, **p<0.05, *p<0.1

4.3 Relaxing the exogeneity assumption in the tests for publication bias

To further establish robustness in our findings, we now proceed to relax the exogeneity assumption we have been holding until this point, which implied that standard errors were not correlated with the original effect if the publication bias was absent. Without this assumption, we can move onto testing for potential endogeneity of the standard error. The first two methods we are going to implement (see table 4.3 for results) are the Instrumental Variable (IV) regression and a method called p-uniform* (van Aert & van Assen 2020).

While constructing the IV regression, we choose the inverse of the number of studies' square root to act as an instrument for several reasons. Firstly, it is because of its correlation to the standard error, which comes from the standard error definition. Secondly, this IV displayed the best overall performance during various tests for its suitability, including the Underidentification test, Weak identification test, Stock-Yogo weak ID test, and while looking at the Sargan statistic. We decide not to go into the details about these tests, for they are but means of selecting the best IV in a regression. Nonetheless, the details and the results are accessible through the code, which we include in the online appendix.

The other method mentioned above - p*uniform - builds on the idea of the uniform and even distribution of p-values around the underlying effect value. It tests for this assumption by observing the distribution of p-values in the sample at various points and evaluates their distribution. Such distribution will appear uneven or quite commonly clustered around specific statistically significant values if publication bias exists in the sample (van Aert & van Assen 2020).

The estimates using these two methods are again in line with most previous models, displaying virtually zero effect of financial incentives on performance and a considerable publication bias. However, none of these estimates are significant at any conventional level, apart from the p-uniform* effect beyond bias (statistical significance at the 5% level). The results can be found below in table 4.3.

Table 4.3: Relaxing the exogeneity assumption

	IV	**p-uniform***
Publication bias	0.512	YES
	(1.121)	(2.017)
Effect beyond bias	0.023	0.021**
	(0.050)	(0.012)
Studies	44	44
Observations	1568	1568

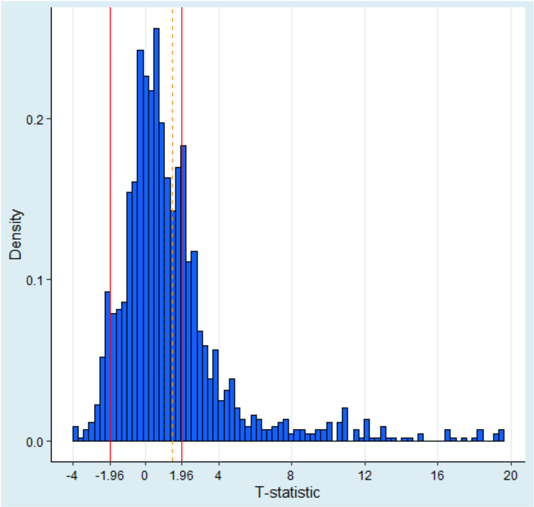
Note: IV = Instrumental Variable Regression; we used the logarithm of the number of studies as an instrument for the standard error. These standard errors, reported in the parentheses, are also clustered at the study level. For p-uniform*, we used the maximum likelihood estimation; the values in parentheses represent p-values. ***p<0.01, **p<0.05, *p<0.1

To finalize our testing for the publication bias, we look at a method developed by Gerber & Malhotra (2008) called the Caliper test. This approach does not assume any relationship between the effect and its standard error, contrary to the previous methods. Instead, it looks at the distribution of t-values obtained by computations using these two statistics. More specifically, Gerber & Malhotra (2008) suggest looking around specific statistically significant values on small enough intervals to detect potential jumps in the distribution. If any particular statistical value tends to be over-reported in the sample, a notable jump will then appear around that value.

From looking at the t-statistic distribution in figure 4.3, it seems clear that a few notable jumps in the number of t-values reported occur at values 1.96 and -1.96 (which correspond to 5% statistical significance). Considering this, we move on to confirm this suspicion empirically and observe the frequency of reported t-values reported around these two values on intervals specified in table 4.4. We can interpret the obtained values as the difference between the number of observations above and below the given threshold. A coefficient of 0.128, for example, would represent that 62.8% estimates are above the threshold, while 37.2% are below. We can obtain this coefficient by subtracting 0.5 from the number of estimates above said threshold (i.e., 0.628 - 0.5) and then comparing it to the original standard error to test for statistical significance.

The tests suggest a significant discrepancy between the number of observations reported on each side of the threshold 1.96. This difference appears even on an interval of 0.05, where we found a sufficient number of observations to support this claim. The tests around threshold -1.96 then suggest a weaker presence of said discrepancy. All in all, we can argue that a particular preference for results at the 5% significance level is present in our sample.

Figure 4.3: T-statistic distribution



Note: The figure displays the distribution of t-statistics of the reported estimates in our data set. The two red lines highlight the critical values 1.96 and -1.96, both at the 5% level of significance. At the same time, the orange dotted line represents the mean t-statistic value in the distribution. We hid the winsorized outliers in the figure for better clarity but included them in the calculations.

Table 4.4: Caliper tests at values 1.96 and -1.96

	Threshold = 1.96	**Threshold = 0**
Caliper width 0.05	0.128*** (0.036) N = 37	0.079** (0.038) N = 24
Caliper width 0.1	0.131*** (0.027) N = 62	0.125*** (0.037) N = 34
Caliper width 0.2	0.207*** (0.025) N = 96	0.183*** (0.032) N = 55

Note: The table shows the results of two sets of Caliper tests. Both of these sets are carried out around a different threshold, as denoted in the table. N = Number of observations found in each of the respective intervals. Standard errors, clustered at the study level, are included in parentheses. ***p<0.01, **p<0.05, *p<0.1

To summarize, the vast majority of the tests conducted in this chapter seem to tell the same story. These tests hint, first, at a small yet positive effect of rewards on motivation (0.012 - 0.035), and second, at the presence of a publication bias in the literature (0.319 - 0.887). Understandably, not all of these results are statistically significant. It is possibly more intriguing that no statistically significant results suggest the opposite, such as a negative effect beyond bias or a negligible publication bias. However, we still need to subject these findings to further testing, as they may be secretly correlated with hidden drivers we have not had the chance to specify. For this reason, we decided to focus the next chapter on uncovering the heterogeneity in the data set, which should allow us to check the robustness of the claims mentioned above.

Chapter 5

Heterogeneity

To better understand the potential drivers behind the rewards-motivation effect, we now look at heterogeneity within our data and studies. Namely, we discuss the variables captured in our data set, how their choice can influence the reported effect, and the pattern of their behavior the existing literature suggests. We further use these variables while applying Bayesian and frequentist model averaging methods to account for model uncertainty and provide an overall robustness check.

5.1 Coding the variables

We first observe and compare the various characteristics within studies. For us to do so, we consider the various specifications of the primary studies and choose our data set variables to capture the behavior of the underlying effect economically. The final selection contains 49 different study characteristics, shown in table 5.1, along with their definitions and summary statistics. To avoid the dummy variable trap, we only use 34 of these later in the model averaging. The variables are also categorized into groups to capture the following specifications of the effect: effect characteristics, methodology, study specifications, reward scheme, task nature, motivation, along with subject and country characteristics. We now detail the reasoning behind the choice of our variables, which we put into direct contrast with the primary studies' approach.

Effect characteristics In terms of specifying the underlying effect itself, our approach varies significantly from the one taken by most of the primary meta-analyses. Eight of the studies focus on the undermining effect and, consequently, intrinsic motivation as their dependent variable (see Chapter 2 for more). On the contrary, Jenkins et al. (1998); Cerasoli et al. (2014), and Van Iddekinge et al. (2018) choose performance as their variable of interest. Of particular interest to us is the study by Cerasoli et al. (2014), which points out that most of the other works operate virtually only under laboratory settings. This setting then makes it challenging for these works to consider how the effect behaves in the field (e.g., in schools, work environment, among others).

When looking at the sample of studies that comprise our data set, a similar, substantial heterogeneity is immediately noticeable. To give but a few examples, Sliwka & Werner (2017) try to observe whether varying wages have an effect on the subjects' speed in counting blocks of numbers during a laboratory setting experiment. Kremer et al. (2009) on the other hand employ a large scale experiment in Kenyan schools to see whether money can improve students' performance during academic exams. Karlan & List (2007) then test for charitable giving by means of mail solicitation in order to uncover the effects of money on altruistic behavior. Having yet a different goal in mind, Fehr & Goette (2007) simply pay bicycle messengers to see whether their delivery numbers will improve.

With this kind of variety in the data, it seems unfeasible to simply lump all of the effects into one category. However, we managed to observe a clear underlying pattern regarding the kind of studies, which allowed us to create a simple variable categorizing them according to their nature. Namely, we created four of these categories, which capture in order: students' GPA, charitable giving, an outcome of a game or a simulation, performance of employees at work. We found that virtually all of the studies collected in our data fitted into one of these four categories, making this setup appear very suitable. By this approach, we also try to remedy one heavily criticized flaw of the PCC: the lumping together of effects, which are seemingly totally different. For example, Glass et al. (1981) point at this and argue, that conclusions drawn when generalizing like this are invalid. However, we believe that our thesis chooses a middle ground in the trade-off between incomparability and excessive generalization and provides an interesting insight into inner workings between effects of different nature.

Besides this critical feature, we decided to code one technical variable regarding the effect characteristics. We call it 'Effect positive,' and in short, it is a dummy capturing whether a higher effect is desirable or not, as we explained in Chapter 3. If we were to justify us choosing this variable, it lets us unify the effect's direction. Furthermore, it shows whether the tasks measuring quantity/quality of the outcome are more frequent than those measuring a decrease in time, GPA, or similar outcomes. An exemplary setup of the positive effect appears in Dohmen & Falk (2011), where more numbers multiplied during a task means a better performance. As for the negative effect, a great illustration is the experiment conducted by Fershtman & Gneezy (2011), where the subjects try to run a 60-meter race in the fastest time possible.

Methodology Regarding methodology, we found a clear pattern in our data. For most cases (87%), the researchers used regression to capture the effect, leading to the seven most common methods captured in the 'method' variable. Ordinary Least Squares (OLS) is undoubtedly the most common and gets used either alone (such as in Nagin et al. (2002); Boyer et al. (2016) or Dwenger et al. (2016)) or combined with other methods (such as Angrist & Lavy (2009), who estimate their models using Logit as well as OLS). If the regression methods that appeared only in 1-2 studies, we classified the case as using the 'Other method,' such as Celhay et al. (2019) or Angrist & Lavy (2009), who used two stage least squares. The remaining cases reported the mean number of the effect, usually denoting the subjects' change in performance, such as when Fryer Jr (2011) observe the number of books read by a student in a given time frame. Table 5.1 then lists all of these methods for clarity.

When it comes to the primary studies' methodological approach, they focus more on their own approach, paying little to no attention to the methods employed in the studies from which they gather their data. In this aspect, our thesis provides a little more insight into our results' origin, albeit of possibly lower importance than other variables.

Study specifications As for characterizing the studies, it is interesting to note the classification of field/lab setting used by Jenkins et al. (1998), who also point out in one of their previous works (Jenkins 1986) that the laboratory setting may yield more substantial effects than its counterpart. Considering other specifications, Cameron & Pierce (1994), for example, include a variable capturing whether the researcher was present at the experiment. However, this specification is again very similar to the field/lab variable.

Apart from the usual variables mentioned in Chapter 3, one other particularly deserves a bit of explanation. We initially distinguished between lab and field setup and between an experiment and a study, creating four categories of the 'Lab/Field study' variable. Nonetheless, after calculating the summary statistics, we found that very few of the experiments we observe are either field or lab studies, so we simplified this variable. The laboratory experiments are in the vast majority of cases conducted in an artificial setting, such as in Gallier et al. (2017); Bradler et al. (2019) or Sliwka & Werner (2017). The field experiments then vary a bit, such as when a blood donation is measured in Lacetera et al. (2012) compared to a when Kirchler & Palan (2018) observe a change in the size of food obtained from a worker after a compliment. However, it still appeared the most optimal to just split the categories into lab/field, as that is where the biggest distinction occurs.

This section also includes one interesting variable, namely 'Crowding-out,' which is equal to one when the study explicitly mentions or utilizes the crowding-out intrinsic motivation theory and was included to capture the researchers' awareness about the theory behind motivation when setting up their experiment. Fehr & Schmidt (2007); Charness & Gneezy (2009) or Homonoff (2018) make for a model example of the awareness about this theory.

Reward scheme If we look at the primary studies, they mainly distinguish between three major reward categories. These are reward type (tangible or verbal), reward expectancy (if one expects the reward or not), and reward contingency (such as whether the subjects receive the reward for simply completing the task, completing it well, maybe during a specific time frame, or given other specifications). We mentioned in Chapter 2 that some authors (Cameron & Pierce 1994; Tang & Hall 1995; Deci et al. 1999) suggest a simple task-contingency to have a detrimental effect on the motivation, while verbal and unexpected rewards should do the opposite. Such classification, precisely like this or in part, is chosen by 9 out of the 11 primary meta-analyses. The only two studies that opt for a different classification are Cerasoli et al. (2014) and Van Iddekinge et al. (2018). In the latter of these studies, the authors choose to focus more on the motivation as something that already exists, or rather something that is but a means of predicting performance. With this approach, they do not put much weight on the origin of the motivation. More interesting reasoning behind reward scheme choice (in regards to our approach) appears in Cerasoli et al. (2014). They see the usual 'contingency continuum' as unfit for their work because it considers a controlled, laboratory environment. If one wants to observe a wide variety of experiments, it should be suitable to choose a different scheme instead. Their study, for example, discerns between different levels of reward salience, by which it hopes to explain the relationship between incentives and performance better.

When defining the reward scheme for our meta-analysis, it thus seemed only fitting to choose a setup more inspired by Cerasoli et al. (2014), given the wide range of the effect we capture. However, we decided to step into a more economical direction and designed the primary reward variable followingly. We first denoted or calculated the treatment group subjects' average earning. Furthermore, we also gathered information about the monthly median household expenditure for each of the necessary countries. Sometimes, we only managed to obtain the latter statistic in yearly intervals, so in that case, we took the statistic for the year in which the study occurred and divided it by 12. If the study took place over several years, we used the mid-year and obtained the information for that year, such as in the case of Kremer et al. (2009) or Lacetera et al. (2012). Using this data, we then divided the logarithm of the treatment group average earning by the logarithm of the median monthly expenditure. We then capture this new measure of payoff in the variable 'Reward scaled.' This approach allowed us to economically quantify the importance of the reward for the subjects, which we value more than the usual psychological approach. On the other hand, one must note that this puts virtually all of the rewards our subjects receive into the category 'tangible,' where the researchers predict a strong 'undermining motivation' effect.

Looking at further specifications in this category, we also found noticeable heterogeneity in how the subjects received the reward, which we codify in two variables. 'All paid' indicates that all subjects participating in the experiment received a reward, making it possible to observe the reward contingency and expectancy. Paying all subjects implies that they are aware of a guaranteed future payment, possibly altering their behavior accordingly, as noted in Greene (2018). Very often, the subjects are guaranteed to receive some kind of pay if they voluntarily participate in a laboratory experiment, such as in the case of Cappelen et al. (2017) or Bradler et al. (2019). Not always receiving a reward is very typical for a school-based setting, where only the best students are awarded a scholarship (Angrist & Lavy 2009; Li et al. 2014). On the other hand, some experiments award the students right after the examination (Levitt et al. 2016), which blurs the lines between this distinction a bit. Apart from the above mentioned scholarship, we also chose this variable to control for lotteries, which appear throughout our sample as a reward (Gallier et al. 2017). Here, a sizeable one-off reward gets compensated with the low probability of receiving it, making the reward schemes more comparable.

The other variable taking care of reward heterogeneity is ’Reward own,’ which equals one if the subjects receive the payoff for themselves. A nice example of a case when the reward is given to someone else than the experiment subject appears in Mellström & Johannesson (2008), where subjects have the choice to donate their payoff (earned by participating in blood donation) to charity. We chose this design with the goal of further accounting for altruistic behavior during the experiment.

Task nature If we take a quick look at the task specifics among the primary studies, the task appeal has been among the main focus of Tang & Hall (1995), who codify among their five primary variables one which they label ’Interest level.’ Similarly to the purpose of our ’appealing’ variable, it serves to specify tasks, which are of interest to the subjects. Jenkins et al. (1998), along with Rummel & Feinberg (1988); Deci et al. (1999); Cameron (2001), choose a similar scheme, distinguishing between extrinsic and boring tasks. A completely new outlook is then proposed by Cerasoli et al. (2014), as they argue that most of the previously mentioned analyses study inherently exciting tasks. According to Cerasoli et al. (2014), numerous field tasks, such as work in an organization or school attendance, are not necessarily appealing to the subjects. When it comes to task performance, theory predicts a stronger relationship between intrinsic motivation and qualitative tasks. On the other hand, this relationship should appear weaker between this motivation and quantitative tasks (Kruglanski et al. 1971; Evans 1979). Such property is, for example, attributed by Deci & Ryan (2000) to the effect of the performance nature on one’s self-determination.

As for our setting, we take inspiration from this theory by distinguishing between quantitative/qualitative performance, cognitive/manual tasks, and, lastly, appealing/non-appealing tasks. With a quick comparison to the primary studies, this setting closely resembles the one chosen by Cerasoli et al. (2014). Most of these variables seem self-explanatory, but we would like to justify the choice of distinguishing between cognitive and manual tasks. We chose this design to further capture small nuances in the nature of some of the experiments, which the other variables were not sufficient enough for. One such specification which the cognitive/manual setup allows is to further classify the subjects’ performance during the lab experiments. We can clearly distinguish between cognitive tasks such as solving puzzles and manual tasks such as clicking on circles, both of which were featured in Takahashi et al. (2013).

Furthermore, using the cognitive/manual distinction, we can also categorize the work in the 'employee' group by nature, such as when employees took part in a cognitive laboratory experiment (De Quidt 2018) versus when they were observed working in a factory (Lazear 2000).

With this setup, we initially expected to find either large values of Variance Inflation Factor (VIF) or correlation between the 'Cognitive task' variable and the 'Effect GPA,' where virtually all students should partake in cognitive tasks. However, neither of these suspicions were confirmed after a closer inspection of the model. We went through the data set and found that the reason for this is the presence of several studies, where the subjects engage in manual tasks. For instance, students were paid in Charness & Gneezy (2009) for gym attendance, in Fershtman & Gneezy (2011) for running a 60-meter race and in Conrads et al. (2016) for attending a conference as voluntary helpers. This fact allowed us to keep in the model the specification mentioned above.

Motivation Among the primary studies which observe the undermining effect, intrinsic motivation is the main focus. However, such motivation may prove to be quite challenging to measure. One of the standard methods, used, for example, by Cameron & Pierce (1994), measures this motivation as free time on task after one stops providing the subject with rewards, along with reported self-interest in the task and the willingness to participate in it without any reward. This reported self-interest is also the primary source of computations to the primary studies that employ Cohen's d (which we have discussed in detail in Chapter 2). The same method appears in the papers of Wiersma (1992) or Cameron (2001). An intriguing variable that might fit into this category, but we decided not to code, would be to distinguish whether extrinsic incentives (i.e., rewards) were present or absent (Cerasoli et al. 2014). This approach would allow us to control for pure non-monetary motivation during specific experiments. Unfortunately, we found it impossible to implement in our sample due to the fact that virtually all studies include a group which received a monetary reward. At most, a middle ground between the type of reward received appears in some studies such as Kirchler & Palan (2018), where only a compliment during a food order is sometimes given as an extra reward by the researcher. However, the subjects are still paid for the service provided, making the distinction unclear and likely impossible.

In understanding the motivation of the subjects behind their behavior, we decided to try a slightly different tactic. To be specific, we classify the experiments into five possible scenarios, in which the subjects are motivated by the following driving effects: altruism, trust, reciprocity, fairness, or purely money. This setup gives us an idea of whether the performance change is induced purely by monetary incentives or may have an underlying driving force. Ariely et al. (2009), for example, designs his experiment to observe whether money has an influence on the interaction between the subjects' pro-social behavior and their internal perspective on the task induced by these rewards. Another reasoning for the approach we chose is that while it completely disregards the standard measure of intrinsic motivation, the economic aspect of motivation gets highlighted instead, which is preferable for our thesis. We also note similarities between some of these variables and ones from different categories (such as 'Effect Charity' and 'Altruism' variables). However, the tests for correlation and VIF suggest that such classification carries some new, hidden information, making this setup appear suitable. Such distinctions appear, for instance, in Konow (2010) or Gallier et al. (2017). In both of these studies the subjects take part in a dictator game (which classifies the effect as 'Game') but have the option to transfer their endowment to charity instead of the recipient, giving ground to altruistic behavior. This clear distinction serves as new information to the model and can be put into direct contrast with, for example, the purely altruistic charitable setting, such as in Karlan & List (2007) or Mellström & Johannesson (2008).

Subject and country characteristics Among primary studies, Tang & Hall (1995), for example, only look at the range of subjects between preschool and college. Nonetheless, they argue for the existence of cognitive differences between how the subjects across and among these groups react to incentives, which corresponds to how age shifts one's perception of a fixed sum of money. Jenkins et al. (1998) classify the subjects into high-school, undergraduate, and graduate students, and lastly, employees. A very similar categorization appears in Cerasoli et al. (2014), where the authors distinguish between four categories: *Child*, *Adolescent*, *College*, and *Adult*.

We took inspiration from the already existing classifications and planned to categorize students into similar groups. However, with the initial categorization, which involved separating students into groups spanning preschool to middle school, high school, and college, the approach suggested a very high VIF in the model averaging between other explanatory variables. We suppose this problem might have arisen from the similar framing of experiments in which students took part. Angrist & Lavy (2009) observing high school students' exam performance in Israel is virtually the same as the experiment conducted by Kremer et al. (2009), where Kenyan elementary school students' exam performance was measured. For this reason, we decided to merge these variables into one, which allowed us to keep the information under the 'Student' variable. The remaining two categories of this variable are called 'Employees' and 'Mix,' which we deem to be the right approach relative to other studies.

In the 'Gender' and 'Mid age' variable, we found many missing values, as a number of studies did not report these characteristics (such as Lacetera et al. (2012); Dwenger et al. (2016) or Dohmen & Falk (2011)). Due to the necessity of having no missing data while performing model averaging, we set the ratio of male/female for missing observations to 50:50. For the mid-age, we filled in the mean of the respective group for each study (i.e., students/employees/mix). Lastly, we created a variable controlling for whether the country where the experiment took place is developed or not. An interesting observation is that in the developing countries, a notable portion of the experiments consisted of measuring students' performance (Kremer et al. 2009; Duflo et al. 2012; Li et al. 2014). No larger correlation or VIF however appeared in the model as a consequence, so it appears this is not always the case. Although this specification allows us to create one more economical specification for our analysis, we have no means of checking its theoretical validity, which happens because no other primary studies implement this variable in their approach.

Table 5.1: Definition and summary statistics of regression variables

Variable	Description	Mean	SD
PCC	Partial correlation coefficient (response variable)	0.044	0.156
Standard error	The standard error of the partial correlation coefficient	0.045	0.046
Effect characteristics			
Effect GPA	=1 if observed effect captured students' performance	0.348	0.476
Effect Charity	=1 if observed effect captured charitable giving	0.276	0.447
Effect Game	=1 if observed effect captured the outcome of a game	0.272	0.445
Effect Work	=1 if observed effect captured performance of workers	0.104	0.305
Positive effect	=1 if the relationship between rewards and the outcome is positive	0.866	0.341
Negative effect	=1 if the relationship between rewards and the outcome is negative	0.134	0.341
Methodology			
OLS	=1 if the authors use Ordinary Least Squares	0.558	0.497
Logit	=1 if the authors use Logit regression	0.063	0.243
Probit	=1 if the authors use Probit regression	0.085	0.279
Tobit	=1 if the authors use Tobit regression	0.034	0.181
Fixed-effects	=1 if the authors use Fixed-effects estimation	0.037	0.188
Random-effects	=1 if the authors use Random-effects estimation	0.027	0.161
Diff-in-diff	=1 if the authors use Difference-in-differences estimation	0.030	0.171
Other method	=1 if the authors use a different method	0.036	0.187

Continued on next page

Table 5.1: Definition and summary statistics of regression variables (continued)

Variable	Description	Mean	SD
Study specifications			
Cross-sectional data	=1 if the data is Cross-sectional	0.454	0.498
Panel data	=1 if the data is Panel	0.546	0.498
Time horizon	The logarithm of the number of days over which the experiment was carried out	4.096	2.667
Average Year	The logarithm of the average year of the experiment's timespan	7.605	0.002
N. of obs.	The logarithm of the number of observations used	7.084	1.960
Lab study	=1 if the experiment took place in a lab	0.224	0.417
Field study	=1 if the experiment took place in a field	0.776	0.417
Journal impact	The logarithm of the journal impact factor from RePEc	5.491	3.201
Study citations	The logarithm of the number of citations the study received	4.876	1.773
Crowding-out	=1 if crowding-out intrinsic motivation theory appears in the study	0.476	0.500
Reward scheme			
Positive framing	=1 if the study rewards its subjects	0.827	0.379
Negative framing	=1 if the study punishes its subjects	0.173	0.379
Reward scaled	The logarithm of the average payoff from the experiment divided by the logarithm of the median monthly expenditure in the corresponding country	0.610	0.299
All paid	=1 if all subjects received a reward (or punishment), =0 if only some received it	0.735	0.441
Reward own	=1 if the subjects received the reward for themselves	0.811	0.391
Reward else	=1 if someone other than the subjects received the reward	0.189	0.391
Task nature			
Quan. performance	=1 if the measured performance was quantitative	0.694	0.461
Qual. performance	=1 if the measured performance was qualitative	0.306	0.461
Cognitive task	=1 if the task involved cognitive work	0.701	0.458
Manual task	=1 if the task involved manual work	0.299	0.458
Appealing task	=1 if the task is appealing	0.479	0.500
Non-appealing task	=1 if the task is not appealing	0.521	0.500
Motivation			
Altruism	=1 if the subjects were motivated by altruism	0.279	0.449
Trust	=1 if the subjects were motivated by trust	0.020	0.140
Reciprocity	=1 if the subjects were motivated by reciprocity	0.098	0.298
Fairness	=1 if the subjects were motivated by fairness	0.156	0.363
Monetary	=1 if the subjects were motivated purely by money	0.447	0.497
Subject and country characteristics			
Students	=1 if the subjects were students	0.607	0.489
Employees	=1 if the subjects were employees	0.079	0.269
Mix	=1 if the subjects were a mix of these two	0.314	0.464
Gender	The logarithm of the ratio of male to female subjects (1 = all male, 0 = all female)	0.528	0.228
Mid age	The logarithm of the average year of the subjects	2.932	0.317
Developed country	=1 if the corresponding country is developed	0.833	0.369
Developing country	=1 if the corresponding country is developing	0.167	0.369

Note: This table presents the summary statistics and descriptions for each of the various study characteristics. SD = standard deviation, GPA = grade point average.

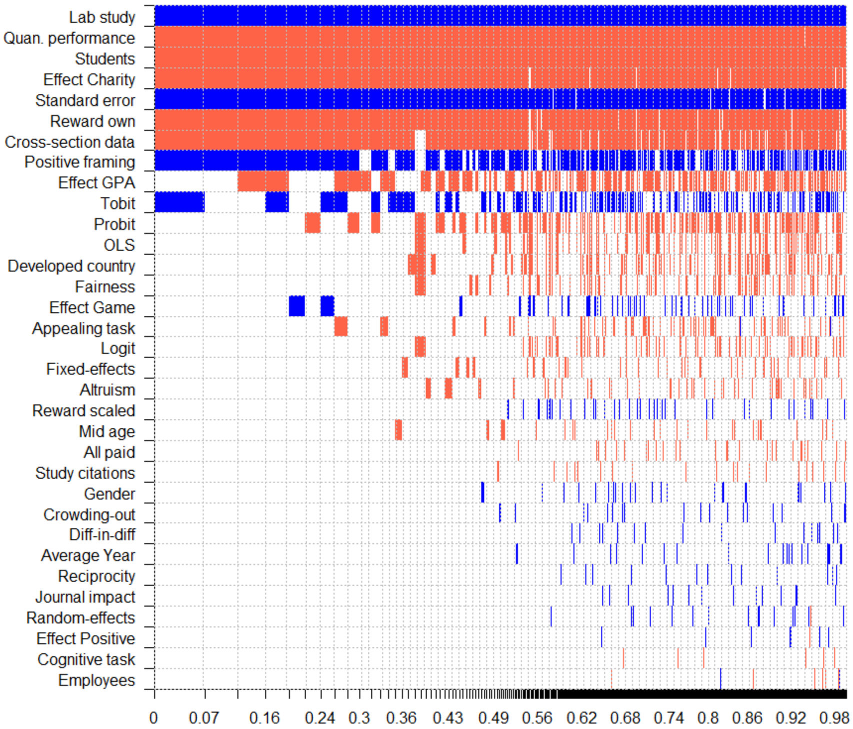
5.2 Model averaging

Having obtained and explained the final data set form and variable setup, we move onto tests that look for heterogeneity more rigorously. One idea is to use simple OLS to observe the effect of the explanatory variables on the PCC. However, this approach would surely suffer from over-specification bias, considering the number of variables we employ. Following the example of Cazachevici et al. (2020), we opt to make use of Bayesian Model Averaging (BMA).

As the name suggests, the BMA approach lies in averaging over many statistically feasible models, where there is particular uncertainty as to which one is the most fitting for estimation. Each model is then assigned a weight called posterior model probability. With it, the method calculates for every variable the probability of being included in the model (posterior inclusion probability), highlighting the importance of every one of these variables (more in Raftery et al. (1997); Hoeting et al. (1999); Amini & Parmeter (2011)). Like Cazachevici et al. (2020), we run the analysis using the *bms* package in R (Zeugner & Feldkircher 2009) using the Markov chain Monte Carlo algorithm. With this algorithm, one can cut a considerable number of models from the analysis without losing information. However, we choose a slightly different setting than could be considered usual. Namely, we decide to use a dilution prior rather than a uniform model prior. The dilution prior (George 2010) serves mainly to control for potential collinearity in the model, which it tries to remedy by multiplying the model probabilities and the determinant of the correlation matrix of the independent variables. This procedure gives larger weights when the correlation between the variables is small, as the determinant will then be nearing 1. In the case of a considerable correlation, the weights will shrink thanks to the multiplication. The main reason we opt for this approach instead of the uniform model prior is the number of similar variables, which may bring collinearity into our model. Another reason is simply the large number of variables we use, which may cause similar problems.

Before doing the actual BMA procedure, we first checked for correlation among our model variables and their VIF. We set up a model using all of the variables from table 5.1 and found the following two problems. A correlation was induced between the number of observation and Standard Error (SE) by definition, which happened because the number of observations is a factor used when calculating SE. We thus removed the number of observations variable, as we prioritize keeping SE in the model. Another conflict arose between the dummy classifying data as cross-sectional/panel and the variable indicating the time horizon. As all of the cross-sectional data span only one day, this is only natural, and one variable implicitly carries the other's meaning. When we looked at the VIF for both of these variables, the cross-sectional/panel dummy

Figure 5.1: Bayesian model averaging results



Note: This figure displays the results of the Bayesian model averaging using the uniform g-prior and the dilution prior. The response variable is the partial correlation coefficient, measured on the horizontal axis in terms of cumulative posterior model probabilities. The explanatory variables are ranked according to their posterior inclusion probability in descending order on the vertical axis. Blue color (dark in grayscale): the variable is included in the model and has a positive sign. Red color (light in grayscale): the variable is included in the model and has a negative sign. Numerical results of the estimation can be found in table 5.2. For a detailed explanation of the variables, see table 5.1.

displayed lower numbers, indicating a better implicit explanation of other variables in the model, so we decided to cut the time horizon variable. The reader can view the exact VIF numbers for each variable in the code.

We then checked the model one last time after taking these measures. From this, we observed that the 'Trust variable' is very noisy for the low number of observations, which relate to it (24 to be exact, which equals roughly 2% of all the observations). In terms of persevering the model's integrity, we found it best to remove this variable altogether, as it produced very inconsistent results. With this done, we finally move onto the actual estimation. The model averaging results are displayed graphically in figure 5.1, along with the numerical results and a Frequentist Model Averaging (FMA) robustness check in table 5.2.

Table 5.2: Model averaging results

Response variable:	Bayesian model averaging			Frequentist model averaging		
Partial Correlation Coefficient	Post. mean	Post. SD	PIP	Coef.	SE	p-value
Constant	-0.337	NA	**1.000**	18.832	27.093	0.487
Standard error	0.439	0.119	**0.987**	0.518	0.132	0.000
Effect characteristics						
Effect GPA	-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
Positive effect	0.000	0.001	0.011	-0.003	0.013	0.816
Methodology						
OLS	-0.005	0.012	0.170	-0.030	0.013	0.022
Logit	-0.007	0.020	0.120	-0.059	0.022	0.009
Probit	-0.015	0.024	0.340	-0.048	0.018	0.008
Tobit	0.027	0.033	0.446	0.034	0.024	0.167
Fixed-effects	-0.003	0.012	0.076	0.027	0.028	0.338
Random-effects	0.000	0.004	0.014	0.008	0.022	0.719
Diff-in-diff	0.001	0.006	0.024	0.060	0.032	0.067
Study specifications						
Cross-sectional data	-0.059	0.021	**0.935**	-0.046	0.017	0.007
Average Year	0.068	0.550	0.022	-2.449	3.559	0.491
Lab study	0.081	0.013	**0.999**	0.100	0.020	0.000
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
Crowding-out	0.000	0.002	0.026	0.008	0.010	0.445
Reward scheme						
Positive framing	0.038	0.024	**0.776**	0.036	0.019	0.068
Reward scaled	0.002	0.010	0.062	0.011	0.023	0.628
All paid	-0.001	0.005	0.038	-0.052	0.014	0.000
Reward own	-0.048	0.014	**0.978**	-0.085	0.019	0.000
Task nature						
Quan. performance	-0.059	0.012	**0.998**	-0.043	0.014	0.002
Cognitive task	-0.000	0.001	0.010	-0.000	0.010	0.962
Appealing task	-0.003	0.010	0.123	-0.036	0.014	0.011
Motivation						
Altruism	-0.001	0.006	0.076	-0.034	0.015	0.023
Reciprocity	0.000	0.003	0.017	-0.003	0.016	0.814
Fairness	-0.004	0.011	0.138	-0.042	0.015	0.005
Subject and country characteristics						
Students	-0.065	0.014	**0.998**	-0.055	0.015	0.000
Employees	-0.000	0.002	0.009	0.004	0.017	0.798
Gender	0.001	0.005	0.035	0.011	0.016	0.500
Mid age	-0.001	0.007	0.056	0.023	0.022	0.307
Developed country	-0.005	0.012	0.169	-0.036	0.015	0.016

Note: This table presents the results of the Bayesian and Frequentist model averaging. Post. mean = Posterior Mean, Post. SD = Posterior Standard Deviation, PIP = Posterior Inclusion Probability, Coef. = Coefficient, SE = Standard Error, GPA = Grade Point Average, OLS = Ordinary Least Squares, diff-in-diff = Difference in Differences. The variables with PIP > 0.5 are highlighted. For a detailed explanation of the variables, see table 5.1.

Besides the direction, size, and inherently the effect significance for each variable, we also present the Posterior Inclusion Probability (PIP), which appears briefly in the theoretical overview above. We would like to quickly remind the reader that it is a sum of all posterior model probabilities the models that include this variable and mention that it is analogous to statistical significance (for more details, see Steel (2020)). In other words, PIP shows the importance of the variable to the average model and how likely it is going to appear in the final model. The higher the number, the higher the importance. In interpreting the importance of each variable given its PIP, we refer to Kass & Raftery (1995), who suggest that values of PIP between 0.5 and 0.75 indicate weak evidence of the effect, values between 0.75 and 0.9 suggest a positive effect, values between 0.9 and 0.99 imply a strong effect and values over 0.99 represent a decisive effect. In table 5.2, we decided to highlight the variables with a PIP of 0.5 or higher.

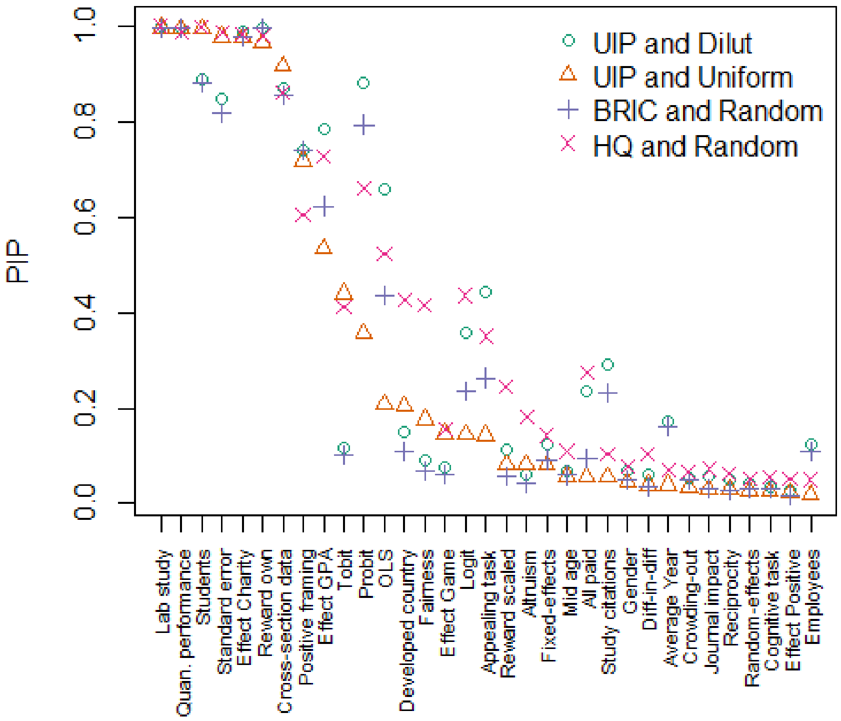
On the right-hand side of table 5.2, we then present a robustness check to these utilizing Frequentist Model Averaging. Similar to Havránek et al. (2021) and Gechert et al. (2021), we use Mallow's criteria as weights (Hansen 2007) and employ the orthogonalization of the covariate space, as suggested by Amini & Parmeter (2012). We do this because the previously used Markov chain Monte Carlo algorithm for reducing the number of models in the computation is not applicable here. Consequently, we can compare our previous results to the ones obtained from this new approach to model averaging and explain the existing heterogeneity a bit further.

For the initial BMA, the results suggest ten significant variables - effect GPA and effect charity, cross-sectional data, lab study, positive framing, reward own, quantitative performance and students. Furthermore, both the constant and the standard error are found significant, representing the underlying effect and publication bias, respectively. However, we cannot draw decisive conclusions about the underlying effect's behavior due to the missing posterior standard deviation. Conversely, we can observe the value of the publication bias variable, which suggests a similar story to the results we obtained in Chapter 4 (i.e., a statistically significant a publication bias in the literature). Looking at the other variables, we would first like to touch upon two of the presented outcomes, namely the highly positive and highly negative coefficients of the variables 'Lab study' and 'Quantitative performance,' respectively (0.081 & -0.059). It is interesting to note that the first finding is in line with Jenkins (1986), and their suggestions. They predict a more substantial effect of rewards on motivation in a laboratory setting, while the second outcome is then in line with the undermining effect and the theory suggested by Kruglanski et al. (1971) and Evans (1979).

To summarize the rest of the noticeable results, we can claim that an undermining effect exists in experiments conducted on school performance and charitable giving (-0.065 & -0.052). Perhaps one could attribute this to the association of money with the subjects' diminishing self-determination during the task, as proposed by Deci & Ryan (1985). In other words, the introduction of monetary incentives may decrease the initial drive the subjects feel towards the activity when doing it for free or out of their own will. Looking further, the experiments featuring cross-sectional data seem to be producing worse outcomes than their panel data counterparts (-0.059). Additionally, if the subjects receive the reward for themselves, the effect appears slightly weaker than when someone else gets that reward (-0.048). The last of the highlighted variables, 'Positive framing,' then indicates that a reward might be more valuable for increasing performance than punishment (0.038), although this coefficient is slightly less statistically significant.

As for the robustness checks, the outcome of the FMA appears to be more or less in line with the BMA approach, presenting a similar direction/significance of the effect in virtually all of the highlighted variables, except for the constant, to which the FMA attributes but a small significance coefficient. Furthermore, we

Figure 5.2: Variables and their inclusion in the model averaging



Note: This figure displays all of the Bayesian model averaging variables plotted against their posterior inclusion probability. PIP = Posterior Inclusion Probability, UIP = Uniform g-prior, Dilut = Dilution Prior, Uniform = Uniform Model Prior, BRIC = Benchmark g-prior, Random = Random Model Prior, HQ = Hannan-Quinn Criterion. For a detailed explanation of the variables, see table 5.1.

confirm our results' robustness by considering various specifications of the BMA model, namely different g-priors and model priors. These alternatives, along with their respective results in graphical form, are included in the Appendix B. Figure 5.2 then graphs the posterior inclusion probabilities of individual variables for all four of these models.

Chapter 6

The best-practice estimate

The last thing we would like to consider in this thesis is a method of estimating the best-practice effect from the BMA model we have obtained in the previous chapter. This method involves taking (for each variable) the coefficients of the said model and plugging in the characteristics which capture the best possible outcome of the effect. Nevertheless, one must keep in mind that this procedure is strictly subjective and should serve more as an extra robustness check than a presentation of new results.

6.1 Modelling the best-practice

When constructing the subjective best-practice estimate, we set the value of variables to the sample mean in most cases. We decided to choose this rather conservative approach mainly because the direction of the underlying effect is quite unclear. Therefore, it is difficult to argue whether increasing one variable (rewards, for example) is desired or not. On the other hand, we were able to identify several variables where the relationship is clear and so we decided to set their values as follows. The standard error should be equal to zero because having publication bias in the sample is not desirable. We are also looking for panel data, rather than cross-sectional, as that way we can retain more information. We then take the year of the most recently published study considering that this should reflect the current practice in the literature the best. Similarly, the number of citations and journal impact factor should also be at their highest values as that suggests higher credibility of the estimates. The rest of the variables are then set to their sample means, as we mentioned previously.

We then compare our set up to three actual studies from our data set, for which we plug the actual values of the respective variables into the BMA model to obtain the best-practice estimate for each of these studies. The only exception is the variable representing publication bias, which we set to zero in order to get unbiased results. As for the studies themselves, the Lazear (2000) study represents a study with the highest amassed number of citations among all other studies. On a similar note, the study by Angrist & Lavy (2009) features the reward scheme with the highest possible payoff for the subjects compared to other studies in the set. Lastly, we chose the study by Takahashi et al. (2013), which we found to be the most representative of the whole data set in terms of the experiment setup. The setup is a straightforward game, where the motivation/performance is directly and easily comparable with the payoff scheme, which we find highly fitting for our topic.

Table 6.1 then presents the results of this estimation, displaying the best-practice estimate along with the respective confidence intervals, which we calculated using OLS with standard errors clustered at the study level.

Table 6.1: Implied best-practice

Best-practice estimate			
Subjective best pratice	0.079 (0.018, 0.140)	0.060 (-0.003, 0.123)	**Lazear (2000a)**
Takahashi et al. (2016)	0.071 (-0.029, 0.171)	0.019 (-0.013, 0.051)	**Angrist et al. (2009)**

Note: The table reports estimates of the best-practice estimate according to three different studies and the author's subjective best-practice. 95% confidence interval bounds are constructed as an approximate using OLS with study level clustered standard errors.

In three of the estimations the implied best practice is slightly above the underlying mean (0.046). However, in most of these estimations the results are reported with considerably wide confidence intervals. In any case, this procedure seems to suggest that the mean implied by various studies in the data set seems to be more or less in line with what we have discovered thus far. For further details, we refer the reader to the original calculations. These can be found in the online appendix, along with the R code and the original data set.

6.2 Economic significance and final remarks

As our last contribution, we take a look at the economic significance of variables, which the BMA model assigned a posterior inclusion probability of 0.5 or higher. These nine variables are displayed in table 6.2, which then represents their ceteris paribus effect on the PCC, i.e., the dependent variable in the BMA model. We first calculate this effect when each of the variables changes by one standard deviation and second when it changes from its minimum to its maximum value. Furthermore, this effect is also presented as a percentage change in the subjective best-practice estimate calculated in the previous section. To give a concrete example, increasing the standard error by one standard deviation causes the PCC to increase by 0.0202 (i.e., 25.52% of the best-practice estimate). All of these results are presented in table 6.2.

Table 6.2: Significance of key variables

	One SD change		Maximum change	
	Effect on PCC	% of BP	Effect on PCC	% of BP
Standard error	0.0202	25.52%	0.2469	312.05%
Effect GPA	-0.0081	-10.24%	-0.0170	-21.48%
Effect Charity	-0.0233	-29.39%	-0.0520	-65.72%
Cross-sectional data	-0.0294	-37.13%	-0.0590	-74.56%
Lab study	0.0338	42.70%	0.0810	102.37%
Positive framing	0.0144	18.19%	0.0380	48.02%
Reward own	-0.0188	-23.73%	-0.0480	-60.66%
Quan. performance	-0.0272	-34.36%	-0.0590	-74.56%
Students	-0.0318	-40.13%	-0.0650	-82.15%

Note: This table presents ceteris paribus effect of several key variables on the partial correlation coefficient. Only those variables with PIP over 0.5 in the BMA model are included. *One SD change* implies how the PCC changes when we increase a specific variable by one standard deviation. *Maximum change* represents the change in the PCC when the variable is increased from its minimum to its maximum. The reference best-practice value is 0.079. SD = Standard Deviation, PCC = Partial Correlation Coefficient, BP = Best-Practice, GPA = Grade Point Average. For a detailed explanation of the variables, see table 5.1.

Out of the nine significant variables, three have a considerable positive effect on the PCC, while the remaining six pull the effect in the opposite direction. Most noticeable is surely the substantial influence of the standard error, which serves as a proxy for publication bias. An increase which is more than threefold when changing this variable from its minimum to its maximum suggests a presence of publication bias in our literature sample. From the rest of the variables we can highlight two extremes, namely the strong positive effect of laboratory setting and strong negative effect of the student subgroup.

All in all, the results we have obtained up to this point seem to be in line with both sides of the theory. The resulting overall effect close to zero seems to be stemming from an interplay of two main factors. On one side, the extrinsic rewards tend to provide our subjects with a boost of motivation, which leads to increased performance. On the flip side, the undermining effect (Deci 1971), possibly caused by the exclusivity of tangible rewards, works in the opposite direction to decrease the inherent enjoyment and self-determination during the task, nullifying the overall performance change to zero. This finding then seems to be in line with the one presented by Jenkins et al. (1998). Surprising is the size of the publication bias. We detected this bias mainly around the 5% significance level. However, we have no means of comparing these results to other papers, as our thesis is the only one considering potential publication bias.

As for the shortcomings of this approach, lumping together several effect types may make it hard to present the previous claims with confidence (Glass et al. 1981). The category capturing the behavior in response to the incentives could allow for more detail, which our models did not allow. Similarly, in the student category coexist both the groups that partook in a lab experiment and the groups observed at school for test performance. One could also implement a more detailed classification for the employees' category, where we seemingly put together all kinds of work. This approach inherently disregards the workers' relationship to their work and whether they enjoy it or not. Such specifications may turn out to be very interesting. Furthermore, our reward scheme is very straightforward due to our focus on the economic aspect of rewards. However important this may initially appear, a step more in the direction similar to Cerasoli et al. (2014) with more consideration towards the rewards' nature may end up being a little more appropriate. Similarly, our distinguishing between the period over which the subjects obtained the reward could also use a bit of refinement if we wanted to perfect our models.

When it comes to the models themselves, in the weighted BMA approach, we could not remove the constant from the model, which would be ideal. Although this constant may have influenced our results, we believe that this should not pose a significant problem due to its resulting negligible size. In different models, various specifications may have also yielded different results, although we hope that the results presented in this thesis represent their respective models quite well. However, we leave it up to the reader to see how the models behave when tweaked slightly, so we include these models along with the R code and the data set in the online appendix.

Chapter 7

Conclusion

Although the effect of rewards on motivation and performance is already well defined in psychology, we present a new outlook on this problem by restricting our approach to purely economic studies. Out of 44 of them, we collect a total of 1568 estimates. We then convert these estimates into partial correlation coefficient to better capture and compare various types of the effect.

By employing a wide range of statistical tests, we observe a close to zero overall effect of incentives on performance (0.012 - 0.035) while suggesting a presence of publication bias. These results stem from computations across 13 various tests, most of which display strong statistical significance. A Caliper test then confirms the preference for statistical significance at the 5% significance level. As for the comparison with literature, our computed effect size corresponds to the findings by Jenkins et al. (1998), while indirectly supporting the validity of the undermining intrinsic motivation theory Deci (1971); Cameron & Pierce (1994); Tang & Hall (1995).

We then define more than 30 different variables and, with the use of Bayesian model averaging and frequentist model averaging, look for heterogeneity in our data set. We also consider various specifications regarding the effect, method, study characteristics, reward scheme, task nature, motivation, and other attributes. As a result, a significant positive relationship appears between performance and the following specifications: positive framing and lab environment. On the other hand, a negative relationship of the same kind is captured for these specifications: charitable giving, cross-sectional data, self-obtained rewards, quantitative performance, and students. These results are all conditional on the subjects receiving a reward. The increase in performance under laboratory setting seems to be in line with the suggestion by Jenkins (1986), while the undermining effect of quantitative performance corresponds to the theory by Kruglanski et al. (1971); Evans (1979). None of our findings in this section were noticeably in contrast with our prediction. In the appendix, we provide a robustness check for these findings by altering the model specifications, namely priors and Zellner's g-priors.

As a bottom line of our thesis, we propose our subjective best-practice estimate and compare it to three other studies from our data set, which helps put our findings into perspective in regards to said literature. We observe similar pattern of behavior across these various specifications, which we then support by a computation focused on economic significance of several key variables from the Bayesian model averaging model. Similarly to the previous results, we again find a noticeable influence of the publication bias on the partial correlation coefficient. We then quantify the ceteris paribus influence of this publication bias along with the rest of the key variables in regards to the best-practice estimate in both absolute numbers and percentage change.

Lastly we consider several potential caveats of our thesis. Primarily, we are aware of the potential problems tied to the effect generalization (Glass et al. 1981), which we employ in order to transform the effect into partial correlation coefficient. We try to partially remedy this flaw by categorizing the effect into four main categories, however this approach still does not fully eliminate the loss of information. Similarly, we simplify several variables which display high correlation numbers in the Bayesian model averaging model, leading to an imprecise identification of the subject groups, as an example. Last but not least, because our focus is on the economic aspect of the effect, we choose a very straightforward reward scheme with lower emphasis on the nature of the reward. This means that the rewards' effect appears in the model on an equal footing with the rest of the explanatory variables. A clearer distinction in the impact of different types of rewards on other variables could certainly be employed.

Bibliography

Aguinis, H., Dalton, D. R., Bosco, F. A., Pierce, C. A., & Dalton, C. M. (2011). Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. *Journal of Management*, *37*(1), 5–38.

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

Allenby, G. M. & Rossi, P. E. (2006). *Hierarchical bayes models.* The handbook of marketing research: Uses, misuses, and future advances. 418-440.

Amini, S. M. & Parmeter, C. F. (2011). Bayesian model averaging in r. *Journal of Economic and Social Measurement*, *36*(4), 253–287.

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

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

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

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

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

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

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

Bom, P. R. & Rachinger, H. (2019). A kinked meta-regression model for publication bias correction. *Research synthesis methods*, *10*(4), 497–514.

Bonner, S. E., Hastie, R., Sprinkle, G. B., & Young, S. M. (2000). A review of the effects of financial incentives on performance in laboratory tasks: Implications for management accounting. *Journal of Management Accounting Research*, *12*(1), 19–64.

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

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

Camerer, C. F. & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of risk and uncertainty*, *19*(1), 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), 29–42.

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

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

Cazachevici, A., Havranek, T., & Horvath, R. (2020). Remittances and economic growth: A meta-analysis. *World Development*, *134*, 105021.

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

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

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

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

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

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

De Long, J. B. & Lang, K. (1992). Are all economic hypotheses false? *Journal of Political Economy*, *100*(6), 1257–1272.

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

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of personality and Social Psychology*, *18*(1), 105.

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

Deci, E. L., Koestner, R., & Ryan, R. M. (2001). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of educational research*, *71*(1), 1–27.

Deci, E. L. & Ryan, R. M. (1985). Cognitive evaluation theory. In *Intrinsic motivation and self-determination in human behavior* (pp. 43–85). Springer.

Deci, E. L. & Ryan, R. M. (2000). The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological inquiry*, *11*(4), 227–268.

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

Doucouliagos, C. & Laroche, P. (2003). What do unions do to productivity? A meta-analysis. *Industrial Relations*, *42*, 650–691.

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

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

Easterbrook, P. J., Gopalan, R., Berlin, J. A., & Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet*, *337*(8746), 867–872.

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

Eisenberger, R., Pierce, W. D., & Cameron, J. (1999). Effects of reward on intrinsic motivation-negative, neutral, and positive: Comment on deci, koestner, and ryan (1999).

Erat, S. & Gneezy, U. (2016). Incentives for creativity. *Experimental Economics*, *19*(2), 269–280.

Evans, G. W. (1979). Behavioral and physiological consequences of crowding in humans 1. *Journal of applied social psychology*, *9*(1), 27–46.

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

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

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

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

Ferriar, J. (1792). *Medical Histories and Reflections...*, volume 2. W. Eyres.

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

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

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

Furukawa, C. (2019). Publication bias under aggregation frictions: Theory, evidence, and a new correction method. *Evidence, and a New Correction Method.* (March 29, 2019).

Gallier, C., Reif, C., & Römer, D. (2017). Repeated pro-social behavior in the presence of economic interventions. *Journal of behavioral and experimental economics*, *69*, 18–28.

Gechert, S., Havránek, T., Irsova, Z., & Kolcunova, D. (2021). Measuring capital-labor substitution: The importance of method choices and publication bias.

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

Gerber, A. & Malhotra, N. (2008). Do statistical reporting standards affect what is published? Publication bias in two leading political science journals. *Quarterly Journal of Political Science*, *3*(3), 313–326.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Sage Publications, Incorporated.

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

Greene, R. J. (2018). *Rewarding performance: Guiding principles; custom strategies.* Routledge.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, *75*(4), 1175–1189.

Havránek, T., Irsova, Z., Laslopova, L., & Zeynalova, O. (2021). Skilled and Unskilled Labor Are Less Substitutable than Commonly Thought.

Havránek, T., Stanley, T. D., Doucouliagos, H., Bom, P., Geyer-Klingeberg, J., Iwasaki, I., Reed, W. R., Rost, K., & Aert, R. C. M. V. (2020). Reporting guidelines for meta-analysis in economics. *Journal of Economic Surveys*, *34*(3), 469–475.

Hedges, L. V. & Olkin, I. (2014). *Statistical methods for meta-analysis.* Academic press.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, 382–401.

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), 177–210.

Hunter, J. E., Ballard, T., Hunter, J., John Edward, H., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*, volume 4. SAGE Publications, Incorporated.

Hunter, J. E. & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings.* Sage.

Ioannidis, J. P., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research, F236–F265.

Ioannidis, J. P. & Trikalinos, T. A. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. *Cmaj*, *176*(8), 1091–1096.

Jenkins, G. D. (1986). Financial incentives. *Generalizing from laboratory to field settings*, *167*, 180.

Jenkins, G. D., Mitra, A., Gupta, N., & Shaw, J. D. (1998). Are financial incentives related to performance? a meta-analytic review of empirical research. *Journal of applied psychology*, *83*(5), 777.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate behavioral research*, *35*(1), 1–19.

Karlan, D. & List, J. A. (2007). Does price matter in charitable giving? Evidence from a large-scale natural field experiment. *American Economic Review*, *97*(5), 1774–1793.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

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

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

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

Kruglanski, A. W., Friedman, I., & Zeevi, G. (1971). The effects of extrinsic incentive on some qualitative aspects of task performance 1. *Journal of personality*, *39*(4), 606–617.

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

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

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

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

Mankiw, N. G. (2014). *Principles of economics.* Nelson Education.

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), 1203–1222.

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

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC medical research methodology*, *9*(1), 1–17.

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

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

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, *92*(437), 179–191.

Rosenthal, R. & Rubin, D. B. (1988). [selection models and the file drawer problem]: comment: assumptions and procedures in the file drawer problem. *Statistical Science*, *3*(1), 120–125.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. *Publication bias in meta-analysis: Prevention, assessment and adjustments*, 1–7.

Rummel, A. & Feinberg, R. (1988). Cognitive evaluation theory: A meta-analytic review of the literature. *Social Behavior and Personality: an international journal*, *16*(2), 147–164.

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of personality and social psychology*, *43*(3), 450.

Schall, D. L., Wolf, M., & Mohnen, A. (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*, 291–300.

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

Stanley, T. D. (2005). Beyond publication bias. *Journal of economic surveys*, *19*(3), 309–345.

Stanley, T. D. (2008). Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. *Oxford Bulletin of Economics and statistics*, *70*(1), 103–127.

Stanley, T. D. & Doucouliagos, H. (2012). *Meta-regression analysis in economics and business*, volume 5. routledge.

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

Stanley, T. D., Doucouliagos, H., Giles, M., Heckemeyer, J. H., Johnston, R. J., Laroche, P., & Nelson, J. P. (2013). Meta-analysis of economics research reporting guidelines. *Journal of economic surveys*, *27*(2), 390–394.

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

Steel, M. F. (2020). Model averaging and its use in economics. *Journal of Economic Literature*, *58*(3), 644–719.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance-or vice versa. *Journal of the American statistical association*, *54*(285), 30–34.

Sudarshan, A. (2014). Nudges in the marketplace: Using peer comparisons and incentives to reduce household electricity consumption.

Work. Pap., Energy Policy Inst. Chicago. http://www. anantsudarshan. com/uploads/1/0/2/6/10267789/nudges_udarshan_2014. pdf.

Takahashi, H., Shen, J., & Ogawa, K. (2013). An experimental examination of compensation schemes and level of effort in differentiated tasks. *Journal of Behavioral and Experimental Economics*, *61*, 12–19.

Tang, S.-H. & Hall, V. C. (1995). The overjustification effect: A meta-analysis. *Applied Cognitive Psychology*, *9*(5), 365–404.

Thornton, A. & Lee, P. (2000). Publication bias in meta-analysis: its causes and consequences. *Journal of clinical epidemiology*, *53*(2), 207–216.

van Aert, R. C. & van Assen, M. A. L. M. (2020). Correcting for publication bias in a meta-analysis with the p-uniform* method. *Manuscript submitted for publication Retrieved from: https://osfio/preprints/bitss/zqjr92018.[Google Scholar].*

Van Iddekinge, C. H., Aguinis, H., Mackey, J. D., & DeOrtentiis, P. S. (2018). A meta-analysis of the interactive, additive, and relative effects of cognitive ability and motivation on performance. *Journal of Management*, *44*(1), 249–279.

Wiersma, U. J. (1992). The effects of extrinsic rewards in intrinsic motivation: A meta-analysis. *Journal of occupational and organizational psychology*, *65*(2), 101–114.

Zeugner, S. & Feldkircher, M. (2009). Benchmark priors revisited: on adaptive shrinkage and the supermodel effect in bayesian model averaging.

Zhou, D., Zhao, S. Q., Liu, S., & Oeding, J. (2013). A meta-analysis on the impacts of partial cutting on forest structure and carbon storage. *Biogeosciences*, *10*(6), 3691–3703.

Zigraiova, D. & Havránek, T. (2016). Bank competition and financial stability: Much ado about nothing? *Journal of Economic Surveys*, *30*(5), 944–981.

Appendix A

A list of studies & Search query

Table A.1: Studies used in the meta-analysis

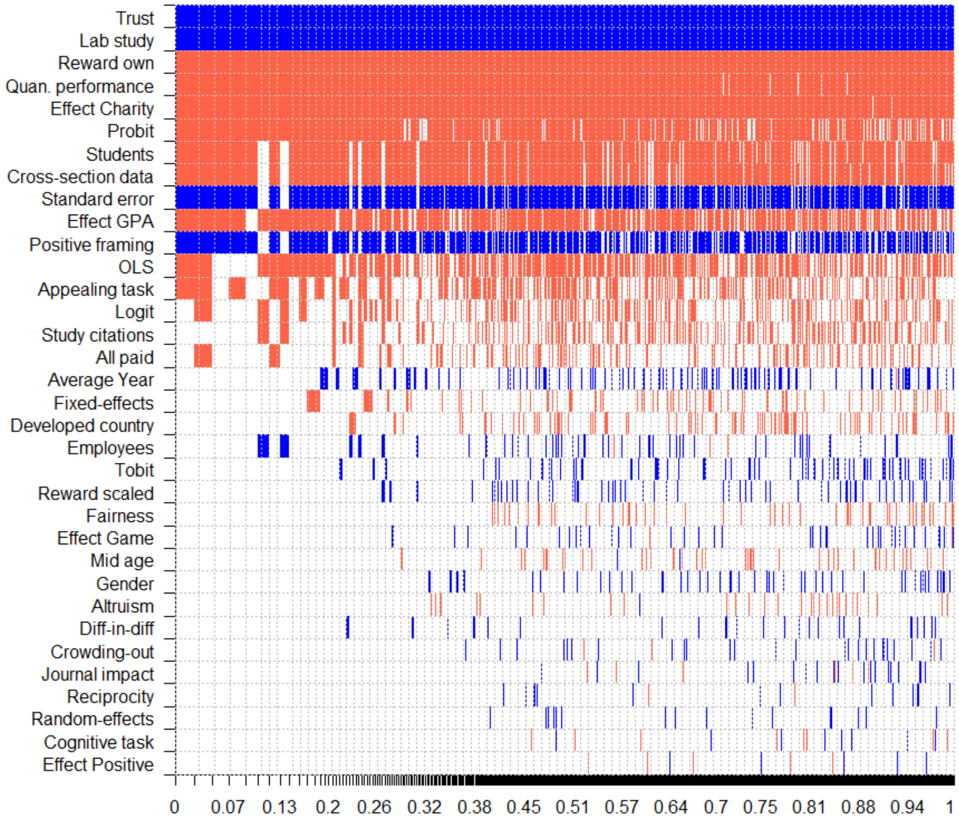
Alberts et al. (2016)
Angrist & Lavy (2009)
Angrist et al. (2009)
Ariely et al. (2009)
Ashraf et al. (2014)
Barrera-Osorio et al. (2019)
Boyer et al. (2016)
Bradler et al. (2019)
Cappelen et al. (2017)
Celhay et al. (2019)
Charness & Gneezy (2009)
Charness & Grieco (2019)
Coffman (2011)
Conrads et al.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

The finalized query, which we used while searching through the Google Scholar repository, had the following form:
("financial rewards" OR "money" OR "financial incentive" OR "financial incentives" OR "monetary incentives") AND ("motivation" OR "performance") effect affect experiment intrinsic extrinsic reward

Appendix B

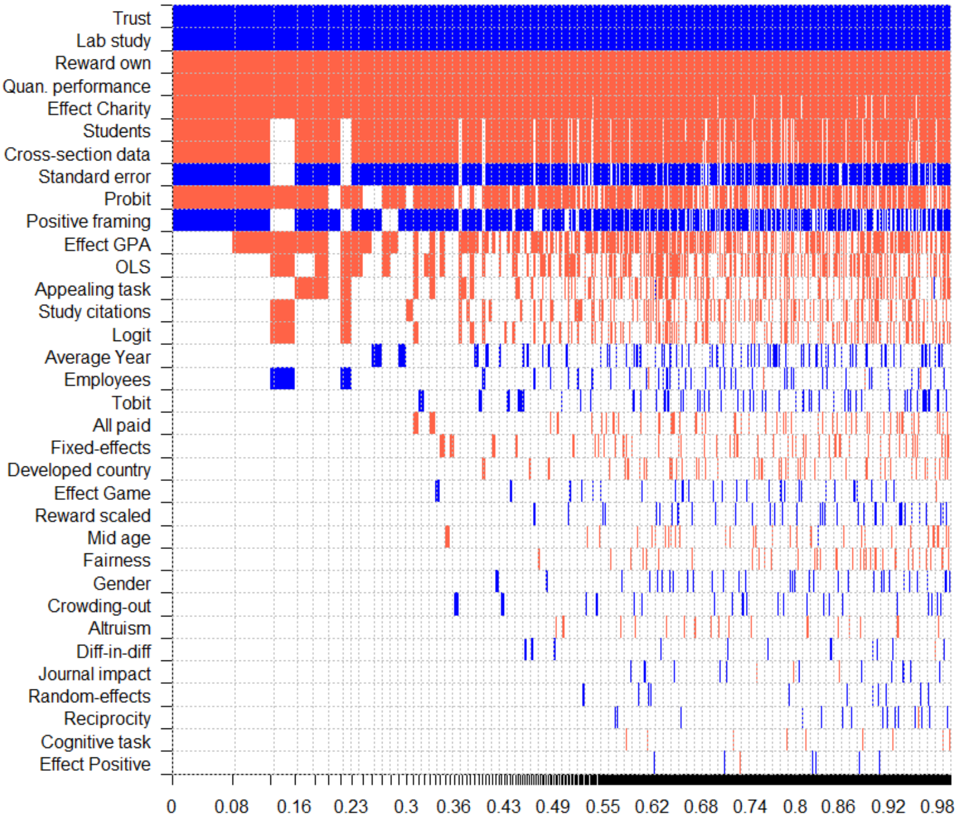
Bayesian model averaging robustness check

Figure B.1: BMA using uniform g-prior and uniform model prior



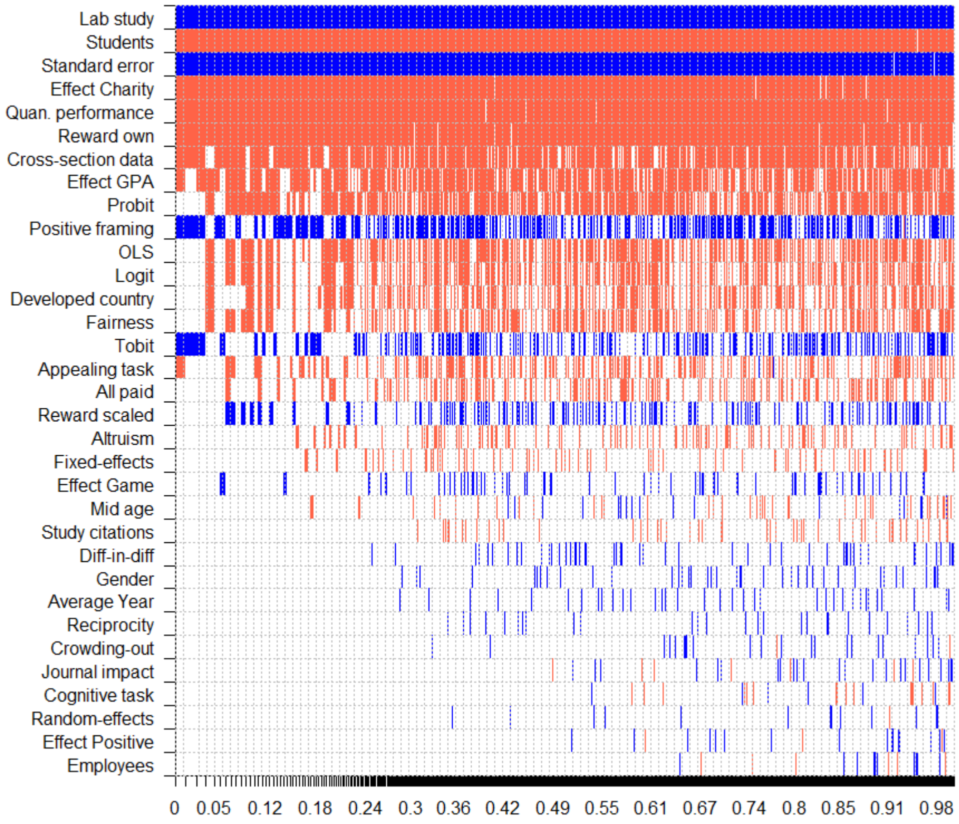
Note: This figure displays the results of the Bayesian model averaging using the uniform g-prior and the uniform model prior setup. BMA = Bayesian model averaging. For further explanation of the procedure and individual variables, see Figure 5.1 and Table 5.1.

Figure B.2: BMA using benchmark g-prior and random model prior



Note: This figure displays the results of the Bayesian model averaging using the benchmark g-prior and the uniform model prior setup. BMA = Bayesian model averaging. For further explanation of the procedure and individual variables, see Figure 5.1 and Table 5.1.

Figure B.3: BMA using HQ g-prior and random model prior

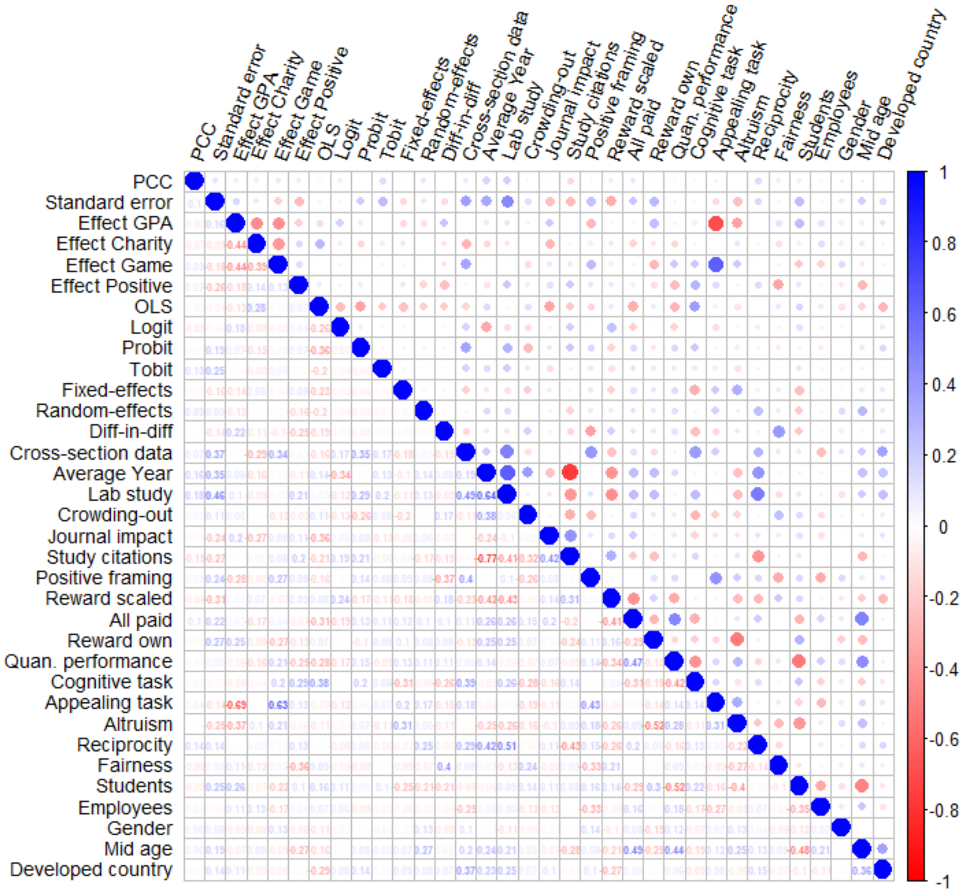


Note: This figure displays the results of the Bayesian model averaging using the Hannan-Quinn criterion g-prior and the uniform model prior setup. BMA = Bayesian model averaging. HQ = Hannan-Quinn Criterion. For further explanation of the procedure and individual variables, see Figure 5.1 and Table 5.1.

On pages above, we present three additional Bayesian model averaging procedures using different specifications to the one discussed in Chapter 5. We chose these specifications to best capture the potential differences in approach and present them in the note under each of the three models. With Zeugner & Feldkircher (2009) providing further detail on the theory behind each of the parameters in our setups, our primary goal in these checks was to control for the large number of observations included. The choice of the parameters (g-prior and model prior) then reflects this aim.

Lastly, we append a correlation table for our main model (see Chapter 5).

Figure B.4: Correlation table for the Bayesian model averaging



Note: This figure displays the correlation table for the Bayesian model averaging approach using the uniform g-prior and the dilution prior, the results of which are shown and discussed in Chapter 5. Blue color (dark in grayscale) indicates positive correlation, while red color (light in grayscale) indicates negative correlation. For a detailed explanation of the variables, see table 5.1.