CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Essays in Behavioural and Experimental Economics

Dissertation thesis

Author: Mgr. Jindřich Matoušek

Study program: Economics and Finance

Supervisor: prof. PhDr. Tomáš Havránek, Ph.D.

Year of defense: 2022

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, July 16, 2022

Jindřich Matoušek

Abstract

The dissertation consists of three papers presenting applications of experimental as well as statistical methods to the topics of behavioural economics. The first paper introduces a series of laboratory experiments in which I apply the experimental methods to a complex decision making problem. The second and third papers present quantitative syntheses of the literature on the classical topics of behavioural economics. The general introduction connects these chapters together. Detailed abstracts for individual papers are presented at the beginning of each chapter.

In the first paper, I experimentally examine two complex multi-unit auction mechanisms with an opportunity to communicate and thus collude while comparing these mechanisms in terms of efficiency. Strikingly, allowing for communication increases efficiency in examined auction formats. A cheap-talk collusive agreement resulted in a better allocation compared to the treatments without communication. I hypothesize that complex auction formats makes the decision-making of bidders too complicated and causes inefficiency, especially in auctions with large numbers of goods.

In the second paper, I provide a meta-analysis of a key parameter estimated by both lab and field experiments in economics — the individual discount rate. I examine the extent to which the variance of the discount rate can be attributed to observable differences in experimental design as well as the selective reporting in the literature. I employ Bayesian and frequentist model averaging to address the model uncertainty and identify the drivers that affect the individual discount rate the most. The results show publication bias against unintuitive results. The corrected mean annual discount rate is less than half the size of the simple mean of the reported values.

In the third paper, I study whether financial incentives motivate people to work better. I take stock of emerging research in economics, collect a total of 1568 estimates from 44 different studies, and codify 39 variables to capture the underlying nature of the effect incentives have on motivation and performance. I perform a meta-analysis on this dataset. A range of statistical tests suggests the overall effect is virtually zero. I employ Bayesian and frequentist model averaging to identify the most prominent effect determinants. Among these, publication bias pushes this effect upwards the most.

Acknowledgments

I would like to sincerely thank my supervisor prof. PhDr. Tomáš Havránek, Ph.D., for his guidance, valuable comments and suggestions. Tomáš Havránek has been a great inspiration, always knowing how to motivate just right when the strengths were coming to an end. Tomáš, it has been a pleasure working with you. Also, I would like to thank my friend and colleague PhDr. Lubomír Cingl, Ph.D., who inspired and helped me to set foot on this academic journey of mine. Last but not least, my deepest gratitude belongs to my family—my parents for their leading example, my wife for her everlasting support, and my little ones for exceptional patience with their father.

Typeset in LaTeXusing the IES Thesis Template.

Bibliographic Record
Matoušek, Jindřich: *Essays in Behavioural and Experimental Economics.* Dissertation thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2022, pages 208. Advisor: prof. PhDr. Tomáš Havránek, Ph.D.

Contents

List of Tables

List of Figures

Chapter 1

General Introduction

Methodology of behavioral economics returns economic thinking to the way it began, with Adam Smith, and continued through the time of Irving Fisher and John Maynard Keynes in the 1930s.
— Richard Thaler (2016, p. 1577)

Policy-, as well as decision-makers on a microeconomic level, turn increasingly often to behavioural economics for understanding how individuals, groups and markets behave. Behavioural economics models human behaviour using a combination of knowledge and insights from various fields, such as economics, psychology, sociology and neuroscience. It incorporates emotions, fairness, reciprocity or social norms into traditional economic models, and applies experimental methods to empirically analyse and test its theories. An experimental examination of empirical procedures can provide important insights since human decision-making is often skewed by various cognitive biases (e.g. loss aversion, Tversky & Kahneman 1974). Moreover, only in an economic laboratory can a proper fully controlled environment be created to examine the impact of a change in a single variable, as the real world is confounded by too many factors and isolation of a causal effect is very difficult, if not in many cases impossible. An experimental approach thus provides a very useful tool that complements other methods of analysis (Samuelson 2005).

In this dissertation thesis, I use experimental as well as statistical methods to understand the behaviour of individuals in situations with specific economic contexts. First, I study how experimental subjects behave in a complex multi-unit auction framework. I conduct several economic experiments to examine a simultaneous multi-round auction format and its extension for combinatorial bidding. By allowing for communication between subjects, I induce a collusive environment and evaluate the efficiency of the auctions with and without the collusion of its participants. Second, I study how people behave within the context of intertemporal decisions, and what are their preferences over the present and future outcomes. From research based on an experimental examination, I move to statistical methods. I take a behavioural literature stream on a single microeconomic parameter that enters into various macroeconomic models—the individual discount rate—and synthesize its true value. Third, I study how people react to incentives and whether motivating them financially increases their performance. Looking at studies that use experimental methods to identify the effect of rewards on motivation I provide a quantitative synthesis of the behavioural economics literature on the topic.

Results of individual dissertation chapters suggest that the design and manageable complexity of the experiment are crucial for its robust results. Besides the individual aspects of an experimental task, it matters also whether the experiment is conducted in the lab or field. Moreover, also the composition of the sample of experimental subjects (the subject pool) may have a systematic impact on the results. Behavioural literature should therefore take great care when designing economic experiments to test its hypotheses since these two factors might question the external validity of some experiments. Using synthetic methods, I find evidence of selective reporting in both examined literature streams. Researchers and editors tend to publish only statistically significant results, yet, the insignificant and unpublished data contains a lot of valuable information that has the potential to add further value to general debate. I elaborate on each study more deeply below while drawing inspiration from their introductions.

In the Chapter 2, I study the behaviour of individuals in complex auction mechanisms with multiple goods for sale. Multi-unit auction mechanisms are one of the most important instruments for the allocation of goods in complex real-life situations. Used for the allotment of spectrum licenses, airport time slots, delivery routes, networking, and furniture, they are one of the few outstanding innovations of modern economics (De Vries & Vohra 2003; Guala 2001). A goal of every auction should be allocating the objects for sale to those who value them the most. Real auctions, however, do not always produce the results predicted by the theory and, more importantly, results within the same auction types but across different settings can vary (Holt 2005).

One of the omnipresent main issues in real auctions is the possibility of collusion among bidding participants. By using coordinated strategies, bidders can keep prices at low levels, thereby decreasing the revenues of the auctioneer. A variety of experimental studies have therefore examined the evolution of collusion in auction mechanisms (Burtraw *et al.* 2009; Bachrach 2010; Zhou & Zheng 2010; Hu *et al.* 2011). Generally, collusion in auctions can emerge either through repeated interaction between bidders, bidding that occurs over multiple objects (Agranov & Yariv 2015), or, perhaps most commonly, through communication. In multi-unit auctions, bidders may in addition coordinate their strategies in an attempt to split the objects for sale and reach a more profitable outcome than would originate from a competitive situation (Kwasnica & Sherstyuk 2013). Despite having several strategic challenges reported by Bichler & Goeree (2017) such as eligibility management, problems with signalling, and most importantly in our case the potential for tacit collusion, the simultaneous multi-round auction formats have been used primarily for the allocation of telecommunication spectrums in recent decades.

Collusion in complex simultaneous auctions of multiple goods has generally not yet been properly examined for mechanisms with more than only a few objects for sale. Although there is evidence that allowing for combinatorial bidding on packages of goods may break collusion in multi-unit auctions, the existing literature differs in its conclusions. The first line of literature represented by Brunner *et al.* (2010) discovers that combinatorial bidding increases the efficiency of simultaneous auctions. The second literature stream represented by Bichler *et al.* (2014) and Goeree & Holt (2010) conclude precisely the opposite: that combinatorial formats with a high number of goods at stake are not computationally manageable for their participants and their efficiency is, therefore, lower compared to basic simultaneous multi-round auctions.

In Chapter 2, together with my colleague I thus experimentally examine the behaviour of individuals in complex multi-unit auction mechanisms and investigate whether a particular design extension where bids are placed on bundles of goods rather than on single units—a package bidding—is capable of preventing the non-competitive practices. We experimentally evaluate the performance of two simultaneous auction mechanisms for selling multiple goods, wherein in one set of treatments we examine the effects of communication on auction efficiency and the auctioneer's revenues, while in the second, orthogonal set of treatments compares the baseline simultaneous multi-round auction format (SMR) with its combinatorial extension that incorporates package bidding (SMRPB), thereby employing a 2x2 design resulting in four treatment cells. To this aim, we create a stylized parametric environment of the 2013 Czech Spectrum Auction to reflect a real-world situation. In other respects, we base our experimental design on the previous literature (Brunner *et al.* 2010).

Our results show that communication was effectively used by subjects to coordinate and split the market. In terms of efficiency, the package bidding format in our setting does not differ from the basic simultaneous multi-round auction, though a significant difference arises favouring the non-combinatorial SMR when allowed for communication among bidders. We deduce that the non-combinatorial format of the auction enhances the stability of collusion within the auction, or, from another perspective, prevents strong bidders from breaking the entire agreement on a specific collusive outcome. In the package bidding format, strong bidders are tempted to break collusion more often. They acquire lower profits resulting from higher prices but may gain from weaker bidders being left with only scraps in return.

In the Chapter 3, I study the behaviour of individuals in intertemporal decisions. Intertemporal trade-offs are key to a host of decision problems at both the private and public levels. Policies addressing climate change, particularly those underpinned by the literature on the social cost of carbon, constitute a typical example of choices for which individual discounting of future costs and benefits plays a crucial role (Tol 1999; Goulder & Stavins 2002; Fujii & Karp 2008; Anthoff *et al.* 2009). Discount rates of individuals also reflect the underlying transaction costs of borrowing money that households face (Kovacs & Larson 2008).

For some of these decisions, it is appropriate to employ the market discount rate, which is detectable from financial time series. For others, however, we must try to recover the underlying discount rates of individuals. Individual discount rates can be either observed from existing data (such as in Lawrance 1991; Dreyfus & Viscusi 1995; Warner & Pleeter 2001) or measured experimentally (Benzion *et al.* 1989; Chapman & Elstein 1995; Coller & Williams 1999; Harrison *et al.* 2010, among others). Controlled experiments provide a natural framework for inferring the subject's preferences. Researchers can vary the experimental parameters to explore time discounting in both laboratory and field conditions. However, no consensus on how to best measure discounting has emerged (Andreoni *et al.* 2015). Discount rates differ across individuals and their estimates vary a great deal throughout the literature (Coller & Williams 1999; Frederick *et al.* 2002).

In Chapter 3, together with my colleagues I, therefore, take stock of the evidence and trace the differences in the reported discount rates to the design of experiments while accounting for model uncertainty. We also control for the effects of potential selective reporting. Focusing on various aspects of existing studies we employ the Bayesian model averaging (BMA; Raftery *et al.* 1997) and frequentist model averaging (FMA; Hansen 2007) to examine which ones matter the most for the differences among the reported estimates.

We find that selective reporting that causes publication bias represents an important factor in the literature. Even though zero or negative discount rate estimates make little sense in most contexts, they should appear in specific contexts. When selective reporting is present, however, insignificant and negative estimates are discriminated against. Our findings indicate that such publication bias is associated with exaggerating the mean reported annual discount rate from 0.33 to 0.80. Aside from publication bias, the differences in results seem to be caused primarily by the experimental design of discounting tasks. We find evidence in line with domain independence (defined as the low correlation between discount rates for different domains) in intertemporal choice (Loewenstein *et al.* 2003; Ubfal 2016): it matters what the experimental subjects should be patient or impatient about. Subjects are more patient with regard to money than health or more exotic contexts (such as vacations, certificates, and kisses from movie stars). The results support the hypothesis that liquidity constraints play a key role in intertemporal choice experiments (Dean & Sautmann 2021). We also find that negative framing is associated with more patience, which corroborates the notion that anticipation of dread is important in intertemporal decisions (Harris 2012).

In the Chapter 4, I study how individuals respond to incentives and how these transfer to their performance. Financial incentives were long perceived as the key to success in motivating people to perform better. By increasing one's pay, the institutions were inducing higher quality of work from their employees. But is it true that financial incentives motivate people to work better? An answer to this question can be viewed from two perspectives. 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). 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 *et al.* 1999). This phenomenon became commonly known as the "crowding-out intrinsic motivation" theory (Deci 1971).

Even though decades of research provided countless experiments that helped put these rewards-motivation findings into a quantitative perspective, the most prominent ideas from the domain do not present identical findings. The variations in results and specific contexts are substantial. 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 toward 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 that involve judgment and find a positive monetary incentive effect. Bonner *et al.* (2000) 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 performance. Gneezy & Rustichini (2000) argue that one has to be paid enough to show an increase in one's performance. Last but not least, Ariely *et al.* (2009) discover that excessive rewards may have a detrimental effect on performance.

Even though the largely heterogeneous literature about the effect of rewards on motivation was synthesized before (Rummel & Feinberg 1988; Cameron & Pierce 1994; Jenkins *et al.* 1998; Deci *et al.* 1999; Cameron 2001; Cerasoli *et al.* 2014; Van Iddekinge *et al.* 2018, among others), none of the studies try to isolate the outlooks of either economists or psychologists by looking at the available literature from strictly separated perspectives. And yet the outlook and more importantly the expectations of an economist would be different from the ones of a psychologist. Also, very little space is devoted to the effect commonly explored in synthetic studies— selective reporting—and yet this phenomenon is widespread in economics as well as other fields in the available literature (Doucouliagos & Stanley 2013; Ioannidis *et al.* 2017). Large heterogeneity among these results together with the absence of answers to specific questions suggests that an additional synthesis of this topic would bring substantial value to the field.

Hence, in Chapter 4 together with my colleague, I synthesize the behavioural economics literature on the topic of rewards and motivation in a quantitative meta-analysis. We focus on the perspective of an economist and look, therefore, at how the effect behaves across economic literature. We aggregate individual studies together and observe the relationships and causalities of the rewards-motivation effect, together with its potential systematic misbehaviour (Hunter *et al.* 1982). Furthermore, we employ advanced methods to uncover potential selective reporting in the literature. Lastly, we examine the differences in the effect and map these to different angles related to effect characteristics, task nature, reward scheme, motivation characteristics, study design, subject pool characteristics, methodology, and publication characteristics. Accounting for model uncertainty, we employ the Bayesian model averaging (Raftery *et al.* 1997, BMA) and frequentist model averaging (Hansen 2007, FMA) to discover which characteristics affect the reported estimates the most.

Together with a noticeable selective reporting, a result demonstrating a publication bias in the literature, we find a minimum overall impact of rewards on motivation. Our results highlight the importance of individual driving factors behind the effect as far as the literature heterogeneity is concerned. Most importantly to the field of experimental economics, rewards have a larger effect on performance in the laboratory rather than in a field setting. The same appears also when the framing of the task is positive. On contrary, usage of a students' sample as experimental subjects reports lower performance. All three characteristics are used in the experimental economics methodology widely. We, therefore, prompt for great care when designing experiments dealing with performance tasks and motivation, since standard experimental setup may provide biased results.

Chapter 2

Collusion in Multi-Object Auctions: Experimental Evidence

Abstract We experimentally examine two complex multi-unit auction mechanisms —a simultaneous multi-round auction and its extension with combinatorial bidding—with an opportunity to communicate and thus collude. The general setting and parametrization originated in the 2013 Czech Spectrum Auction. Our results suggest that the package bidding format does not bring higher efficiency. Strikingly, allowing for communication increases efficiency in examined auction formats. A cheap-talk collusive agreement resulted in a better allocation compared to the treatments without communication. We hypothesize that combinatorial bidding makes the decision-making of bidders too complicated and causes inefficiency, especially in auctions with large numbers of goods.

This paper was published in *Eastern European Economics [2018, 56, pp. 28-56, DOI: 10.1080/00128775.2017.1394161]*. I am grateful to Peter Katuščák and Fabio Michelucci for useful discussions, and Tomáš Havránek, Mario Holzner and Michal Pakoš as well as two anonymous referees for valuable comments. Supplemental data for this article can be accessed on the `publisher's website`. I acknowledge support from the CERGE-EI Foundation under a program of the Global Development Network (GDN RRC 1464) and Grant Agency of Charles University (GAUK Grant No. 848217).

2.1 Introduction

Multi-unit auction mechanisms are one of the most important instruments for the allocation of goods in complex real-life situations. Used for the allotment of spectrum licenses, airport time slots, delivery routes, networking, and furniture, they are one of the few outstanding innovations of modern economics (De Vries & Vohra 2003; Guala 2001) and are also increasingly popular also outside of the world's richest economies. The main concern of every auctioneer should be the efficiency of the type of the auction employed, that is, allocating the objects for sale to those who value them the most. This is not a simple task, especially when real auctions do not always produce the results predicted by the theory. Moreover, the theoretical literature suggests that even results within the same auction types but across different settings can vary (Holt 2005). In general, an experimental examination of empirical procedures can provide important insights, since human decision making is often skewed by various cognitive biases (e.g. loss aversion, Tversky & Kahneman 1974). Moreover, only in an economic laboratory can a proper fully controlled environment be created in order to examine the impact of a change in a single variable, as the real-world is confounded by too many factors and isolation of a causal effect is very difficult, if not in many cases impossible. An experimental approach thus provides a very useful tool that complements other methods of analysis (Samuelson 2005).

One of the omnipresent main issues in real auctions is the possibility of collusion among bidding participants. By using coordinated strategies, bidders can keep prices at low levels, thereby decreasing the revenues of the auctioneer. A variety of experimental studies therefore has examined the evolution of collusion in auction mechanisms (Burtraw *et al.* 2009; Bachrach 2010; Zhou & Zheng 2010; Hu *et al.* 2011). Generally, collusion in auctions can emerge either through repeated interaction between bidders, bidding that occurs over multiple objects (Agranov & Yariv 2015), or, perhaps most commonly, through communication. In multi-unit (spectrum) auctions, bidders may coordinate their strategies in an attempt to split the objects for sale and reach a more profitable outcome for themselves than that originating from a competitive situation (Kwasnica & Sherstyuk 2013). There is strong evidence that ascending auctions are particularly vulnerable to collusive behavior of bidders, despite several design extensions developed in order to prevent non-competitive practices, including package bidding (Klemperer 2004). As a result, an important aspect of spectrum auctions is their high complexity. Therefore it is advised that only highly sophisticated players participate in the auction, which is carefully set up by the auctioneer; otherwise it may not function as intended.

The 2013 Czech Spectrum Auction was set up according to the current standards of multiunit auctions with all the pros and cons such standards deliver. However, its execution raised public awareness rather than sufficient funds. After its first version was cancelled by the auctioneer in 2012 because of ”too high prices”, its repetition a year later actually resulted in suspiciously low prices and revenues (CTO 2013a). The situation in the Czech telecommunication market actually matched survey results finding that young, large companies in the Eastern European region that do not face significant competition perceive corruption more favorably, thus suggesting that they engage in such practices (Sahakyan & Stiegert 2012). This conclusion is not exclusive to Eastern European countries. Bichler *et al.* (2017) review the 2015 German Spectrum Auction with only three incumbent operators bidding; a setting very similar to that of the 2013 Czech Spectrum Auction. They show that in the beginning of the auction, bidders were actively searching for a way to allocate the available spectrum that all bidders could agree to at low prices. The bidders were essentially teaching each other what they should bid. We were broadly inspired by the case of the 2013 Czech Spectrum Auction and investigate the effects of potential collusion due to communication in the specific setting of the Czech auction that, in our opinion, suffered from a major design flaw. We do so since this case provides a reasonable example of a real-world controversial situation with an unproven suspicion of collusion that can be translated into laboratory conditions.

Despite having several strategic challenges reported by Bichler & Goeree (2017) such as eligibility management, problems with signalling, and the potential for tacit collusion, simultaneous multi-round auction formats represent a framework that has been used for allocation of spectrums most frequently in recent decades. Mochon & Saez (2017) identified key variables of spectrum auctions in which the combinatorial clock format was used. Several of these variables are rather general for the whole class of spectrum auctions—spectrum packaging policy, reserve price, demand limit, activity rule, and pricing rule—and were crucial in the 2013 Czech Spectrum Auction as well.

Next, we investigate whether in such an environment one particular design extension is capable of preventing the non-competitive practices: we add the feature of package bidding, where bids are placed on bundles of goods rather than on single units. Collusion in complex simultaneous auctions of multiple goods has generally not yet been properly examined for mechanisms with more than only a few objects for sale. On the one hand, large scale auctions may offer enough possible combinations for bidders to find profitable collusive allocation, but on the other hand it can create a coordination problem so that it is too complicated for colluding bidders to cooperate successfully (Kwasnica & Sherstyuk 2013). Although there is evidence that allowing for combinatorial bidding on packages of goods may break collusion in multi-unit auctions, the existing literature differs in its conclusions. The first line of literature represented by Brunner *et al.* (2010) discovers that combinatorial bidding increases the efficiency of simultaneous auctions. The second literature stream represented by Bichler *et al.* (2014) and Goeree & Holt (2010) concludes precisely the opposite: that combinatorial formats with a high number of goods at stake are not computationally manageable for their participants and their efficiency is therefore lower compared to basic simultaneous multi-round auctions.

Our contribution to the literature is that, to the best of our knowledge, we are the first to experimentally evaluate the performance of two simultaneous auction mechanisms for selling multiple goods, where in one set of treatments we examine the effects of communication on auction efficiency and the auctioneer's revenues, while the second, orthogonal set of treatments compares the baseline simultaneous multi-round auction format (SMR) with its combinatorial extension that incorporates package bidding (SMRPB), thereby employing a 2x2 design resulting in four treatment cells with parameters creating a stylized environment of the 2013 Czech Spectrum Auction. In other respects, we base our experimental design on the previous literature (Brunner *et al.* 2010).

Our design involves more than fifty units of goods of four heterogeneous complementary types. Bidders' individual values consist of common and private value components.[1] Although restricted by the limit on individual activity, the four bidders participating in each auction possess enough possible combinations of ways to divide the goods for sale and thus are able to form a stable collusive equilibrium.[2] The possibility of communication is introduced through a simple chat window. We do not use any binding commitments or transfer promises in our design.

[1] This structure is standard and aims to capture the fact that a bidder has typically (i) an intrinsic preference for the sold good that depends on her own abilities to make a profit on the auctioned good, and (ii) beliefs about the price for which the good could be resold in the future to others (Noussair & Seres 2020)

[2] Our design provides, without restrictions, over 36 billion possible outcomes: $C(4;6) \cdot C(4;24) \cdot C(4;14) \cdot C(4;9) = 36.756.720.000$ possible combinations.

Our results show that communication was effectively used by subjects in order to coordinate and split the market. The package bidding format in our setting does not differ in terms of efficiency from the basic simultaneous multi-round auction, though a significant difference arises favoring the non-combinatorial SMR with allowed communication. We deduce that the non-combinatorial format of the auction enhances the stability of collusion within the auction, or, from another perspective, prevents strong bidders from breaking the entire agreement on a specific collusive outcome. In the package bidding format, strong bidders are tempted to break collusion more often. They acquire lower profits resulting from higher prices, but may gain from weaker bidders being left with only scraps in return.

Our results thus imply that on the one hand, the 2013 Czech Spectrum Auction was designed surprisingly appropriately in terms of the employment of a non-combinatorial format. Even though a collusive market should be unacceptable from a policy maker's perspective, if it happens to be a real situation, a simple non-combinatorial format may deliver more efficient results than its extension with package bidding, since communication among bidders can increase the efficiency of the auction. This finding fits the literature from both theoretical and experimental sources: it has been suggested that collusion does not always lead to lower efficiency. In case the valuation of bidders is composed of private and common value components, it may happen that even when all bidders behave rationally, the auctioned good may be allocated inefficiently (Goeree & Offerman 2003). Recently a model has been developed and an experimental test of this claim has been carried out in a double-auction setting (Noussair & Seres 2020). Bidders can cooperate on beneficial strategies, split the market and reach a stable equilibrium. Regardless of the number of possible combinations, bidders tend to collusive payoff-maximizing strategies by which they gain not only higher profits but, unintentionally, also more total revenue for the auctioneer.

What we see as a design shortcoming of the 2013 Czech Spectrum Auction is the absence of a mechanism that would control for the allocation of maximum possible units of goods, used e.g. in Brunner *et al.* (2010). This part of the design seems to be crucial since if employed, it could prevent a situation when many goods are not allocated and consequently low auction efficiency results.

2.2 Literature

2.2.1 Concepts of Multi-Unit Auctions

The SMR auction format was originally developed in the early nineties by the US Federal Communications Commission (FCC) for their spectrum auctions. The description of the process of designing, testing and implementing the FCC auctions as one of the few cases of complex economic engineering is available in Guala (2001).

To win a specific set of goods that has a particular value to the bidder as a whole can become complicated in non-combinatorial auctions, of which the SMR is a standard representative. Bidders with high value complementarities may have to bid more for some combinations of licenses than the licenses are actually worth individually to them. When only a part of a desired package is won, the bidder can incur large losses. This situation exposes bidders to a great risk, and may lead to conservative bidding during the auction and therefore to lower revenues and inefficient allocation of the auctioned goods (Brunner *et al.* 2010).

Brunner *et al.* (2010) deal with the different formats in flexible combinatorial spectrum auctions in environments including complementarities. They compare a widely used simultaneous multi-round auction with three other formats: the simultaneous multi-round format with package bidding (SMRPB), a combinatorial clock (CC) auction, and the "Resource Allocation Design" (RAD) auction. They use a series of laboratory experiments to evaluate these alternative multi-unit auction formats. Their results suggest that all three combinatorial auction procedures are more efficient than the SMR auction format when value complementarities are present. As the interrelation of auction objects is a common feature of many auction types, this finding is crucially important for practice. In addition, all the formats in their setting reach different results in terms of efficiency as well as sellers' revenue.

Bichler *et al.* (2013) cast doubt on the mainstream literature results mentioned above and provide results favoring non-combinatorial auction formats in efficiency. They compare the combinatorial clock auction to SMR. They analyze the efficiency of the auction methods and the auctioneer's revenue. They also examine bidding behavior in both cases. Their experiments are based on two value models resembling single and multi-band spectrum auctions, which often offer thousands of possible bundles. The efficiency of the CC auction is significantly lower than that of the SMR in the multi-band model in their case. Moreover, auctioneer's revenue is lower in both value models for CC. The second recent paper dealing with high numbers of auctioned goods in the same multi-band models is Bichler *et al.* (2014). They find that the simplicity of bid language has a substantial positive impact on the efficiency of the auction. Moreover, the simplicity of the payment rule has a substantial positive impact on auction revenue. The CC auction scores the worst in both dimensions in their experiment, favoring the SMR auction format. These results are in contradiction to Brunner *et al.* (2010) or, for example, Cramton (2013), who prefer combinatorial bidding auction formats.

Goeree & Holt (2010) suggest that even though the combinatorial auction can solve the problem of goods packaging by allowing competition among the bidders to determine the market structure, the decision-making problem in large complicated auctions that all rational bidders have to solve after each round can be computationally hard to manage. Specifically, they claim that: "*...bidders will not be able to reproduce the outcome of a round to understand why their bids did not win, unless they solve a non-deterministic polynomial-time hard problem quickly*"[3] (Goeree & Holt 2010, p. 149). They then propose a new hierarchical package bidding (HPB) combinatorial auction format which should be computationally manageable. The general result of their article is that the proposed HPB format is a "*paper & pencil*" package auction format. It is simple to implement, transparent and easily verifiable by the bidders (Goeree & Holt 2010).

2.2.2 Collusion in Auctions

Collusion in auction mechanisms has been studied for years (Robinson 1985; Crawford 1998; Kwasnica & Sherstyuk 2013). Bidding rings are a theoretically well-described method of collusion (Krishna 2009). Traditionally, various single-unit auction mechanisms in which collusion has implied explicit communication among bidders have been subject to research interest. Experimental studies further confirm that collusion can and actually does occur when communication is present in single-unit auctions. There is strong evidence that ascending auctions are particularly vulnerable to collusive behavior and also very likely deter entry into the auction (Klemperer 2004). Furthermore, the tacit form of collusion, during which participants silently coordinate on low-

[3] We observed exactly this type of situations during the execution of the experiment.

price outcomes, has been widely observed and theoretically described. Bidders simply tend toward collusive payoff-maximizing strategies (Skrzypacz & Hopenhayn 2004).

Recent auction literature concentrates on sequential and multi-round auction formats which appear to most often lead to bidder conspiracies (Kwasnica & Sherstyuk 2013). In multi-unit auctions, bidding agents can silently split objects and keep the competition at low levels throughout the auction. A large number of bidders is not a sufficient condition to hinder collusion as long as the bidders can share a sufficiently large number of the goods sold in the auction among themselves. Moreover, depending on the parameter setting of each individual auction, the multi-unit nature of auctions usually introduces complexities into the environment and the outcomes are therefore uncertain (Kwasnica & Sherstyuk 2013).

Phillips *et al.* (2003) document the impact of practices that may facilitate low final prices in repeated English auctions with multiple units. They create artificial laboratory markets using English auctions with a symmetric structure of bidders. By employing two market sizes (two and six bidder structures) they control for competitive and rivalrous environments. Three practices are identified as potentially facilitating collusion among bidders: (i) knowledge about the number of units for sale; (ii) familiarity through repeated interaction; and (iii) communication. According to their results, repeated interaction should allow buyers to learn the bidding strategies of their opponents even without communication. Moreover, if the agents can talk or exchange information, agreements become easier and bid prices lower.

Miralles (2010) analyzes a generalization of Campbell's self-enforced collusion mechanism in simultaneous auctions. While Campbell (1994) based his collusion mechanism on complete comparative cheap-talk and endogenous entry with only two bidders, Miralles (2010) examines cases of more than two bidders with prior symmetric design. He focuses on self-enforced and simple mechanisms without side-payments or trigger strategies. He uses a pre-play cheap-talk, which is *"clearly difficult to prosecute by competition authorities"* (Miralles 2010, p. 525). Two important results arise from the analysis: (i) a cheap-talk equilibrium exists if the number of objects is large enough; and (ii) a partial cheap-talk equilibrium, in which *"each bidder splits the objects into two sets, the favorite one and the rest, and lets the other bidders know about that split"* (Miralles 2010, p. 526), always exists.

Agranov & Yariv (2015) experimentally study collusion through communication in one-shot first- and second-price sealed bid auctions with two bidders. The results of their research suggest that communication alone can dramatically affect auction outcomes. They document two strategies in their simple one-by-one cheap-talk auction environment. The *reveal-collude* strategy in which the players reveal their valuations and consequently collude can potentially be applicable to our multi-unit case. Regardless of the strategy players used in their experiment, communication led to significant price drops, reducing auction revenues by up to 33%.

2.3 Hypotheses

We expect the experimental parameters to be different in comparisons within and between formats and therefore construct null hypotheses within and between formats separately. We examine four similar parameters in all treatments, i.e. the average total prices paid by one bidder, the efficiency of the auction, the auctioneer's revenue and the final profits of bidders. All pairs of respective parameters should be equal between individual treatments under null hypotheses. Table 2.1 summarizes the comparison of partial hypotheses *within formats* and setting *without* and *with* communication in the auction. Table 2.2 summarizes the comparison of partial hypotheses *between formats* and setting *without* and *with* communication in the auction.

Table 2.1: Comparison of Partial Hypotheses Within Formats

			H0			**HA**	
		Basic		Comm	Basic		Comm
SMR	Total prices	P	$=$	P	P	$\neq$	P
	Efficiency	E	$=$	E	E	$\neq$	E
	Auctioneer's revenue	R	$=$	R	R	$\neq$	R
	Final profits	π	$=$	π	π	$\neq$	π
SMRPB	Total prices	P	$=$	P	P	$\neq$	P
	Efficiency	E	$=$	E	E	$\neq$	E
	Auctioneer's revenue	R	$=$	R	R	$\neq$	R
	Final profits	π	$=$	π	π	$\neq$	π

Note: Basic - without communication; Comm - with communication

2.3.1 Outcome Expectations

Based on the previous literature, we hypothesized that the experimental parameters would have the following outcomes within the basic and communication treatments for both SMR and SMRPB auction formats: we expected total

Table 2.2: Comparison of Partial Hypotheses Between Formats

			H0			**HA**	
		SMR		SMRPB	SMR		SMRPB
Basic	Total prices	P	$=$	P	P	$\neq$	P
	Efficiency	E	$=$	E	E	$\neq$	E
	Auctioneer's revenue	R	$=$	R	R	$\neq$	R
	Final profits	π	$=$	π	π	$\neq$	π
Comm	Total prices	P	$=$	P	P	$\neq$	P
	Efficiency	E	$=$	E	E	$\neq$	E
	Auctioneer's revenue	R	$=$	R	R	$\neq$	R
	Final profits	π	$=$	π	π	$\neq$	π

Note: Basic - without communication; Comm - with communication

prices and the auctioneer's revenue to be higher in the baseline treatment compared to the cases when communication is allowed. We further expected the communication to increase the final profits of bidders in both formats (as e.g. in Agranov & Yariv 2015), where a shift occurred in rent distribution from auctioneer revenues to bidders' surpluses due to the stable collusion equilibrium. In the extreme case of coordinated collusion we expected prices to stay at the reserve base, similarly as in Valley (1995) in the case of a double oral auction. Our expectations regarding the efficiency of basic versus communication treatments were ambiguous. It would be reasonable to expect that the artificial market in which communication between players is naturally forbidden should be generally more efficient than a version in which the communication is enabled. However, cheap-talk between auction bidders can result in an allocation closer to the equilibrium, i.e. allocating the goods to the players with the highest valuations, and therefore in higher efficiency compared to the basic treatment. Table 2.3 describes expectations about outcomes within the formats of our experiment.

Moreover, we were interested in whether allowing for combinatorial bidding in SMRPB auction would break collusion. The SMRPB format would in that case increase competition among bidders. Prices would go up, approaching a more competitive level from the SMR basic treatment. Respective changes in rents distribution would appear.

We hypothesized that the experimental parameters in terms of comparisons between SMR and SMRPB formats would have very similar outcomes for both basic and communication treatments. Based on the results provided by Goeree & Holt (2010) we expected total prices, efficiency, and the auctioneer's revenue to be higher in the basic combinatorial SMRPB treatment than in basic SMR. We expected the final profits of bidders to end up lower in the basic SMRPB

Table 2.3: Outcome Expectations Within Formats

		Basic		Comm
SMR	Total prices	P	$>$	P
	Efficiency	E	?	E
	Auctioneer's revenue	R	$>$	R
	Final profits	π	$<$	π
SMRPB	Total prices	P	$>$	P
	Efficiency	E	?	E
	Auctioneer's revenue	R	$>$	R
	Final profits	π	$<$	π

Note: Basic - without communication; Comm - with communication

than in basic SMR treatment. We expected the same outcomes in terms of total prices, auctioneer's revenues and final profit in cases of treatments introducing communication; that is, total prices and auctioneer's revenues would end up higher in SMRPB with communication than in SMR with communication and final profits of bidders would be lower in the case of SMRPB with communication than in the case of SMR with communication. Our expectations regarding the comparison of efficiency of SMR with communication and SMRPB with communication was again ambiguous. We employed a setting with high complementarities among goods, which should favor the combinatorial SMRPB format over SMR in efficiency in basic treatment (Goeree & Holt 2010). The efficiency in communication treatment was, however, uncertain. Table 2.4 describes our expectations about outcomes between the formats of our experiment.

Table 2.4: Outcome Expectations Between Formats

		SMR		SMRPB
Basic	Total prices	P	$<$	P
	Efficiency	E	$<$	E
	Auctioneer's revenue	R	$<$	R
	Final profits	π	$>$	π
Comm	Total prices	P	$<$	P
	Efficiency	E	?	E
	Auctioneer's revenue	R	$<$	R
	Final profits	π	$>$	π

Note: Basic - without communication; Comm - with communication

2.4 Methodology

2.4.1 Experimental Design

In a fully computerized laboratory experiment, we employ the simultaneous multi-round auction format (SMR) and compare it to its combinatorial version, the simultaneous multi-round package bidding (SMRPB) format, to see the effect of package bidding on efficiency and revenues, and to evaluate the original policy format with its most natural extension. Next, we incorporate the dimension of communication by implementing a simple chat window into both SMR and SMRPB auction formats and thereby allow for coordinated strategies in the experiment. The treatment matrix of the experimental design is shown in Table 2.5.

Table 2.5: Treatment Matrix

	SMR	**SMRPB**
I.set	no communication	no communication
II.set	communication all	communication all

The design is broadly inspired by the specifications of the 2013 Czech Spectrum Auction. Auctioning of the radio spectrum in the Czech Republic was executed in 2012 and 2013 through a standard Simultaneous Multi-Round auction (CTO 2012b; 2013d). The former Czech telecommunications market was comprised of three incumbents earning excessive profits and attracting strong suspicions of collusion (CTO 2012a). Two auctions were eventually conducted, since the first auction was terminated by the regulator[4] due to ”*unrealistically high bids*” at 20 billion CZK, which was 2.3-times the total reserve prices for all the goods (CTO 2013a, p. 1). Even though a potential newcomer participated in the first auction, the distribution of market power remained the same; only three original incumbents participated in the second auction. No competition seemed to be present as bidders split the objects for sale almost at the reserve prices, resulting in low final prices and revenues for the auctioneer (CTO 2013b;c). Moreover, all three market incumbents were accused of tacit collusion by the CTO in 2012 (CTO 2012a). The collusion was ultimately unproven (Bányaiová 2012) but the relevant market was not effectively competitive (CTO 2012a). Furthermore, the Czech telecommunications market did not appear competitive even several years later. The European Commission

[4] The Czech Telecommunications Office - CTO

started a formal investigation of network sharing in the Czech Republic in 2016 (European Commission 2016) and the Czech Office for the Protection of Competition announced that it would start an inquiry into mobile operators at the end of March 2017 (CTO 2017). The Czech Spectrum Auctions may therefore serve as a good representative of a simultaneous auction with a high suspicion of collusion.

Researchers in experimental economics commonly refer to real life situations when conducting experiments on combinatorial auction formats. Abbink *et al.* (2005) explore design alternatives for the British 3G/UMTS auction using two symmetric tetrads of bidders in their experiments; and more recently Bichler *et al.* (2013) use a band plan with two bands of blocks, which can be found in several European countries, in their base value model. In our design we take the number and types of goods auctioned, the number of players, and the rules of the SMR auction from the 2013 Czech Auction setting, while the valuation is modelled according to the literature. We do not claim that our parametric setting is either the only one or the best one possible. Rather it was a readily available representation of a real life situation with a remarkable development.

The parameters were further adjusted and simplified in order to be applicable to the experimental design.[5]

There were four heterogeneous types of goods in our experiment; each type (A; B; C; D) had multiple homogeneous units in stock (6; 24; 14; 9), respectively. Each of the four players who competed in the tender was assigned her own personal valuations for each type of the goods sold in the auction. These were determined randomly by the procedure specified below. At the end of the auction, players either earned profit or incured losses in the experiment.

2.4.2 Basic Auction Formats

The SMR auction format is a generalization of the ascending English auction designed for a simultaneous allocation of multiple objects. The auction proceeds in a sequence of rounds in which the bidders submit their bids separately for individual items. The process continues until nobody is willing to submit a higher bid for any item. The SRMPB is a combinatorial auction format originally designed to prevent the exposure risk of bidders. The provisional winning bids in each round are calculated according to maximization of the revenues

[5] Table 2.6 describes the actual parameters used in the experiment.

for the seller. Each bidder can have only one provisionally winning bid in each round at most (Brunner *et al.* 2010).[6]

Each bidder is eligible to act in only a limited number of possible actions during each round. This number of actions is constrained by the amount of activity points at her disposal. This rule ensures that if the bidder wants to play seriously and to win her share of the desired goods at the end, she must maintain the activity throughout the entire auction, rather than acting in the last round. Otherwise, the activity points are lost and the player is not eligible to bid in the subsequent rounds. The volume of activity points each player has at her disposal at the beginning of the auction represents her budget constraint. When implementing the activity rule, we follow the procedures used, for example, in Cramton (2013) and Brunner *et al.* (2010).

In order to prevent signalling via the determination of prices, a simple system of proportional ascending bidding was introduced into the simultaneous multi-round auctions. There is a one-level raising algorithm in the program used for our experiment. Bidders can either keep their previous bid or raise their bid for some goods by 20% of the respective reserve price.[7] Bidders can withdraw their provisionally winning bids in cases in which they would win an unwanted (e.g. incomplete) set of goods in an auction round. All bidders can withdraw their provisionally winning bids in at most 2 rounds of the auction.

The winner determination algorithm (WDA) in the case of excess demand is, in both SMR and SMRPB formats, provided by a random mechanism. All four types of goods are handled separately in SMR, while in the SMRPB the items are handled in packages.[8] At the end of each round, bidders receive information about the provisionally winning bids in the current round. The identity of the provisionally winning bidders is known. The bidders also have complete information about their own bids.

[6] An exclusive XOR (logical exclusive-or) rule is imposed on bids made in the auction rounds.

[7] Even though we are aware that setting the level at 20% was rather too high, we could not make it lower since the whole task would have taken too long in the experimental sessions and therefore the experiment would not have been feasible.

[8] The winner determination algorithm applied in the program is simple. Each player involved in the problem is assigned a random number. The player with the smallest number wins and is allocated the item. There is a mechanism sorting the packages according to their highest price in the SMRPB. The loser resulting from WDA in SMRPB auction is put into a subsequent place and if she satisfies the conditions for winning her package out of the remaining goods, she wins it. This process continues down to the bottom if needed.

2.4.3 Communication

The communication channel is introduced via a simple chat window through which the multilateral undirected communication is implemented. No verbal contact between participants was allowed during the experiment. All communication was monitored and recorded. This approach has already been used in, for example, Phillips *et al.* (2003).

When implementing communication, we were inspired by Phillips *et al.* (2003); Lopomo *et al.* (2005); and Miralles (2010). Two minutes for communication were provided prior to each auction, similarly as in Miralles (2010). Since the number of objects in the auction is large, the comparative cheap-talk equilibrium should, according to Miralles (2010), exist. The chat window was also available during the whole auction phase as in Phillips *et al.* (2003).

Analogously to Lopomo *et al.* (2005), there was only a limited amount of information available to the bidders in the pre-auction phase of communication. The bidders did not know their exact valuations for goods and therefore were able to communicate only on the collusive mechanisms they could employ, not directly on their own private values. There was, however, no ex post budget balancing allowed in our experiment; the effect of collusion on the auction efficiency is therefore ambiguous beforehand. The only information revealed to the participants during the pre-auction communication phase was the types and numbers of auctioned goods. All three features facilitating collusion from Phillips *et al.* (2003) were therefore satisfied.

2.4.4 Valuations and Complementarities

There were four types of players in the experiment. Such a market structure can be found in various countries and industries, and was modelled, for example, in experiments by Bichler *et al.* (2013) or by Abbink *et al.* (2005) who used two symmetric tetrads of bidders in their experiments. The valuations of goods for each player-type were randomly drawn from a publicly known interval using the uniform distribution (the same interval for all players, so there was an ex-ante symmetric setting for each auctioned good) only once prior to the experiment. Then, in each session, each experimental subject was randomly assigned to one player type. This prevented any additional external variation between the treatments possibly caused by random draws of valuations on place.

The valuations of goods were based on two components. The first represents the common-value component (CVC), while the second represents the private-value component (PVC) of each particular unit of goods. The common value arises from the overall market potential and is the same for all players, while the private value stems from private expected profits depending on the individual potential of the bidder's business concept (Abbink *et al.* 2005). The bidder's total valuation of goods was therefore the sum of her CVC and PVC, as in Abbink *et al.* (2005).

The CVC of the signal was randomly drawn from the integer interval for each type of goods. Players did not have information about the exact random draw of CVC, nor did they know the interval boundaries from which it was drawn. Each bidder received instead an independent private signal on the CVC and was informed about the fact that these signals were determined by uniform random draws from the integer interval $[CVC - \alpha; CVC + \alpha]$, as in Abbink *et al.* (2005, p. 511).[9]

The PVC of the signal was randomly drawn from the integer interval $[-\beta; +\beta]$ for each type of goods. The parameter β was proportionally lower to the CVC component and corresponded to one-tenth of the common value component, i.e. $0.1 \cdot CVC$. Each bidder was informed of her own true PVC.

When modeling the complementarities in player-type valuations of goods, we followed Brunner *et al.* (2010). The interrelations among goods are modeled in a linear manner. If the player acquires multiple goods, then the value of each unit of goods increases by a factor of $[1 + \gamma(K - 1)]$, where K stands for the number of types of goods won and the γ is the synergy factor. The player should, therefore, be motivated to win all four types of goods.

Since we assume a high level of complementarities among the types of goods, we set the synergy factor γ equal to 0.1. This setup ensures that if a bidder wins all four types of goods, his valuation of all of them rises by 30%.

2.4.5 Efficiency Measurement

We measure and compare the efficiency levels of individual auction formats with different collusive properties. We use a simple measure of efficiency based on the actual surplus attained by all bidders in an auction (S_{act}) and divided by the maximum possible surplus (S_{opt}) of an auction with given parameters as did Brunner *et al.* (2010). This measure should be comparable across the

[9] The exact intervals for CVC in Abbink *et al.* (2005) were $[1000; 1500]$ for the CVC interval and $[CVC-200; CVC+200]$ for the independent private signal known to the bidders. We employ quite a similar setting in our experiment.

four treatment cells, since our value draws are constant across all player types and treatments.

$$S_{act} = \sum_{i=1}^{n} \sum_{j=1}^{m} q_j^i * (CVC_j^i + PVC_j^i) * (1 + \gamma * (K^i - 1)) \qquad (2.1)$$

where i is the number of players ($n = 4$), j is the number of types of goods in the auction ($m = 4$), CVC, PVC are common and private value components, $K^i = \sum_{j=1}^{m} q_j,\ \forall q_j > 0$ is the number of types of goods won and γ is the synergy factor equal to 0.1, both from Subsection 2.4.4.

The maximum possible surplus is found by solving a constrained maximization $S_{opt} = max\ S_{act}(q)$, subject to the number of goods sold in each type, complementarities among goods and players, and the activity limit each player has at her disposal. The optimal allocation resulted in the surplus $S_{opt} = 19145.1$.[10] The efficiency measure is then calculated as follows:

$$E = \frac{S_{act}}{S_{opt}} \qquad (2.2)$$

2.4.6 Revenues, Average Prices and Final Profits

We define auctioneer revenues as the sum of final prices for all individual goods sold in the auction. Average prices are calculated as the average price paid for one sold item in a given auction treatment. This variable therefore indicates a price level that is reached in a respective treatment.

The final profit was determined for each player at the end of the auction by the difference of her total valuation for all goods won in the auction and the total price paid for those goods. There was no endowment assigned to the players since the budget constraint is irrelevant in the experimental design. Bidders either earned profits or incurred losses in the experiment.

2.4.7 Number of Goods Sold in the Auction

In order to compare how many goods out of the total were sold in an auction, we set up an auxiliary index θ. We define the index as a weighted percentage

[10] The optimal allocation is provided in the appendix to this paper in Table A.1.1

share of goods sold in the auction in order to take into account that some types of goods are more valuable than others:

$$\theta = \left(\frac{\sum_{j=1}^{m} A_j}{A_{total}} + \frac{\sum_{j=1}^{m} B_j}{B_{total}} + \frac{\sum_{j=1}^{m} C_j}{C_{total}} + \frac{\sum_{j=1}^{m} D_j}{D_{total}} \right) \cdot \frac{1}{n}, \in < 0; 1 > \tag{2.3}$$

where A, B, C, D are individual types of goods, n is the number of players in the auction and *total* is the maximum total amount of a given type of good that is available in the auction. Following the set-up of the 2013 Czech Spectrum Auction, the goods that were not sold during the final round of the auction are not distributed to the remaining players.

2.4.8 Parametrization

We estimate a common value component of each particular type of goods as its reserve price multiplied by a parameter $\delta = 1.65$ calculated according to the Equation 2.4.

$$\delta = \frac{B_{upp} - B_{low}}{2} \cdot \frac{1}{B_{low}} + 1, \tag{2.4}$$

where B_{upp} and B_{low} are upper and lower bounds of the interval in which the totals for all goods in the auction fluctuated. The B_{low} is therefore equal to the sum of all reserve prices of all goods auctioned off and B_{upp} is the estimate of the value reached in the real auction at the external cancellation by the regulator.[11] The common value component is the same for each player. The private value component is different for each player and is determined by a random draw from the interval $[-0,1 \cdot CVC; +0,1 \cdot CVC]$.

The activity points per one unit of goods used in the experiment are determined by taking the respective activity per block in a spectrum interval from the real parameters of the 2013 Czech Spectrum Auction and rounding it up to integers. The total activity in the experiment is therefore slightly higher than in the real situation, but it is more convenient for the purpose of the experiment. Each player has her initial activity based on the $\frac{1}{4}$ of the total activity in the experiment, while the precise activity endowment is determined

[11] $B_{low} = 8.719 \cdot 10^9 CZK$; $B_{upp} = 20 \cdot 10^9 CZK$ - see also 2.4.1.

Table 2.6: Actual Parameters of Goods Used in the Experiment

Category of Goods	**A**	**B**	**C**	**D**	**Total**
Goods in stock	6	24	14	9	53
Reserve price per unit	1400	40	180	40	-
Total reserve price per category	8400	960	2520	360	12240
CVC	1820	50	140	50	-
PVC interval	±182	±5	±14	±5	-
Activity per unit	10	1	1	1	-
Total activity for category	60	24	14	9	107

Table 2.7: Final Parameters Determined by Random Draws

Players	**Valuations for goods**				**Activity**
	A	**B**	**C**	**D**	**endowment**
Type I	1886	53	138	48	27
Type II	1727	53	140	47	28
Type III	1900	47	130	46	26
Type IV	1865	48	138	50	25

by a random draw from the interval $[-3; +3]$,[12] which is added to the $\frac{1}{4}$ of total activity in the experiment.

The following tables summarize the final experimental parameters. Table 2.6 shows the actual parameters of individual categories of goods used in the experiment.[13] Table 2.7 shows the final individual valuations for goods and the final endowment of activity points of each player. Table 2.8 summarizes the common value component and its private signal intervals.

Table 2.8: Common Value Component and Private Signal Intervals

Goods	**CVC**	**CVC variance**	**CVC private signal interval**
A	1820	200	[1620 ; 2020]
B	50	5	[45 ; 55]
C	140	10	[130 ; 150]
D	50	5	[45 ; 55]

[12] Each tail of this interval represents a rounded 10% of the $\frac{1}{4}$ of total activity in the experiment.

[13] Blocks A3 and B1 from the 2013 Czech Spectrum Auctions were adjusted compared to the original settings in order to be homogeneous with other blocks in respective categories. One specific real block A3 of 2x10 MHz was split into two blocks of 2x5 MHz, which is in accordance with other category A blocks. One specific real block B1 of 2x15 MHz was split into fifteen blocks of 2x1 MHz, which is in accordance with the other category B block. The reserve prices, activity points per block etc. were also homogenized according to this principle. The setting with the original price vector had to be changed after the pilot experiment since the whole task would have been too long and therefore unfeasible. To accommodate for this, the reserve prices were multiplied by the coefficient 1.3.

2.4.9 General Procedure of the Experiment

We conducted a computerized laboratory experiment with four experimental sessions. We engaged 24 subjects per session, which resulted in 96 subjects in total. The experiment was performed in the Laboratory of Experimental Economics at University of Economics in Prague[14] and was computerized using the Z-TREE program (Fischbacher 2007).

The subjects of the experiment were invited through the ORSEE system of the Laboratory of Experimental Economics (Greiner 2004). Additional criteria were imposed on the selected subject pool in order to ensure that they would understand the task and would be capable of taking part in the experiment; we gave preference to economics majors with previous experience in auction experiments. The experiment was conducted in Czech.

Subjects were paid according to their performance in the experimental treatment. Each treatment lasted approximately two hours and the average pay for the whole treatment was expected to be on average 500 CZK[15] per subject, which was above the students' regular hourly wage rate. Prior to the experiment itself, we ran a pilot-version to test the structure of the experiment and the functioning of the programs, and to calibrate the tasks.

Instruction Procedures

The complexity of the required task to be done in the laboratory was expected to be highly demanding. We were not able to train subjects specifically before the experiment or to carry out the complicated procedures used for example in Abbink *et al.* (2005); Brunner *et al.* (2010) or even Bichler *et al.* (2013). This was mainly due to the necessity of high over-recruitment rates in the case of such training and the tightly constrained funding of the research. Therefore, we used a simpler procedure.

The participants received an invitation five days prior to the experiment and three days prior were asked to fill in an online questionnaire based on the partial instructions available online. This online material consisted of general instructions common to all treatments of the experiment. The instructions were concluded with a 5-question quiz. Those who filled in the questionnaire correctly were preferably invited to the lab. There were no difficulties with the

[14] (LEE at VŠE); `www.lee-vse.cz/eng`

[15] The resulting levels of competition during the experimental auctions and random draws determining the treatments for payments resulted in a lower average payoff than 500 CZK: on average 400 CZK per subject.

online questionnaires, and the rate of successful completion was over 95%. The whole procedure regarding the instructions in advance and the questionnaire was described in the invitation email for the experiment and was therefore publicly known.

General Procedure

For each session, a group of 24 participants came to the lab and randomly drew seat numbers. Each subject was seated at the respective computer station with no possibility to see anybody else's screen or to talk to each other. This rule was strictly enforced for the entire experiment. The participants were provided with (i) a set of written general instructions for the experiment (the same set they had already seen in the online questionnaire); (ii) treatment-specific supplement to the instructions;[16] (iii) consent form;[17] (iv) pencil and blank sheet of paper for notes. Participants had 15 minutes for self-study of the instructions when they arrived at the laboratory. A computerized questionnaire with several control questions was launched for all subjects after this time expired. A practice auction round was conducted in order to be sure subjects understood the experimental interface, how to read their parameters, how to enter bids on the screen, and that they were acquainted with the auction procedures.

To prevent misunderstandings and make the task easier, only one type of auction (one treatment) was performed in each session, i.e. a between subject design was used. In each session, each participant was randomly assigned to one of four player types, which remained stable across the whole session. This ensured that no additional external variation caused by random draws was present. The players were then randomly assigned to groups. Each group consisted of four players, each of a different type. Each player took part only in one auction format, while there were three auctions performed within the session and therefore within the auction format. Before the three real auctions, a practice auction was conducted to ensure understanding of the task and to accustom the subjects to the auction interface. All groups were randomly re-matched with the condition of stranger matching at the beginning of each auction within one session.

There was a predefined exchange rate of experimental currency units (ECU)

[16] English instructions are published in Section A.3 of the Appendix, complete sets of instructions are available at `http://ies.fsv.cuni.cz/cs/staff/matousek`

[17] If the participant refused to give consent for the experiment, she was paid the show-up fee and sent away.

and real money in the experiment of 1 CZK for 3 ECU. Participants knew this exchange rate in advance from the instructions. The payment from the experiment was not aggregated over all auctions executed in a session, but rather depended on one specific session round determined by a random draw.

When the participants accomplished the experimental task and the auction was over, they were called separately to an adjacent room, where they were paid in private and then left.

Experimental Task

Subjects participated in the auctions within the treatments they attended. The objective of the task was to win the desired goods in the auction and to gain a profit, which was then converted to real money at the end of the session. At the very beginning of each auction a chat window was displayed for two minutes in the two treatments with communication. After the chat window, a screen with experimental parameters was shown for one minute in all treatments. Then the first auction round began. The auction itself progressed in a series of simultaneous rounds where players were bidding for the collections of goods of their interest. Bidding was accomplished by adding the goods to the bidding basket. Players could submit their baskets within the auction round time limit of two minutes. There was an auction interface with parameters for all goods; bidding basket; the player's personal account; history of past rounds and, in respective treatments, also a chat window displayed on the auction round screen.

After all players had submitted their bids, the system executed all background tasks and the summary of the auction round was displayed. Each player received complete information about her resulting situation in the current auction round and from the previous auction round (her provisionally winning goods). The history of past rounds and, in respective treatments, also a chat window were displayed on the summary round screen. There was a button to open the bid withdrawal interface implemented in this stage of the round. By entering this interface, players could withdraw any of their provisionally winning goods in SMR.[18] When the one minute time limit for the summary phase ran out or when all players clicked the proceed button, the next round began. The whole process was repeated until the final round of the auction in which no player submitted any higher bid.

[18] The whole package in SMRPB treatments.

Table 2.9: Sample Statistics Between Treatments

Variable	**SMR**	**SMR Comm**	**SMRPB**	**SMRPB Comm**
θ	0.2767	0.6436	0.2528	0.3760
	(0.1739)	(0.1940)	(0.2062)	(0.2151)
AVG prices	579.90	451.90	598.38	557.52
	(296.46)	(441.40)	(533.15)	(388.60)
Efficiency	0.34	0.68	0.30	0.49
	(0.21)	(0.18)	(0.22)	(0.23)
Auctioneer's rev.	5865.11	10689.78	5783.56	7565.11
	(3414.89)	(2435.19)	(4080.25)	(3031.87)
Final profit	174.57	597.00	-23.38	500.09
	(356.86)	(416.17)	(361.96)	(669.71)

Displayed: mean (standard deviation)
Note: SMR/SMRPB - without communication; Comm - with communication

2.5 Results

Table 2.9 shows sample statistics of the most important variables between treatments. The results of three out of four auction treatments (SMR, SMR Comm and SMRPB Comm) are different from each other. There is little difference between basic SMR and SMRPB. Generally, treatments allowing for communication among players score better in almost all experimental parameters than their non-collusive counterparts.

Table 2.10 summarizes the results of partial hypotheses *between formats* and in the setting *without* (Basic) and *with* (Comm) communication allowed in the auction. A comparison of basic formats without the communication channel does not allow rejection of the null hypotheses, and shows that both formats are statistically identical[19] in terms of the number of goods sold in the auction represented by index θ ($p = 0.20$), average prices paid by the bidders ($p = 0.20$), Efficiency ($p = 0.48$) and Auctioneer's revenues ($p = 0.94$).[20] However, a difference arises in the average final profits gained by the players of those treatments ($p = 0.02$). Comparing the SMR and SMRPB formats both with communication generally favors the SMR format without package bidding. Statistical differences arise in index θ ($p = 0.00$), average prices paid by the bidders ($p = 0.00$), Efficiency ($p = 0.00$) and Auctioneer's revenues ($p = 0.00$). The final profits of bidders are slightly higher for non-combinatorial SMR with marginal significance ($p = 0.10$).

Table 2.11 summarizes the results of partial hypotheses *within formats*

[19] Although SMR performs slightly better.

[20] Since the Shapiro-Wilk test rejected its null hypothesis about the normal distribution of the data for all variables, we use the Wilcoxon-Mann-Whitney test for the analysis of variable differences (W-M-W test).

Table 2.10: Comparison of Partial Hypotheses Between Formats - Results

		Means			**p-value**	**Z-stat**
		SMR		SMRPB		
Basic	θ	0.2767	=	0.2528	0.2008	-1.279
	AVG prices	579.90	=	598.38	0.2048	-1.268
	Efficiency	0.34	=	0.30	0.4817	-0.704
	Auctioneer's rev.	5865.11	=	5783.56	0.9490	0.064
	Final profits	174.57	$\neq$	-23.38	0.0242	-2.255
Comm	θ	0.6436	$\neq$	0.3760	0.0000	-6.461
	AVG prices	451.90	$\neq$	557.52	0.0000	4.093
	Efficiency	0.68	$\neq$	0.49	0.0000	-4.157
	Auctioneer's rev.	10689.78	$\neq$	7565.11	0.0000	-6.653
	Final profits	597.00	=	500.09	0.1024	-1.633

Note: Basic - without communication; Comm - with communication

and in the setting *without* (Basic) and *with* (Comm) communication allowed in the auction. Comparing the two basic formats against their representatives with communication channels resolutely favors the collusive treatments. When comparing the SMR format, statistical differences arise in the number of goods represented by index θ ($p = 0.00$), average prices ($p = 0.00$), Efficiency ($p = 0.00$), Auctioneer's revenues ($p = 0.00$) and Final profits ($p = 0.00$).Very similar results occur when we compare Basic and Communication treatments for combinatorial SMRPB format. There are statistical differences in index θ ($p = 0.00$), Efficiency ($p = 0.00$), Auctioneer's revenues ($p = 0.00$) and Final profits ($p = 0.00$) within this format. Results do not show that average prices are different within the SMRPB format ($p = 0.54$), even though its means are still quite similar.

Table 2.11: Comparison of Partial Hypotheses Within Formats - Results

		Means			**p-value**	**Z-stat**
		Basic		Comm		
SMR	θ	0.2767	$\neq$	0.6436	0.0000	-8.444
	AVG prices	579.90	$\neq$	451.90	0.0000	4.349
	Efficiency	0.34	$\neq$	0.68	0.0000	-8.122
	Auctioneer's rev.	5865.11	$\neq$	10689.78	0.0000	-8.315
	Final profits	174.57	$\neq$	2672.44	0.0000	-6.551
SMRPB	θ	0.2528	$\neq$	0.3760	0.0010	-3.294
	Total prices	1445.89	=	1891.28	0.1524	-1.431
	AVG prices	598.38	=	557.52	0.5480	0.601
	Efficiency	0.30	$\neq$	0.49	0.0000	-4.573
	Auctioneer's rev.	5783.56	$\neq$	7565.11	0.0021	-3.070
	Final profits	-23.38	$\neq$	500.09	0.0000	-4.491

Note: Basic - without communication; Comm - with communication

Treatment-specific differences were not the only influence on the results of the experiment. Not all players were of the same strength. The second type of player had only a limited chance to outplay others in terms of final profits due to the parameter setting. There appear to be significant differences in efficiency achieved among individual player types. The first type of player had a statistically higher rate of average cumulative efficiency (by 6%, 2.5 times more) and final profit (by 49%) than others. The situation was the opposite for the third player, since her average cumulative efficiency was significantly lower (by 6%, 7 times less) together with the total price paid (by 24.7%); in other words, the third player bought fewer goods on average

2.6 Chat Analysis

The recorded chat content was independently coded into nine categories by two researchers and the conflicts (less than 2%) were resolved by discussion. Table 2.12 reports the results of the chat content. Columns 1 and 2 provide the respective frequencies and percentages of the message contents across the two auction formats. The χ^2 test rejects the null hypothesis that no association exists between the columns ($\chi^2(8) = 94.5; p = 0.000$). Most messages sent were about collusion, specifically about building collusion, agreement on collusion, specific collusive offers, disagreement with collusion and threats to enforce collusion (32%, 16%, 7%, 5% and 2% respectively). This clearly shows that collusion was the main topic of the conversations, with the aim to split the market. Disagreement with this aim was relatively rare. Questions about and explanations of the auction rules appeared in 10% of cases, while messages indicating no understanding appeared only in 3% of messages, which suggests that, especially in the initial stage of the auction, some subjects were not sure about the whole auction procedure, but this was only a small proportion who quickly learned, and overall the rules were then well understood. Interestingly, the revelation of private attributes or self-identification was rare, happening in only 6% of cases. Sophisticated subjects could, e.g., reveal their real identities and form a collusive agreement for splitting the total payoff after the experiment, or first choose a general collusive mechanism and then act according to their private values. The data shows that these sophisticated strategies were again rather rare, with only one person revealing their real identity.

Table 2.12: Chat Content

	Type of auction								
	SMR			SMRPB			Total		
Message content	No.	Col %	Cum %	No.	Col %	Cum %	No.	Col %	Cum %
Building of collusion strategy	220	24.9	24.9	380	39.3	39.3	600	32.4	32.4
Specific collusion deal offer	87	9.8	34.7	47	4.9	44.1	134	7.2	39.6
Agreement on collusion	169	19.1	53.8	124	12.8	56.9	293	15.8	55.5
Questions on / explanation of auction rules	108	12.2	66.1	86	8.9	65.8	194	10.5	65.9
Threats/collusion is broken	11	1.2	67.3	32	3.3	69.1	43	2.3	68.3
Self identification - parameter values, profit, identity	59	6.7	74.0	51	5.3	74.4	110	5.9	74.2
No understanding of collusive stategy	31	3.5	77.5	19	2.0	76.3	50	2.7	76.9
Disagreement with collusion, non-cooperation	25	2.8	80.3	68	7.0	83.4	93	5.0	81.9
Other, not related to auction	174	19.7	100.0	161	16.6	100.0	335	18.1	100.0
Total	884	100.0		968	100.0		1852	100.0	

2.7 Discussion

Our results generally evince different outcomes than we originally hypothesized. Table 2.13 summarizes the total outcomes of our experiments with regard to our prior expectations between formats, i.e. comparing SMR Basic versus SMRPB Basic treatments and SMR Comm versus SMRPB Comm. Our results do not confirm those of Goeree & Holt (2010), who reported that combinatorial SMRPB format should outperform basic SMR in an environment with high complementarities among goods. Both our auction formats end up with statistically the same results in terms of average prices, efficiency, and also the auctioneer's revenue. When communication plays into the setting, the results are even contradictory. The simpler SMR performs significantly better. Allowing bidders to communicate about their actions resulted in a better allocation of goods and therefore also higher efficiency of treatments with collusive agreements.

When we compare basic treatments across auction formats, i.e. SMR Basic vs. SMRPB Basic, we do not see any major differences. Approximately the same number of goods was sold in both treatments, as weaker bidders fell behind players with higher valuations. The revenues for the auctioneer are therefore generally low in comparison to the treatments with communication. We ascribe this fact to the high competition in these treatments (as compared to those allowing for communication) that favors only strong bidders. Moreover, the overall level of efficiency achieved in our parametric setting is largely lower due to the absence of a revenue maximizing mechanism that would aim to allocate as many units of goods as possible, e.g. as in Brunner *et al.* (2010). In case of an excess supply at the end of the auction, an algorithm to allocate leftover goods based on the history of the auction was initialised in their design and if not successful, the auction clock was even restarted for unsold goods to allow the bidders to compete for those goods again. Any Pareto-optimal mechanism could have been employed in the real 2013 Czech Spectrum Auction to reach a higher overall level of revenues, and in our case consequently also the efficiency of the auction.

Comparison of the treatments with communication between formats, i.e. SMR Comm vs. SMRPB Comm, results in much better allocation in SMR than in SMRPB. The communication among players in the strategically simpler SMR results in a better allocation of goods. More goods were sold in the auction, resulting in higher efficiency. Many cases of coordinated collusion appear during both auction treatments (32.4%), in which prices frequently remain very low or even at the reserve base. This fact supports the results of Valley (1995) in the case of a double oral auction. Players usually set a collusive agreement on splitting the goods in some way that is favorable for all. In the SMRPB treatment, many strong players try to divert from these agreements in an attempt to win more and gain higher profits. Prices gradually rise as the agreements are broken and weaker players again fall behind, thus leaving a significant portion of the supply unallocated. Such behaviors suggest that, even though the combinatorial auction format seems to have the potential to break collusion, without a revenue maximizing mechanism that would allocate the remaining goods, the efficiency of such auction will be generally low. This argument is underlined by the fact that overall, fewer units of goods are sold on average in the SMRPB with collusion than in the SMR with collusion.

Focusing on the overall outcomes within formats also brings different results

Table 2.13: Outcome Expectations Between Formats - Results

		SMR		SMRPB
Basic	AVG prices	P	$=$	P
	Efficiency	E	$=$	E
	Auctioneer's revenue	R	$=$	R
	Final profits	π	$>$	π
Comm	AVG prices	P	$<$	P
	Efficiency	E	$>$	E
	Auctioneer's revenue	R	$>$	R
	Final profits	π	$=$	π

Note: Basic - without communication; Comm - with communication

Table 2.14: Outcome Expectations Within Formats - Results

		Basic		Comm
SMR	AVG prices	P	$>$	P
	Efficiency	E	$<$	E
	Auctioneer's revenue	R	$<$	R
	Final profits	π	$<$	π
SMRPB	AVG prices	P	$=$	P
	Efficiency	E	$<$	E
	Auctioneer's revenue	R	$<$	R
	Final profits	π	$<$	π

Note: Basic - without communication; Comm - with communication

than were originally expected. Table 2.14 summarizes the results of Outcome Expectations Within Formats, i.e. comparing SMR Basic versus SMR Comm treatments and SMRPB Basic versus SMRPB Comm. Despite our original expectations, auctioneer's revenues were higher in the treatments with communication. Enabling bidders to talk increases not only their final profits, but also the profits of the auctioneer together with the efficiency of overall auction. A shift in the rent distribution from auctioneer to the bidders, as in Agranov & Yariv (2015), therefore does not occur. The multi-object nature of the auction with such number of goods in store leaves enough space for both bidders and the auctioneer to profit. The average prices within formats are higher in the basic treatment for SMR and results are the same in SMRPB auctions, which is in line with this argumentation.

Significant differences arise in all experimental parameters within the SMR auction format when we contrast the SMR Basic and SMR Comm treatments. Allowing for communication opened a number of (non-binding) possibilities for players to collude. They were able to make an agreement and split the goods at stake among themselves. Successful collusive agreements resulted in one of the best allocations of goods among all experimental treatments: 34 items were sold[21] on average in the SMR Comm treatment, resulting in the lowest average price per one item at 451.9 experimental currency units. Auctioneer's revenues reached the highest levels in SMR Comm treatment. An interesting fact about relative efficiency can be tracked by comparing SMR Basic with SMR Comm treatments. The relative efficiency is actually higher with the total number of goods sold in the auction, since the players were able to split the goods more accurately and therefore reach an allocation that is more efficient.

There is also a high degree of differences within SMRPB formats. When

[21] Out of 53 in total, calculated as $\theta * 53$

comparing SMRPB Basic with SMRPB Comm treatments and therefore allowing for communication, players did not let the prices go up. By splitting the goods in stakes, they were able to buy more goods altogether and thus to increase their profits substantially. With more goods sold at lower prices, higher revenues occurred for the auctioneer. Some players tried to divert from the collusive agreement in the package-bidding format in an attempt to win more goods and gain higher profits. Such strategies increased prices of goods and in the final consequence decreased the number of goods sold. This bidder behavior explains the difference between average prices and the number of goods sold in SMRPB Comm and in SMR Comm. Final profits resulting from the SMRPB Comm increased with respect to the Basic SMRPB, but were far lower than in the non-combinatorial SMR Comm treatment. The communication also had a positive impact on efficiency also within SMRPB formats.

2.8 Conclusion

In this paper we report on an original experimental evaluation of two complex auction mechanisms with the possibility of communication between bidders before and during the course of the auction process. We use the experimental approach because it provides a useful tool to isolate the causal effects of specific design features on the auction outcomes, which is very difficult if not impossible to do when using observational data only. Evaluating the effects of changes in design features is also especially important in the auction setting because theoretical predictions may differ from actual behavior (Holt 2005). The evaluation of auctions from observational data is relatively easily done on small and highly frequented markets, such as online auctions (Stern & Stafford 2006; Lucking-Reiley *et al.* 2007), but multi-unit auctions were specifically developed for sales of items involving substantial amounts of money, such as the licences for using a broadband spectrum, and they do not happen often enough to allow for standard quantitative evaluation (Klemperer 2004; Bichler *et al.* 2017). Another approach is the application of agent-based models and simulating the behavior of the bidders under the specified auction mechanism (Choi *et al.* 2010; Farnia *et al.* 2015). However, the behavioral patterns that stem from the communication between participants may easily create an environment too complex to be modelled by such simulations. Therefore an experimental approach with real subjects provides important results that are complementary to other methods of analysis (Samuelson 2005; Croson & Gächter 2010).

Specifically, we investigate two simultaneous auction formats: (i) the Simultaneous Multi-Round auction (SMR) format and (ii) its extension allowing for combinatorial bidding, the Simultaneous Multi-Round Package Bidding (SMRPB) auction format. The parametric setting of our experiments was inspired by the 2013 Czech Spectrum Auction. Four bidders participate in an auction for multiple heterogeneous types of goods in each experimental treatment with private and common value components. The total number of auctioned goods exceeds fifty in each auction. We study four fundamental variables in the experiment: the relative efficiency of auction formats, the average price per item paid by bidders, their final profits, and the auctioneer's revenue.

Our results show that the allocation mechanisms work in our environment much better in terms of all four experimental parameters when bidders can communicate: not only the final profits of bidders, but also the auctioneer's revenues are increased when collusion occurs during the auction. Our analysis of chat reveals that the main reason seems to be that the bidders are able to reach agreements that are profitable for all. We note an important caveat of our results: they probably stem from the specific rules and parametric setting of our artificial market. The complexity of a multidimensional space given by such a number of goods for sale provides enough space for all bidders to make substantial profits. Moreover, our experiment shows suggestive evidence that the combinatorial SMRPB auction format may break collusion among cartel members, but for the price of the winner's curse. This result is in line with, but does not strengthen, the statement of Kwasnica & Sherstyuk (2013) that the multi-unit nature of sale may facilitate collusion through splitting the objects (Kwasnica & Sherstyuk 2013, p. 475).

We contribute to the literature by experimentally evaluating the claims of Groenewegen (1994) and Goeree & Offerman (2003) that collusion via communication may actually yield allocative efficiency when the value of the auction good is composed of private and common components. This question carries tremendous practical relevance, both in developing countries where the institutional framework may not be fully developed, and the auction participants may already collude on a long-term basis, as well as in developed countries, since in the current ever-connected world the auctioneer cannot be sure that communication among the bidding participants is prevented. In contrast to Noussair & Seres (2020), who work with the second-price sealed-bid auctions, we do so for the more complicated spectral auctions.

Furthermore, we contribute to the literature on the effectiveness of the package-bidding format, as the current literature is not conclusive in whether it actually improves auction outcomes: while Brunner *et al.* (2010) argues that package bidding improves the efficiency, Goeree & Holt (2010) question the combinatorial formats because they are not computationally manageable for their participants, and Bichler *et al.* (2014) suggest that with the number of goods in stock exceeding 30, the number of possible bidding combinations is immense and makes the bidder optimization problem unacceptably difficult. Our results do not show that the package-bidding format provides significantly different efficiency from the standard SMR format, which is in contradiction to Brunner *et al.* (2010). We argue that in such complex situations, the strength of the simplicity of bidding in the SMR auction format wins: the clear and simple design of the SMR makes the decision problem of players easier and manageable in comparison to its combinatorial SMRPB counterpart. The inappropriate bidding strategies in complex combinatorial mechanisms probably do not allow for complete utilization of the allocation potential of the auction formats and therefore cause inefficiency.

Regardless of the auction format, we consider the major flaw of the 2013 Czech Spectrum Auction design to be the absence of any revenue maximizing mechanism. If such a mechanism had been employed, there would not have been such a large portion of supplied goods left unsold and therefore efficiency would have been extensively improved. We suggest the use of any such mechanism in multi unit auctions, especially those with large volumes of goods for sale, to ensure that the whole lot of auctioned goods is sold. Having this feature in our experimental design thus limits the generalizability of our results.

Our study implies the following policy recommendations: When there is a suspicion of potential collusion during the preparation of an auction, policy-makers may prefer non-combinatorial auction designs, as they may produce higher efficiency and revenues. Even though combinatorial bidding was introduced to simplify the decision making problem of bidding for bidders who are exposed to combinatorial risk, our results suggest that this might actually further complicate their decisions. This holds especially true for auctions with a high volume of heterogeneous goods for sale, and is supported not only by our research, but also by Bichler *et al.* (2013). Furthermore, Bichler *et al.* (2014) state that the efficiency of simple auction formats increases with high volumes of goods in stock. This result was not confirmed for non-communication treatments in our research, but holds true for the case where communication was present.

For further investigation we recommend different set-ups or even different designs for the allocation of multiple heterogeneous goods, where the volume of auctioned goods reaches a certain threshold, and thus the number of possible combinations becomes too complicated for participants to analyze. One way to overcome this problem may be to create pre-defined bundles of goods and to auction them as single units of goods in SMR auctions (a ”controlled” SMRPB auction). Another possibility is to auction the goods using a standard combinatorial auction, but to split their volume into several smaller bundles and to auction them sequentially. Both proposed solutions, as well as the estimation of the threshold level of the volume of goods that determines the auction sufficiently ”simple” to analyze are indeed matters for future research.

Last but not least, we suggest that auctioneers in high-stakes public auctions use an experimental evaluation of the auction design before it is implemented, as in Abbink *et al.* (2005) and Bichler *et al.* (2013), so that potential flaws are revealed and the real auction is carried out correctly. Such an evaluation would also help the auctioneer find the revenue-maximizing (or efficiency-maximizing) design alternative, depending on their preferences.

Chapter 3

Individual Discount Rates: A Meta-Analysis of Experimental Evidence

Abstract A key parameter estimated by lab and field experiments in economics is the individual discount rate—and the results vary widely. We examine the extent to which this variance can be attributed to observable differences in methods, subject pools, and potential publication bias. To address the model uncertainty inherent to such an exercise we employ Bayesian and frequentist model averaging. We obtain evidence consistent with publication bias against unintuitive results. The corrected mean annual discount rate is 0.33. Our findings also suggest that discount rates are independent across domains: people tend to be less patient when health is at stake compared to money. Negative framing is associated with more patience. Finally, the results of lab and field experiments differ systematically, and it also matters whether the experiment relies on students or uses broader samples of the population.

This paper was published in *Experimental Economics [2022, 25, pp. 318-358, DOI: 10.1007/s10683-021-09716-9]*. I am grateful to the editors, Marie Claire Villleval and Roberto A. Weber, and two anonymous referees for helpful comments. An online appendix with data and code is available at `meta-analysis.cz/discrate`. I acknowledge support from the Czech Science Foundation (grant #21-09231S) and Charles University (project Primus/17/HUM/16).

3.1 Introduction

Intertemporal trade-offs are key to a host of decision problems at both the private and public levels. For some of these decisions, it is appropriate to employ the market discount rate, which is detectable from financial time series. For others, however, we must try to recover the underlying discount rates of individuals—rates that also reflect the underlying transaction costs of borrowing money that households face (Kovacs & Larson 2008). Policies addressing climate change, particularly those underpinned by the literature on the social cost of carbon, constitute a typical example of choices for which individual discounting of future costs and benefits plays a crucial role (Tol 1999; Goulder & Stavins 2002; Fujii & Karp 2008; Anthoff *et al.* 2009).

Individual discount rates can be either observed from existing data (such as in Lawrance 1991; Dreyfus & Viscusi 1995; Warner & Pleeter 2001) or measured experimentally (Benzion *et al.* 1989; Chapman & Elstein 1995; Coller & Williams 1999; Harrison *et al.* 2010, among others). We focus on the latter: experiments. Controlled experiments provide a natural framework for exploring time discounting in both laboratory and field conditions by enabling researchers to vary the parameters in order to infer the subject's preferences. However, despite decades of work and dozens of experiments devoted to eliciting time preferences, no consensus on how to best measure discounting has emerged (Andreoni *et al.* 2015). It is safe to say that the discount rate differs across individuals and its estimates vary a great deal throughout the literature, sometimes by orders of magnitude (Coller & Williams 1999; Frederick *et al.* 2002).

In this paper we take stock of the evidence and aim to trace the differences in the reported discount rates to the design of experiments while accounting for model uncertainty. We also control for the effects of potential selective reporting, a phenomenon found to be widespread in economics and other fields (Doucouliagos & Stanley 2013; Ioannidis *et al.* 2017). Focusing on aspects related to study design, methodology, and subject pool characteristics, we collect a set of 22 explanatory variables and employ Bayesian model averaging (BMA; Raftery *et al.* 1997) and frequentist model averaging (FMA; Hansen 2007) to examine which ones matter the most for the differences among the reported estimates. Model averaging techniques estimate many regressions with various combinations of the 22 variables and then weight the models according to data fit, parsimony, and collinearity.

The closest work to our own is the meticulous meta-analysis by Imai *et al.* (2021a), who employ a similar methodology but focus on the present-bias parameter estimated using the convex time budget protocol. They find that the literature implies the present-bias parameter to lie between 0.95 and 0.97 on average and describe the sources of heterogeneity: for example, experiments that use monetary rewards tend to find little evidence of present bias. Other related recent studies include Brown *et al.* (2021), who meta-analyze the estimates of loss aversion, Imai *et al.* (2021b), who estimate the degree of publication bias in laboratory experiments in economics, and a series of important works evaluating the replicability of experiments in economics and other social sciences (Camerer *et al.* 2016; 2018; Altmejd *et al.* 2019).

Our results are consistent with the notion that selective reporting (which causes publication bias) represents an important factor in the literature. When selective reporting is present, insignificant and negative estimates are discriminated against. A zero or negative discount rate, of course, makes little sense in most contexts. Nevertheless, given sufficient noise in the experimental setup, we should sometimes observe insignificant estimates and sometimes very large positive estimates. If non-positive estimates (which are unintuitive) are discarded but large positive estimates (for which it is difficult to determine whether they are intuitive or not) are kept, harmful publication bias arises. This outcome is paradoxical because selective reporting can be beneficial at the micro level: for an individual study, it is most likely a wise choice not to build the story around negative or insignificant estimates of the discount rate. However, at the macro level, the discarding rule is asymmetrical since large estimates are typically not omitted. Our findings indicate that such publication bias is associated with exaggerating the mean reported annual discount rate from 0.33 to 0.80.

Aside from publication bias, which manifests as a correlation of the discount rate estimates with their standard errors, the differences in results seem to be caused primarily by the experimental design of discounting tasks. We find evidence in line with domain independence (defined as the low correlation between discount rates for different domains) in intertemporal choice (Loewenstein *et al.* 2003; Ubfal 2016): it matters what the experimental subjects should be patient or impatient about. Subjects are more patient with regard to money than health or more exotic contexts (such as vacations, certificates, and kisses from movie stars). The results support the hypothesis that liquidity constraints play a key role in intertemporal choice experiments (Dean & Sautmann 2021), since health and kisses from movie stars are more difficult than money to transfer over time (Bleichrodt *et al.* 2016). We also find that negative framing is associated with more patience, which corroborates the notion that anticipation of dread is important in intertemporal decisions (Harris 2012).

Our results offer three broad implications for economics experiments in general. First, it matters whether the experiment is conducted in the lab or in the field. Lab experiments yield systematically larger discount rates, indicating greater impatience. Second, the composition of the sample of experimental subjects (the subject pool) has a systematic impact on the results. Experiments working exclusively with students show less evidence for patience than experiments using mixed population samples. Taken together, these two results might question the external validity of some experiments. Third, we show that it does not matter systematically for the reported discount rates whether experiments use real or hypothetical rewards.

Three caveats of our results are in order. First, we are unlikely to cover all experiments ever conducted on the discount rate. Nevertheless, a meta-analysis does not have to collect the entire universe of available studies; it is important only to avoid selecting studies based on their results. Second, fewer than two-thirds of the collected estimates are reported together with a measure of uncertainty from which we can directly compute standard errors. We address this problem partially by resampling standard errors at the study level for observations with missing data. (Limiting our attention to the studies that explicitly report precision would not change our main results.) Third, although we control for the differences in many features of study design, experiments involve unique methodological as well as procedural details that are difficult to codify but that can cause differences in the results of individual studies. Some of these unobserved features might be correlated not only with the reported discount rate but also with the reported standard error, which might make our results concerning publication bias spurious. We partially address this problem by using study fixed effects, caliper tests, p-uniform*, and by employing the number of observations in primary studies as an instrument for the standard error.

The remainder of the paper is structured as follows. Section 3.2 reviews the basic concepts of discounted utility models and discusses the methods of discount rate elicitation. Section 3.3 describes our approach to data collection and presents an overview of our dataset. Section 3.4 examines the extent of publication bias using meta-regression and other meta-analysis techniques. Section 3.5 investigates the sources of heterogeneity in the estimated discount rates using Bayesian model averaging. Section 3.6 concludes the paper. Supplementary data, codes, statistics, and diagnostics for the BMA and robustness checks to all analyses presented in the main body are available in Section B.1, Section B.2, and online at `meta-analysis.cz/discrate`.

3.2 Estimating the Discount Rate

In this section we do not attempt to provide a comprehensive review of the methodology used to measure discounting but briefly describe the basic concepts that are necessary for the understanding of our meta-analysis. For a more detailed treatment, we refer the reader to the authoritative works by Frederick *et al.* (2002), Andersen *et al.* (2014), Cheung (2016), and Cohen *et al.* (2020).

The theory of intertemporal choice and discounting dates back to Irving Fisher's *Theory of Interest* (Fisher 1930) and Paul Samuelson's *Note on Measurement of Utility*, in which he postulated the discounted utility model (Samuelson 1937). His model was widely accepted together with its central idea of concentrating various decisions about intertemporal choice into a single parameter—the discount rate. Several modifications to the original discount function have been introduced to capture various features, such as hyperbolic (Ainslie 1975; Mazur 1984) or quasi-hyperbolic (Phelps & Pollak 1968; Laibson 1997) discounting functions.

The discounted utility model captures the time preferences of an individual—more specifically, an individual's preference for immediate utility over delayed utility, represented by her intertemporal utility function $U^t(c_t, ..., c_T)$, which can be described by the functional form presented in Equation 3.1:

$$U^t(c_t, ..., c_T) = \sum_{k=0}^{T-t} D(k) \cdot u(c_{t+k}) \ , \tag{3.1}$$

where $D(k)$ is the discount function and $u(c_{t+k})$ is an instantaneous utility function that can be interpreted as an individual's well-being in period $t+k$. The discount function $D(k)$ represents the relative weight that the individual places in period t on her well-being in period $t+k$ and encompasses parameter δ, which represents the individual's discount rate. This discount function can have different functional forms.

The standard exponential model, a well-known functional form used in the majority of practical applications, follows:

$$D^E(k) = \frac{1}{(1+\delta)^k} \ , \qquad k \geq 0 \tag{3.2}$$

where the discount rate d is $d^E(k) = \delta$. The key feature of this model is that the discount rate $d^E(k)$ is constant over time, i.e., the rate at which an individual discounts future well-being between today and tomorrow is identical to the rate at which she discounts well-being between one month from today and one month from tomorrow. In contrast, a widely documented situation in which an individual has a declining rate of time preference is described as hyperbolic discounting, which generally means that the implicit discount rate over longer time horizons is lower than the implicit discount rate over shorter time horizons. A typical case from the family of hyperbolic discounting functions proposed by Mazur (1984) is described in Equation 3.3:

$$D^H(k) = \frac{1}{1+\delta k} \ , \tag{3.3}$$

where the hyperbolic discount rate $d^H(k) = (1+\delta k)^{\frac{1}{k}} - 1$ (Andersen *et al.* 2014).[1] Phelps & Pollak (1968) further introduced a quasi-hyperbolic specification of the discount function for use in a social planner problem:

$$D^{QH}(k) = \begin{cases} 1, & \text{if} \quad k = 0 \\ \frac{\beta}{(1+\delta)^k} \ , & \text{if} \quad k > 0 \end{cases} \tag{3.4}$$

where $\beta \leq 1$ and the quasi-hyperbolic discount rate $d^{QH}(k) = \left(\frac{\beta}{(1+\delta)^k}\right)^{-\frac{1}{k}} - 1$.[2] A characteristic feature of the quasi-hyperbolic specification is the discontinuity at time $t = 0$. This specification was applied by Laibson (1997) to model individual agent behavior.

Several experimental methods are available to elicit time preferences in both laboratory and field settings, such as lotteries, choice lists, and bidding; however, there is no consensus on how to best measure discounting (Andreoni *et al.* 2015). The basic method for eliciting individual discount rates is conceptually simple—asking subjects questions about whether they prefer an amount

[1] In a hyperbolic specification, the discount rate is the value of $d^H(k)$ that solves $D^H(k) = 1/(1+d^H)^k$, i.e., the equation $1/(1+\delta k) = 1/(1+d^H)^k$.

[2] Again, in the quasi-hyperbolic specification, the discount rate is the value of $d^{QH}(k)$ that solves $D^{QH}(k) = 1/(1+d^{QH})^k$, i.e., the equation $\beta/(1+\delta)^k = 1/(1+d^{QH})^k$.

of money today (option A) or the same amount + $\$X$ tomorrow (option B). By changing X, a researcher can infer bounds for the subject's individual discount rate.[3] Experiments therefore involve a series of questions aligned in lists, such as in the classical choice list design of Coller & Williams (1999) or Harrison *et al.* (2002). Modifications to this basic method are further used to elicit preferences more precisely, such as variations in the delay between options A and B, the domain in which preferences are revealed (money, health, etc.), and the magnitude or the nature of the reward (hypothetical or real).

Several types of elicitation methods are routinely used in the experimental literature (Frederick *et al.* 2002): i) choice, ii) matching, iii) rating, and iv) pricing. The most common type of elicitation is the choice method, where subjects are presented alternative options and are asked to simply choose between them. This method provides discount rate intervals pre-generated by the experimenter rather than precise estimates of the discount rate for specific individuals. The matching method, in contrast, provides an exact inference of the individual's discount rate since she reveals her true indifference point by filling the blank field to equate two intertemporal options. In rating tasks, subjects evaluate individual options by rating their attractiveness on a predefined scale, while in pricing tasks, subjects specify their willingness to pay for individual options in which they either obtain or avoid a particular outcome. In contrast to choice and matching tasks, rating and pricing tasks allow the researcher to manipulate the time variable between subjects since immediate and delayed options are evaluated separately.

Each method described briefly above has its strengths and limitations. When subjects are asked to evaluate multiple options at once in a standard choice list, the earlier choices inevitably influence the choices made later. This procedural limitation—the anchoring effect—can be partially addressed by employing titration procedures and exposing subjects to a sequence of different opposing anchors (Frederick *et al.* 2002). The timing of an outcome was found to have a much lower effect when evaluating a single option compared to a situation when two options occurring in different times are evaluated against each other at once (Loewenstein 1987). The timing of two evaluating options is further argued to cause the more general problem of an additional risk or transaction costs imposed on the future option. The recent literature, repre-

[3] The point of the first switch to option B gives a measure of the upper bound of her discount rate. We assume linear utility here for simplicity and discuss relaxing of this assumption later.

sented by Harrison *et al.* (2005), Andersen *et al.* (2014), and others, deal with this risk by employing a front-end delay, thereby shifting the immediate option to the nearer future and imposing transaction costs on the instant payoff.

Harrison *et al.* (2005) argue that standard choice tasks often executed through multiple price lists (MPL) have three possible disadvantages: i) they elicit only interval responses; ii) they allow subjects to switch back and forth while moving down the list; and iii) they can be subject to framing effects. Harrison *et al.* (2005) therefore introduces an *iterative Multiple Price List* (iMPL) that allows the subjects to iteratively specify their choices through refined options within an interval chosen in the last option.

The inference of discount rates from the experimental task depends on the utility function presented in the discounted utility model (Equation 3.1). This function, however, is unobserved and therefore usually assumed to be linear, generating biased estimates for individuals with non-linear utility functions (Cheung 2016). Recent papers by Andersen *et al.* (2008; 2014) use the *joint elicitation strategy* to measure time preferences by controlling for non-linear utility. Using the equivalence of utility for risk and time, these authors use a series of binary choices to infer the discount function conditional on the utility function elicited through Holt & Laury (2002)'s risk preference task. Further modifications of the design to measure time preferences by controlling for non-linear utility include, among others, the work of Laury *et al.* (2012), who interact risk with time using a lottery to be paid out with probability p_t in time t and with probability p_{t+k} in time $t+k$, where $p_t \leq p_{t+k}$ and p_{t+k} vary through the choice list. Further experiments measuring time preferences while controlling for non-linear utility are conducted by Takeuchi (2011), who employs separate choices under risk and over time using matched pairs of payoffs; Andreoni & Sprenger (2012b), Andreoni & Sprenger (2012a), and Andreoni *et al.* (2015), who examine risk and time preferences through individual elicitation methods—convex time budgets and double multiple price list tasks—and Attema *et al.* (2016), who introduce a *direct method* to measure discounting that is not dependent on the knowledge or measurement of utility.

An alternative method for inferring discount rates was devised by Chabris *et al.* (2008b), who not only derive intertemporal preferences from standard choice tasks but also adopt an approach of using response times from these choices, i.e., how long it actually takes the subjects to choose between option A and option B. The authors assume that *"subjects should take longest to decide when the two options are most similar in their discounted values"* and therefore argue that the inference from response times should, in principle, work (Chabris *et al.* 2008a, p. 7). The results of Chabris *et al.* (2008a) suggest that choice-based and response-time-based estimates are nearly identical in their setting.

3.3 The Dataset

The first step of a meta-analysis is the collection of primary studies. To this end, we search Google Scholar for the literature on discounting and then examine the references of the retrieved studies to search for other usable studies (this method is called "snowballing" in the meta-analysis context). We use Google Scholar because it provides powerful fulltext search. Specifically, we employ the following query: `discount method experiment "discount rate" OR "discount factor."` The query is designed to yield the well-known experimental studies on discounting among the first hits, while being sufficiently inclusive. We go through the first 300 studies returned by the search and examine the abstract of each paper. If the abstract suggests at least a remote possibility that the paper contains estimates of the discount rate, we download the paper and inspect it; this way we inspect 178 studies. Next, we collect the references of these studies and download the 30 papers that are most often quoted in the literature but are not returned by our baseline Google Scholar search.

We apply three inclusion criteria. Each study included in our dataset must be an experiment, either lab or field, and must report an estimate of the discount rate (or the discount factor in a way that allows re-computation to the discount rate). Next, we exclude estimates of the discount rate derived from very short delays (several hours)—these are extreme cases for which it is often difficult to find use in practice. Finally, we include only studies published in peer-reviewed journals. The major reason for the last inclusion criterion is feasibility, but we also hope that peer review sets a bar for quality. Moreover, journal articles generally contain fewer typos and other mistakes in the presentation of results compared to unpublished manuscripts

We terminate the search for studies on January 15, 2020. Our final dataset covers 56 studies comprising 927 estimates of the discount rate. Of these, 715 were reported explicitly as discount rates, and the remaining 212 estimates were reported as discount factors that we recomputed to rates according to the corresponding discounting formulas. All discount rates are annualized.

Table 3.1: Studies used in the meta-analysis

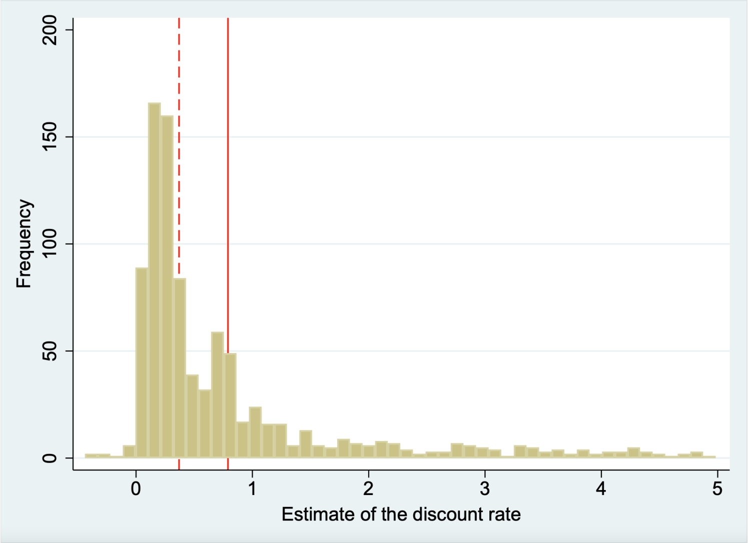
Abdellaoui *et al.* (2010)
Andersen *et al.* (2006)
Andersen *et al.* (2008)
Andersen *et al.* (2010)
Andersen *et al.* (2013)
Andersen *et al.* (2014)
Andreoni & Sprenger (2012b)
Andreoni *et al.* (2015)
Attema *et al.* (2016)
Bauer & Chytilová (2010)
Bauer & Chytilová (2013)
Bauer *et al.* (2012)
Benzion *et al.* (1989)
Booij & van Praag (2009)
Brown *et al.* (2009)
Burks *et al.* (2012)
Cairns & der Pol (1997)
Carlsson *et al.* (2012)
Cassar *et al.* (2017)
Castillo *et al.* (2011)
Chabris *et al.* (2008a)
Chabris *et al.* (2009)
Chapman & Elstein (1995)
Chapman & Winquist (1998)
Chapman (1996)
Chapman *et al.* (1999)
Chesson & Viscusi (2000)
Coller & Williams (1999)
Deck & Jahedi (2015a)
Deck & Jahedi (2015b)
Dolan & Gudex (1995)
Duquette *et al.* (2012)
Field *et al.* (2013)
Finke & Huston (2013)
Hardisty *et al.* (2013)
Harrison *et al.* (2002)
Harrison *et al.* (2010)
Hausman (1979)
Ifcher & Zarghamee (2011)
Kirby & Marakovic (1995)
Kirby & Marakovic (1996)
Kirby *et al.* (1999)
Loewenstein (1987)
McClure *et al.* (2007)
Meier & Sprenger (2010)
Meier & Sprenger (2013)
Meier & Sprenger (2015)
Newell & Siikamäki (2015)
Olivola & Wang (2016)
Read & Read (2004)
Sutter *et al.* (2013)
Tanaka *et al.* (2010)
Thaler (1981)
Voors *et al.* (2012)
Warner & Pleeter (2001)
Zauberman *et al.* (2009)

The oldest study in our sample was published in 1979,[4] and our meta-analysis thus spans four decades of research in the area. An overview of primary studies included in the meta-analysis is presented in Table 3.1; the full dataset (together with estimation codes for R and Stata) is available in an online appendix at `meta-analysis.cz/discrate`. We follow the reporting guidelines for meta-analysis compiled by Havranek *et al.* (2020).

Apart from the key variables for our analysis—the estimated discount rate and its standard error—we codify additional explanatory variables to control for the sources of variation in our data sample. We control for the length of the time horizon presented to the subjects, i.e., the delay of the experimental task. Moreover, we include a dummy variable describing whether the reported estimate relates to hyperbolic or exponential discounting. We further control for whether the study employs front-end delay; if it is performed in the lab or in the field; if payoffs used in the study are hypothetical or real, i.e., paid out at the end of the experiment; what the stakes of the experiment are in terms of the maximum payoff related to median personal expenditure; which elicitation method (choice, matching, and rating) and domain (money, health, and others) is used to identify the estimate; and whether the framing of the task is positive (gaining), negative (losing) or neutral. We also control for the characteristics of the subject pool: whether it contains students or a more general sample of the population; the gender of the subjects it includes (exclusively males, females, or

[4] The oldest paper we use is Hausman (1979), which is not an experiment in the strict sense but is still based on real choices. The paper estimates discount rates from trade-offs between upfront capital costs and future savings of operating costs, looking at purchasing decisions of air conditioners.

Figure 3.1: Histogram of discount rate estimates



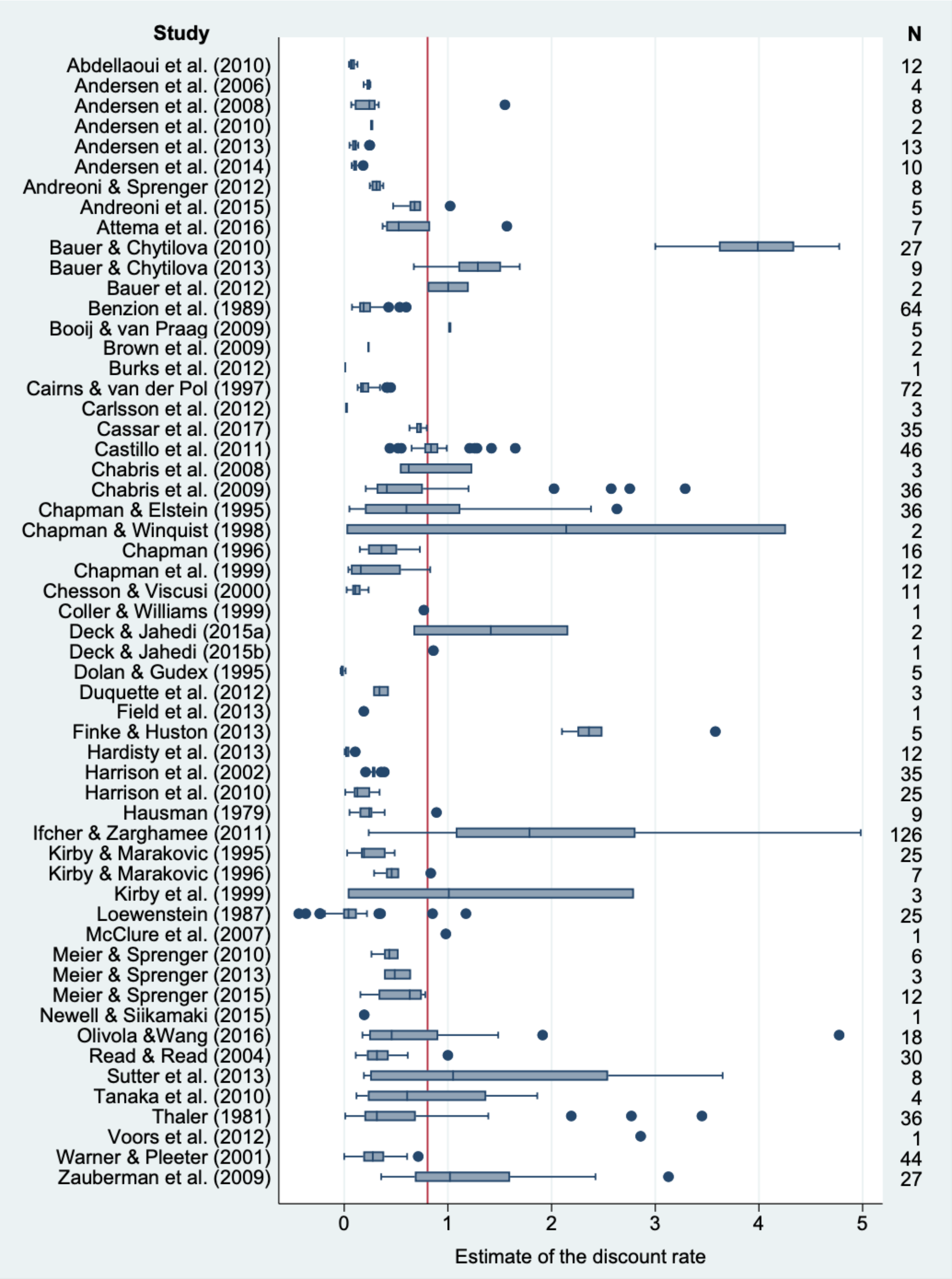
Notes: The figure depicts a histogram of annualized discount rate estimates reported by individual studies. Extreme values are omitted from the graph but included in all regressions. The solid line denotes the sample mean; the dashed line denotes the sample median.

both); and the continent from which the subject pool was drawn. Additionally, we control for study age and the number of Google Scholar citations weighted by the number of years since the first version of the study appeared in Google Scholar. We describe these variables in more detail in Section 3.5, which also includes the corresponding Bayesian model averaging analysis.

The estimated discount rates in our dataset have a mean of 0.80 and a standard deviation of 0.97. A histogram of the estimates is presented in Figure 3.1: the distribution is apparently skewed, with a median value of 0.37. Negative values of the discount rate estimates are rare, though present, and often the matter of negative framing (for example, choosing to pay a fine or experience an illness now rather than later). The distribution thus offers several outliers on both sides. We address the potential influence of these outliers on our analysis by winsorizing at the 5% level (the main results are robust to changes in the winsorization level; without winsorizing, the minimum reported discount rate is −0.4, the maximum is 13.7).

To be able to employ modern meta-analysis methods, we need measures of precision for individual estimates. Nevertheless, the standard errors of the discount rate estimates are reported only for 539 of the 927 estimates in our dataset. Researchers in the field sometimes mention that the discount rates

Figure 3.2: Within- and between-study variation of discount rate estimates



Notes: The figure shows a box plot of annualized discount rate estimates reported in individual studies. Extreme values are omitted from the graph but included in all regressions. N = the number of estimates reported in the study.

they report are large and robust to various changes in the specifications, which constitutes the implicit apology for not reporting precision. As a robustness check (available in the working paper version of this article), we exclude these studies from the dataset and focus only on those for which standard errors can be obtained directly. However, doing so reduces the power of our estimations and does not affect our main results. Therefore, in the baseline case, we also use studies that do not report precision explicitly. To approximate precision at least at the study level, we apply the bootstrap resampling technique. We then combine the explicitly reported standard errors with the standard errors obtained by bootstrapping at the study level.[5] The substantial within- and between-study heterogeneity of discount rate estimates, the rationale for a meta-regression analysis, is apparent from Figure 3.2.

3.4 Publication Bias

The selective reporting of some estimates (typically those that are intuitive and statistically significant) has been identified as a serious threat to the credibility of empirical economics (Ioannidis *et al.* 2017).[6] When estimation noise is large, and therefore standard errors are large, researchers have incentives to preferentially report large point estimates that become statistically significant. Nansen McCloskey & Ziliak (2019) liken selective reporting to the Lombard effect, in which speakers increase their vocal effort in the presence of noise. Selective reporting (which is conventionally called publication bias but is not confined to published papers) thus manifests as a correlation between point estimates and their standard errors.

The general prior among economists and psychologists is that the discount rate is positive. People are impatient; they value the present more than the

[5] Specifically, our approach follows the meta-analysis of Havranek *et al.* (2015b) on the social cost of carbon. In the social cost of carbon literature standard errors are also sometimes not reported but individual studies report many different estimates, which allows the reader to gauge the uncertainty that surrounds individual estimates within studies. For each study we use 1,000 iterations for bootstrapping so that the mean of bootstrapped values equals the mean of the estimates reported in the study. From the bootstraps we then approximate the standard error at the study level and use it for all estimates within the given study. When the standard error is explicitly reported for an estimate, we use the reported standard error.

[6] Other recent papers documenting publication bias in various fields in economics include Blanco-Perez & Brodeur (2020); Brodeur *et al.* (2016; 2020a); Campos *et al.* (2019); Doucouliagos & Paldam (2011); Duan *et al.* (2020); Geyer-Klingeberg *et al.* (2019); Havranek (2010); Havranek & Irsova (2010); Havranek & Kokes (2015); Irsova & Havranek (2010); Nelson & Moran (2020); Tokunaga & Iwasaki (2017); Ugur *et al.* (2018; 2020); Valíčkova *et al.* (2015); Xue *et al.* (2020); Zigraiova & Havránek (2016).

future. In contrast, a negative estimate of the discount rate means that an individual is willing to accept an offer in the future with a lower value than what is available now, indicating an extraordinary preference for such a state of the world. Negative (and positive but insignificant) estimates are rare in our sample but do occur, which suggests that any potential publication bias in the literature is occasional and not universal. We do not claim that the average discount rate should be zero or even negative. However, the crux of the publication bias problem is the following: with sufficient imprecision and liberal elicitation techniques, we always obtain insignificant or negative estimates from time to time. For the same reason we also obtain large positive estimates. If negative and zero findings are often discarded (they are obviously implausible), while large positive estimates are often retained (it is less obvious whether they are far from the true value), the literature as a whole presents distorted results. The typical reported estimate is biased upwards.

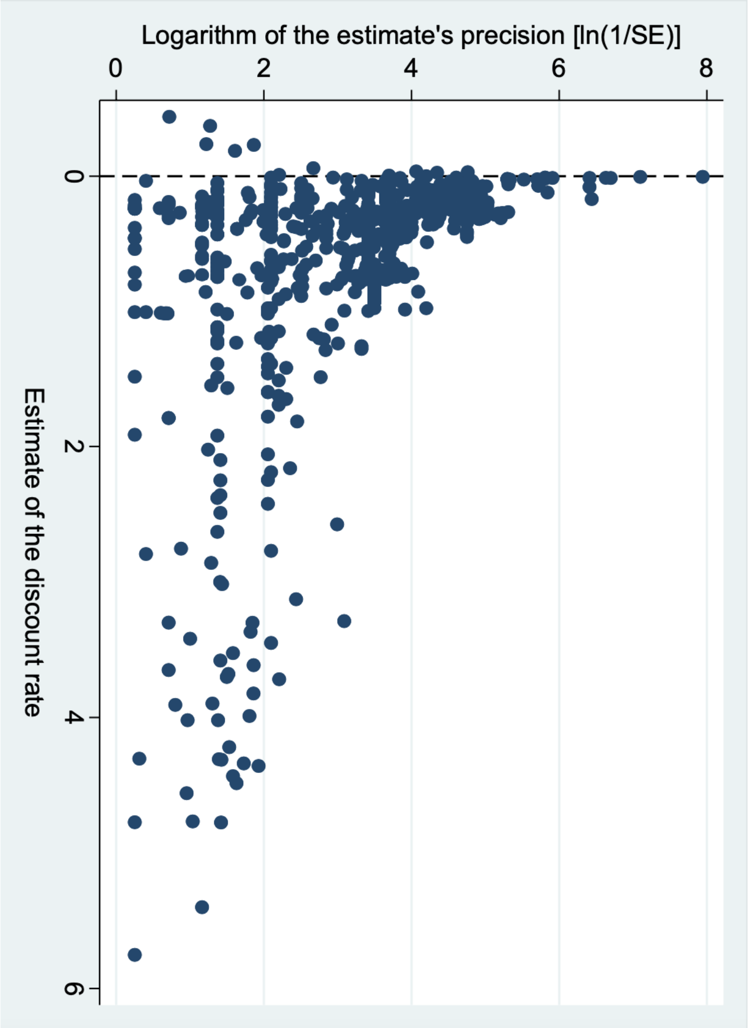
The idea of publication bias is illustrated by Figure 3.3, the so-called funnel plot (Egger *et al.* 1997). The horizontal axis depicts the magnitude of the estimate, while the vertical axis depicts the estimate's precision. With no publication bias, the most precise estimates should be close to the underlying average effect. With decreasing precision, we obtain increasing dispersion, which creates the shape of an inverted funnel. However, in the absence of publication bias, there is no reason for asymmetry in the funnel. If, in contrast, imprecise negative estimates are discarded but imprecise large positive estimates are reported, we obtain asymmetry—which is precisely what we see from the figure. The funnel plot can thus serve as a visual check of publication bias (Stanley & Doucouliagos 2010; Rusnak *et al.* 2013).

Next, we examine the correlation between the discount rate estimates and their standard errors quantitatively to test for the presence of publication bias (the so-called funnel asymmetry test, Egger *et al.* 1997):

$$\hat{\delta}_{ij} = \delta_1 + \gamma_1 \cdot SE(\hat{\delta}_{ij}) + u_{ij}. \tag{3.5}$$

Here, the $\hat{\delta}_{ij}$ is the i-th estimate of the discount rate from the j-th study, $SE(\hat{\delta}_{ij})$ is the corresponding standard error, γ_1 measures publication bias, and δ_1 is the mean discount rate corrected for the bias; u_{ij} is a disturbance term. The first part of Table 3.2 shows the results of the funnel asymmetry test; we always cluster standard errors at the study level. The first column in the table shows a simple OLS regression; the second column presents a weighted

Figure 3.3: Funnel plot suggests publication bias



Notes: The figure depicts the funnel plot of annualized discount rate estimates. Extreme values are omitted from the graph but included in all regressions.

least squares specification (with precision as the weight) which addresses the apparent heteroskedasticity of Equation 3.5.

The results presented in Panel A of Table 3.2 are consistent with the finding of publication bias: the correlation between estimates and standard errors is statistically significant at least at the 10% level in both specifications and the corrected mean is smaller than the simple uncorrected mean (0.26–0.52 vs. 0.80). But, as Stanley & Doucouliagos (2014) show, while the linear funnel asymmetry test is a valid tool for testing the presence of publication bias, it is not a good estimator of the underlying corrected mean. The reason is that selective reporting is a more complex function of the standard error, and

Table 3.2: Funnel asymmetry tests indicate publication bias

PANEL A: Linear models		
	OLS	Precision
Standard error	0.535^{***}	1.031^{**}
(*publication bias*)	(0.0299)	(0.449)
Constant	0.518^{***}	0.259^{***}
(*effect beyond bias*)	(0.114)	(0.0373)
Observations	927	927

PANEL B: Non-linear models				
	WAAP of Ioannidis *et al.* (2017)	Stem-based method of Furukawa (2021)	Selection model of Andrews & Kasy (2019)	Endogenous kink of Bom & Rachinger (2019)
Effect beyond	0.331^{***}	0.282^{***}	0.252^{***}	0.145^{***}
bias	(0.0131)	(0.00915)	(0.0140)	(0.00321)
Observations	927	927	927	927

Notes: The table reports the results of regression $\delta_{ij} = \delta_1 + \gamma_1 \cdot SE(\delta_{ij}) + u_{ij}$, where δ_{ij} denotes the i-th annualized discount rate estimated in the j-th study, and $SE(\delta_{ij})$ denotes its standard error. Panel A shows estimation by OLS and weighted least squares where estimates are weighted by precision, the inverse of their standard error. Panel B shows the recently developed non-linear estimation techniques; WAAP stands for the Weighted Average of the Adequately Powered estimates. Standard errors, clustered at the study level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Monte Carlo simulations have shown that a linear approximation does not suffice (Stanley 2008). For this reason, in Panel B of Table 3.2 we employ more advanced non-linear techniques.

The first non-linear technique presented in Table 3.2 is the Weighted Average of Adequately Powered estimates (WAAP) due to Ioannidis *et al.* (2017). The technique computes the statistical power of each estimate and uses only those whose power exceeds 80%. From these "adequately powered" estimates Ioannidis *et al.* (2017) compute a weighted average with weights proportional to the precision of the estimate. From this technique we obtain a mean discount rate of 0.33, which lies between the two estimates we obtained in Panel A (but as we have noted, estimates of the underlying effect derived from linear models in Panel A are not reliable). The second non-linear approach we use is the stem-based technique by Furukawa (2021). The "stem" in the title of the methods refers to the stem of the funnel plot; the technique focuses on the most precise estimates. It follows Stanley *et al.* (2010), who suggest that "*discarding 90% of the [most imprecise] published findings greatly reduces publication selection bias and is often more efficient than conventional summary statistics.*" (Stanley *et al.* 2010, p. 70). Instead of discarding an arbitrary portion of estimates, which is generally suboptimal, Furukawa (2021) optimizes the trade-off between efficiency (which decreases when estimates are discarded) and bias (which increases when less precise estimates are included). The cut-off percentage is thus determined endogenously in the model, and in our case it yields an estimate of 0.28 for the mean discount rate.

The third non-linear technique is the selection model developed by Andrews & Kasy (2019). The selection model assumes that the probability of publication changes abruptly after reaching pre-defined thresholds for the t-statistic (in our case: 0, 1.65, 1.96, 2.33). The technique then computes how much estimates from each bracket are over- or under-represented in the literature, and reweights them accordingly. The selection model gives us an estimate of 0.25 for the mean discount rate. Finally, the fourth non-linear specification we employ is the Endogenous Kink technique introduced recently by Bom & Rachinger (2019). The logic of the estimator is similar to both the linear funnel asymmetry test and the stem-based technique by Furukawa (2021): it also assumes that highly precise estimates are unbiased, but fits the publication bias function using two linear segments. The first segment is horizontal (no bias, therefore no relation between estimates and standard errors for the most precise estimates) and the second segment has a slope equal to the correlation between estimates and standard error for less precise estimates. Bom & Rachinger (2019) show how the "kink" (that is, the point where both segments join) can be identified. The technique yields an estimate of 0.15 for the mean discount rate.

In sum, Table 3.2 gives us significant estimates for publication bias (Panel A) and estimates of the corrected mean discount rate in the range 0.15–0.33 (Panel B). We prefer to focus on the most conservative estimate from Panel B, 0.33. These results indicate that publication bias exaggerates the mean reported discount rate more than twofold, from 0.33 to 0.80 (the simple uncorrected mean). But again we have to note that our results hinge on the assumption that in the absence of publication bias there is no correlation between estimates and standard errors; even the selection model by Andrews & Kasy (2019) uses this assumption for identification. There are two reasons why the assumption might not hold in the case of the discounting literature, and we thank two anonymous referees of this Journal for articulating the reasons. First, researchers are likely to design the experiment in a way that is tuned to detect discount rates near zero and does not uniformly cover the entire interval of possible rates. Consequently, smaller discount rates are likely to be measured with greater precision, and thus the correlation between estimates and standard errors can arise even in the absence of publication bias. Second, negative estimates of the discount rate can be missing from the literature simply

Table 3.3: Caliper tests for different ranges of discount rate estimates

Caliper test for $\delta \in \langle -0.5, 0.5 \rangle$	OLS	Precision
Standard error	0.0919^{**}	0.473^{**}
(*publication bias*)	(0.0367)	(0.190)
Constant	0.214^{***}	0.184^{***}
	(0.0139)	(0.0188)
Observations	538	538

Caliper test for $\delta \in \langle -1, 1 \rangle$	OLS	Precision
Standard error	0.205^{***}	0.949^{**}
(*publication bias*)	(0.0398)	(0.409)
Constant	0.325^{***}	0.232^{***}
	(0.0444)	(0.0313)
Observations	717	717

Caliper test for $\delta \in \langle 0.25, 0.75 \rangle$	OLS	Precision
Standard error	0.0835^{**}	0.536^{*}
(*publication bias*)	(0.0395)	(0.288)
Constant	0.429^{***}	0.371^{***}
	(0.0351)	(0.0428)
Observations	313	313

Caliper test for $\delta \in \langle 0.5, 1.5 \rangle$	OLS	Precision
Standard error	0.125^{***}	0.199^{**}
(*publication bias*)	(0.0126)	(0.0786)
Constant	0.801^{***}	0.764^{***}
	(0.0295)	(0.0341)
Observations	244	244

Notes: The table reports the results of regression $\delta_{ij} = \delta_1 + \gamma_1 \cdot SE(\delta_{ij}) + u_{ij}$, where δ_{ij} denotes the i-th annualized discount rate estimated in the j-th study, and $SE(\delta_{ij})$ denotes its standard error. The regressions only include estimates within the bounds indicated by the caliper. The table shows estimation by OLS and precision weighting. Standard errors, clustered at the study level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

because elicitation techniques used by the researchers do not allow for negative values: for instance, if experimental subjects are always offered a larger sum of money in the future compared with the immediate option.[7]

While we see no bulletproof way how to measure the quantitative importance of these two caveats for our results, a useful exercise is to conduct a

[7] If the correlation between estimates and standard errors is driven by this second caveat, certainly it cannot be called publication bias. (The issue is also discussed by Nobel *et al.* 2020, p. 11.) But it can still represent another type of research bias that should be corrected in meta-analysis: suppose an extreme case in which the true discount rate is negative. If an experiment does not consider the possibility of negative discounting, it will inevitably produce estimates biased upwards. A similar bias will arise on average in a more plausible setting in which the true discount rate is positive but small, because most elicitation designs will allow large positive outliers, but not negative ones.

caliper test inspired by Gerber & Malhotra (2008) and Brodeur *et al.* (2020b). Caliper tests are typically employed to identify a systematic break related to publication bias at a particular psychologically important threshold (such as 0 for the point estimate or 1.96 for the t-statistic). For example, Brodeur *et al.* (2020b) show how, for many quasi-experimental techniques commonly used in economics, estimates that are just significant at the 5% level (that is, have t-statistics slightly larger than 1.96) are more likely to get published than estimates that are just insignificant. The essence of the caliper test is thus to compare the number of estimates just below and just above a particular threshold: given a sufficiently narrow caliper, there should be no difference. In this paper we use a different tactic and employ calipers of varying width to constrain our baseline linear regression (of estimates on their standard errors) in an attempt to address the important caveats mentioned earlier.

We use two groups of calipers. First, we focus on small estimates, both positive and negative. If the correlation between estimates and standard errors persists when large positive outliers are excluded, the finding of publication bias is probably not fully driven by the authors designing experiments in a way that is tuned to detect discount rates near zero. Second, we focus on positive estimates approximately around the mean and median of the reported discount rates. If the correlation between estimates and standard errors persists when only safely positive estimates are considered, the finding of publication bias is not fully driven by the impossibility of negative discount rates in many experimental designs. The results of caliper tests of funnel asymmetry are shown in Table 3.3. Note that here we cannot interpret the means corrected for publication bias (the constant in the regression), because the calipers are arbitrary slices of the data. We can interpret the slopes in this regression, and they all suggest a positive correlation between estimates and standard errors. It is important to point out, however, that we still have to assume that the standard error is exogenous within individual calipers. If there is a mechanical relationship between the estimates and standard errors within calipers in the absence of publication bias, caliper tests fail to address the two caveats.

Another way to approach this problem is to use techniques that do not need the assumption of zero correlation between estimates and standard errors in the absence of publication bias—or, in the case of one technique, at least not between studies. Table 3.4 shows the corresponding results. In the first column we apply p-uniform*, a brand new technique to test publication bias and estimate the corrected mean. The technique was developed by van Aert

& van Assen (2021) for psychology, but it can be applied to an experimental economics setting as well. (In fact, it is probably better suited to experimental economics than the traditional publication bias tests that are designed to aggregate regressions because in experimental research the exogeneity assumption for the standard error is unlikely to hold.) At the basis of p-uniform* lies the statistical principle that p-values should be uniformly distributed *at the mean underlying effect size*: i.e., when testing the hypothesis that the estimated coefficient equals the underlying value of the effect (not necessarily zero). The reported t-statistics and p-values, of course, in almost all cases correspond to tests that relate the estimated coefficient to zero. It follows that if the reported p-values are uniformly distributed, the literature is consistent with a zero underlying effect. The idea of p-uniform* is to find a coefficient at which the distribution of p-values is uniform; this is done by recomputing p-values for various potential values of the underlying effect and then comparing the resulting distribution to the uniform one. Similarly the technique's test for publication bias evaluates whether p-values are uniformly distributed at the simple mean reported in the literature. Technical details and more discussions are available in van Aert & van Assen (2021). The results in Table 3.4 show evidence of publication bias significant at the 1% level. The mean corrected discount rate is small (0.18) but imprecisely estimated.

Table 3.4: Relaxing the exogeneity assumption

	p-uniform*	Instrument	Fixed effects
Publication bias	YES***	0.316*	0.875***
	(*0.007*)	(0.183)	(0.0154)
Effect beyond bias	0.176	0.633***	0.341***
	(*0.663*)	(0.158)	(0.00806)
Observations	927	927	927

Notes: In the first column the table reports the results of the p-uniform* test for publication bias developed by van Aert & van Assen (2021); p-values are reported in parentheses. For the remaining two specifications, which show regressions along the lines of the first panel of Table 3.2, standard errors are reported in parentheses and are clustered at the study level. The second column reports an instrumental variable specification (where the instrument for the standard error is the inverse of the square root of the number of observations in a study), and the third column reports a study-level fixed effects specification. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In the second column of Table 3.4 we use the inverse of the square root of the number of observations as an instrument for the standard error following Stanley (2005), Havranek (2015), and Astakhov *et al.* (2019): some method choices in the primary studies can influence both the discount rate and the standard error, which would make our OLS results spurious. (There can also exist a more direct mechanical relationship between estimates and standard errors, as we discussed in the context of the caliper test.) The number of observations is a natural instrument, because it correlates with the standard error by definition. Nevertheless, while not the product of the estimation technique (in contrast to the standard error), in the studies estimating the discount rate the number of observations can be still correlated with the choice of the technique. Therefore the instrumental variable technique cannot be expected to fully address the exogeneity problem. The results in Table 3.4 indicate publication bias significant at the 10% level and an underlying mean discount rate of 0.63. Finally, in the last column of the table we explore whether publication bias appears within studies. This specification still needs the exogeneity condition to hold within individual studies, but relaxes it between studies as the latter source of variation in discount rate estimates is not used. Once again we obtain evidence of publication bias, now significant at the 1% level, and underlying mean effect smaller than the uncorrected simple mean (0.34 vs. 0.8). Overall we prefer this fixed effects estimation because it is simple, elegant, and its results are consistent with the most conservative non-linear technique presented earlier.

The Appendix harbors four sets of further robustness checks. First, in Table B.1.1 we cluster standard errors at the level of authors instead of studies. Several researchers have co-authored many of the studies in our dataset, and consequently the results of these studies do not have to be independent of each other. We have identified 31 clusters for which no co-authors overlap. The results are almost identical to the baseline case, with the exception of the IV specification, in which we lose statistical significance. Second, in Table B.1.3 we exclude estimates for which the discounting model is not explicitly specified. Once again the results are similar, but we obtain smaller estimates of the mean discount rate corrected for publication bias.

Third, in Table B.1.4 we run funnel asymmetry tests with the discount rate in the absolute value. Aside from the standard error, on the right-hand side we include the interaction of the standard error and a dummy variable that equals one for negative values. In consequence, this specification reveals different mechanisms of selective reporting for positive and negative estimates. For positive estimates, our findings are consistent with publication probability increasing with an increasing t-statistic. For negative estimates, our findings are consistent with the opposite: insignificant negative estimates tend to be easier to publish, probably because they are more feasible. Fourth, in Table B.1.2 we investigate how publication bias differs between medians and means of individual-specific discounting. To this end, we include an interaction of the standard error and a dummy variable that equals one for median estimates. Medians comprise 15% of the data set, and the results of the table show mixed findings. According to most techniques, there is little difference in the extent of publication bias between means and medians. Our preferred fixed effects specification, however, indicates that median estimates are substantially less biased than mean estimates.

In sum, this section has shown that, similarly to the rest of the empirical research in economics, the experimental literature estimating discount rates is affected by publication selection bias. The finding holds when we relax the classical meta-analysis assumption that estimates and standard errors are independent in the absence of publication bias (the assumption is unlikely to hold in the experimental literature) and apply a battery of recently developed techniques. We find that the mean reported discount rate (0.80) is exaggerated, and our median estimate suggests that the underlying mean discount rate corrected for publication bias is around 0.33. But of course discount rates vary across individuals and experimental context, an issue to which we turn next.

3.5 Heterogeneity

The substantial differences in the estimates of the discount rate reported in the experimental literature have already been stressed by several previous studies (Frederick *et al.* 2002; Percoco & Nijkamp 2009; Andersen *et al.* 2014; Cheung 2016). As Frederick *et al.* (2002, p. 352) puts it: *"While the discounted utility model assumes that people are characterized by a single discount rate, this literature reveals spectacular variation across (and even within) studies."* Figure 3.2 shows strong differences in the results at the study level. In this section we try to explain the differences by regressing the estimated discount rates on their standard errors together with 21 additional explanatory variables that reflect observable variation in the context in which researchers obtain the estimates. We start from the linear model of publication bias, which is the reason why we retain the standard error variable in the regression. Therefore the second goal of this section is to find out whether our previous findings concerning publication bias prove robust to controlling for heterogeneity.

The first option for estimating such an extended model is simply running a regression with all the collected variables. The problem is that not all the variables are equally important; some are probably redundant, and including all variables would substantially diminish the precision of our point estimates for the effects of the important variables. However, we do not know ex ante which variables are redundant. A common approach would be to eliminate potential redundant variables in a step-wise fashion (sequential t-tests); but in doing so, we can never be sure that we have arrived at the best underlying model. Furthermore, the theory can help us stress some particular variables, but we still do not want to completely ignore the remaining ones. In other words, we face extensive model uncertainty, which is a typical feature of meta-regression analysis. The formal response to model uncertainty in the Bayesian setting is Bayesian model averaging (Raftery *et al.* 1997), our first method of choice.

Bayesian model averaging (BMA) tackles the problem of uncertainty by estimating models with all possible combinations of explanatory variables in the dataset[8] and constructing a weighted average over the estimated coefficients across all these models. The weights used for averaging stem from posterior model probabilities derived from Bayes' theorem and are analogous to information criteria in frequentist econometrics. Posterior model probabilities (PMPs) measure how well the particular model fits the data, conditional on model size. BMA produces posterior inclusion probability (PIP) for each variable, which is the sum of the posterior model probabilities for the models in which the variable is included. Recent applications of Bayesian model averaging in meta-analysis include, for example, Irsova & Havranek (2013); Babecky & Havranek (2014); Havranek & Irsova (2017); Cazachevici *et al.* (2020); Zigraiova *et al.* (2021). More details on BMA, including a formal derivation, can be found in Raftery *et al.* (1997) or Eicher *et al.* (2011).

The application of BMA, however, is not straightforward since estimating the millions of possible model combinations is infeasible. A solution is to approximate the whole model space by applying the Markov chain Monte Carlo algorithm that walks only through the models with high posterior model probabilities (Madigan *et al.* 1995). For approximation we use the BMS package for R developed by Zeugner & Feldkircher (2015). Bayesian model averaging is sensitive to the estimation framework, particularly to the use of *priors* representing the researcher's prior beliefs on the probability of each model (the model prior: how much confidence we place in the prior that, for example, all

[8] If the matrix of explanatory variables X contains K potential variables, this means estimating 2^K variable combinations, i.e., 2^K models. This estimation results in $2^{22} = 4,194,304$ models in our case.

models have the same probability) and regression coefficients (Zellner's g-prior: how much confidence we place in the prior that, for example, all regression coefficients are zero). In the baseline specification we follow the two priors suggested by Eicher *et al.* (2011). First, the unit information prior (UIP) for Zellner's g-prior, which assigns the prior that coefficients are zero the same weight as one observation of data. Second, the uniform model prior, which gives each model the same prior probability, irrespective of the number of variables included in the model. Such intuitive priors are agnostic in the sense that they are easily overridden by data, and Eicher *et al.* (2011) show that they yield good predictive performance.

On top of the uniform model prior we use the dilution prior suggested by George (2010). In this prior the relative weight of each model is further multiplied by the determinant of the correlation matrix of the variables included in the model. The dilution prior is designed to address collinearity: models with high collinearity will have small determinants of the correlation matrix, and therefore little weight in our implementation of BMA.[9]

3.5.1 Variables

The explanatory variables we have collected are listed in Table 3.5; we include the description of each variable, its mean, standard deviation, and the mean weighted by the inverse of the number of estimates reported per study, which effectively equalizes the impact each study has on the statistics. For ease of exposition, we divide the explanatory variables into 4 categories: estimation characteristics, experimental characteristics, subject pool characteristics, and publication characteristics.

Estimation Characteristics

The variation among the reported discount rate estimates can stem from the theoretical assumptions of the intertemporal choice model used in the experi-

[9] A robustness check using the BRIC g-prior suggested by Fernandez *et al.* (2001) and the beta-binomial model prior according to Ley & Steel (2009) can be found in Section B.2; our main results would not change if we opted for this alternative set of priors. A detailed discussion of the priors used in the robustness checks is beyond the scope of the paper; for more details, see Zeugner & Feldkircher (2015). For example, the beta-binomial model prior gives the same weight to each model size (a certain number of variables included in the model), not the same weight to each model. The reason is that moderate model sizes are over-represented: there are many models that have 2^{10} variables, but only one model that has 2^{22} variables.

Table 3.5: Description and summary statistics of regression variables

Variable	Description	Mean	SD	WM
Discount rate	The reported estimate of the discount rate.	0.798	0.973	0.710
Standard error	The standard error of the discount rate estimate.	0.522	1.149	0.214
Estimation characteristics				
Hyperbolic discounting	= 1 if the discounting type is hyperbolic.	0.402	0.491	0.368
Exponential discounting	= 1 if the discounting type is exponential.	0.143	0.351	0.199
Delay	The logarithm of the time horizon of the task.	-0.255	2.222	-0.782
Front-end delay	= 1 if the immediate option is shifted to the future, thereby imposing transaction costs on the instant payoff.	0.338	0.473	0.364
Lab experiment	= 1 if a controlled laboratory experiment is used instead of a field experiment.	0.650	0.477	0.549
Experimental characteristics				
Real reward	= 1 if the reward subjects received is real instead of hypothetical.	0.629	0.483	0.754
Matching task	= 1 if matching is used for elicitation.	0.243	0.429	0.149
Health domain	= 1 if the experiment concerns health questions.	0.055	0.228	0.055
Other domain	= 1 if the experiment concerns questions other than health or money (such as vacation or a kiss from a movie star).	0.082	0.274	0.100
Negative framing	= 1 if the framing of the experimental task is presented as negative, i.e., “losing.”	0.086	0.281	0.072
Neutral framing	= 1 if the framing of the experimental task is presented as neutral.	0.023	0.149	0.031
Stakes	The ratio of the logarithm of the highest payoff possible in the experiment to the logarithm of the median monthly expenditure in the country where the experiment was conducted.	0.817	0.373	0.753
Subject pool characteristics				
Sample size	The logarithm of the sample size used for the experiment.	4.889	1.617	5.035
Students	= 1 if the subject pool consists of students only.	0.528	0.500	0.445
Males only	= 1 if the subject pool contains males only.	0.029	0.168	0.027
Females only	= 1 if the subject pool contains females only.	0.030	0.171	0.054
North America	= 1 if the experiment is conducted in North America.	0.588	0.492	0.589
Asia	= 1 if the experiment is conducted in Asia.	0.058	0.234	0.107
Africa	= 1 if the experiment is conducted in Africa.	0.030	0.171	0.036
Publication characteristics				
Citations	The logarithm of the number of citations the study received in Google Scholar normalized by the number of years since the first draft of the study appeared in Google Scholar.	2.691	1.278	2.776
Publication year	The standardized publication year of the study.	0.000	1.001	0.283

Notes: SD = standard deviation, WM = mean weighted by the inverse of the number of estimates reported per study. The variable *Stakes* is only available for 777 observations; statistics for all other variables are calculated using the full sample of 927 observations. Data on median expenditure are obtained from World Bank (2020).

mental task presented to subjects, that is, mainly from the type of the discounting model and the time horizon that subjects face in their decision. The studies included in our dataset use the hyperbolic discounting model most frequently (373 observations; 40% of the data), followed by the exponential discounting model (133; 14%). Special cases of discounting models such as exponential mixture share, quasi-hyperbolic discounting, or mixed general model occur rarely in our dataset. Due to a lack of information reported in primary studies, we cannot identify the precise type of the discounting model in some of the cases and use this "unidentified" group as a reference category. The time horizon of the decisions presented to the subjects spans from one week to 50 years, while the mean value is 4.07 years. We also take into account whether the study uses front-end delay. With front-end delay the immediate option is shifted to the future, thereby imposing transaction costs on the instant payoff. Last but not least, we control for the general estimation setup—that is, whether the study employs a controlled laboratory experiment or a field experiment.

Experimental Characteristics

The results of any experiment can be affected by procedural subtleties. The second set of explanatory variables therefore comprises experimental and behavioral characteristics of the task presented to the subject pool. Psychological research suggests that there should be no systematic difference observed between real and hypothetical payoffs in discounting experiments (Johnson & Bickel 2002; Kuhnberger *et al.* 2002; Locey *et al.* 2011). The recent literature, however, provides more ambivalent results stating that hypothetical conditions yield patterns of discounting that mirror those for real effort tasks, but these may change with repeated exposure to the decisions. The nature of the payoffs provided with the repetition of those tasks therefore needs to be taken into account when designing discounting studies (Malesza 2019). We therefore control for this payoff effect by extracting the information on the nature of the reward from primary studies; 53% of the discount rates are computed for hypothetical payoffs. For a subsample of estimates, we are able to collect data on the size of the maximum payoff available in the experiment. We relate the maximum payoff size to World Bank data on household median monthly expenditure in the country, and the resulting variable is labeled "Stakes." Note that this variable is not included in the baseline model, because doing so would imply disregarding all the observations for which the variable is not available.

Following the reasoning of Frederick *et al.* (2002) and others, we control for the variation in the estimates caused by the elicitation method used in the experiment. We include a dummy variable for matching tasks, taking choice tasks as the reference category present in 66% of cases. An important behavioral aspect of the corresponding task is represented by the domain over which the intertemporal decision is made. The majority of observations utilize monetary payoffs (87%); we therefore use them as the natural reference category in this regard. We codify the remaining domains by using dummy variables, distinguishing between the health domain and other domains—typically, more exotic ones (e.g. vacation, certificate, or a kiss from a movie star).

The design of any experiment is seldom immune to the issues of framing effects that refer to the finding that subjects often respond differently to different descriptions of the same problem (Tversky & Kahneman 1981). The majority of discounting tasks are presented (framed) as positive decisions, e.g., choices between an amount of money today and a greater amount tomorrow (89.1%). There are, however, also negative framings of the tasks present in our dataset (8.6%). For example, Chapman & Winquist (1998) and Hardisty *et al.* (2013) use monetary losses in their experiments. Other studies with negative framing operate with the health domain (Dolan & Gudex 1995; Read & Read 2004). Neutral framing applies for only 2.3% of the observations.

Subject Pool Characteristics

We describe the subject pool characteristics of an individual study by several variables. First, we control for the size of the subject pool by coding the number of subjects used for deriving the estimate; the mean is 271. Second, we control for the composition of the subject pool by incorporating dummy variables reflecting whether the pool consists exclusively of male or female subjects. The majority of studies, however, use non-exclusive subject pools consisting of both males and females (94.1%).

A general concern of any experimental study is its external validity, i.e., the extent to which its results can be generalized to other situations. Economic experiments are often criticized for using university students (typically economics majors) as experimental subjects—a pool of people with specific characteristics not always generalizable to the whole population (Marwell & Ames 1981; Carter & Irons 1991; Frank *et al.* 1993). The behavior of decision makers recruited from natural markets has been examined in a variety of contexts, and it has typically not differed from that exhibited by more standard (and far less costly) student subject pools (Davis & Holt 1993, p. 17).[10] We control for the potential effect of a subject pool composed exclusively of student subjects. In addition, as recommended by an anonymous referee, we include an interaction of the student and lab experiment dummy variables. These two variables are correlated, because lab experiments often rely on students, and students, who are commonly familiar with lab experiments, may potentially behave differently in lab and field settings. Finally, the heterogeneity in the reported discount rates may stem from different cultural characteristics of populations. The primary studies do not give us much information to build on systematically, but at least we can control for continents out of which the subject pool was recruited. The majority of studies recruit subjects from European countries (32.4% obs.) and North America (58.8%). We also experimented with including dummy variables for each individual region, but doing so creates collinearity problems.

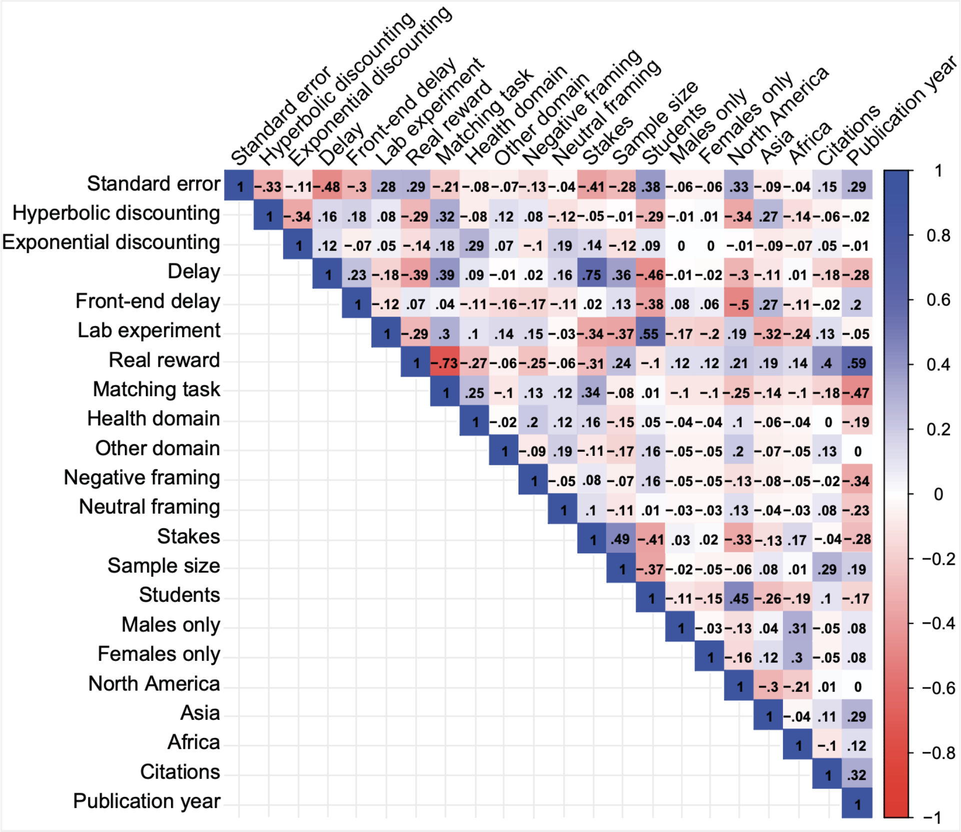
Publication Characteristics

We do not exclude any journal articles based on their supposedly poor quality, but we try to control for it—even poor-quality studies can bring useful information, especially if their results differ from those of high-quality studies. Some of the aspects related to quality are captured by the data and method characteristics described above. However, other quality aspects are surely more difficult to observe. Therefore we use two rough proxies: the age of the study and the number of citations. These are no perfect controls for quality, but other things being equal, newer and highly cited studies tend to be more reliable. For computing the age of the study we do not use the year of journal publication; due to different publication lags in different economics and psychology journals, such a measure would be of limited use. Therefore, we use the date of the first appearance of a draft of the paper in Google Scholar. For citations, we also rely on Google Scholar and compute the number of per-year citations that the primary study has obtained since the first draft appeared.

Figure 3.4 shows the correlations between the variables we consider. Several patterns emerge that are informative for understanding the types of experiments observed in the data. For example, lab experiments tend to use matching tasks with hypothetical rewards and rely on students. Recent and highly cited studies typically employ real rewards. Recent studies are also less

[10] Recent evidence on this problem is provided by Depositario *et al.* (2009).

Figure 3.4: Correlation matrix



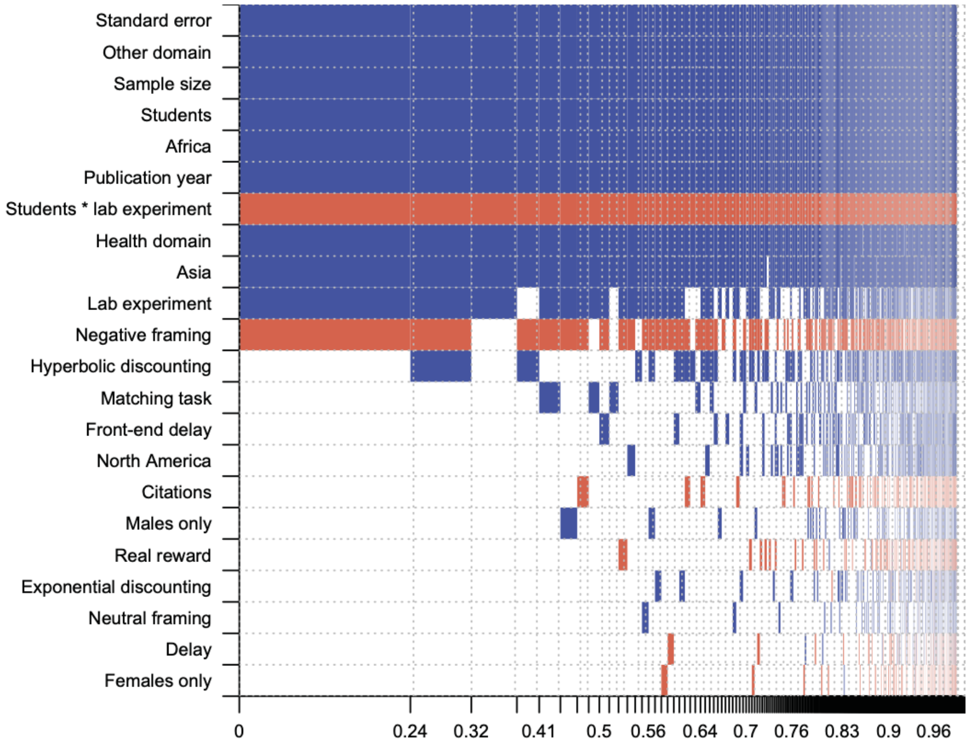
Notes: The figure presents Pearson correlation coefficients for the variables reported in Table 3.5. Correlations for *Stakes* are computed using the 777 observations for which the variable is available. For all the other variables the figure shows correlations calculated at the full sample of 927 observations.

likely to use negative and neutral framing compared to older studies. Payoffs in experiments tend to be smaller when students are used.

3.5.2 Results

The results of the BMA estimation are visualized in Figure 3.5. The variables are displayed on the vertical axis and sorted by posterior inclusion probability. PIP can be thought of as a Bayesian analogy of statistical significance—we therefore see the most "significant" variables at the top of the figure. The horizontal axis denotes individual regression models sorted according to the posterior model probability, from left to right. The PMP represents how well the model fits the data relative to its size; the width of the columns is proportional to the PMP. The colors of individual cells denote the sign of the corresponding regression coefficients. Blue (darker in grayscale) depicts a positive sign, while red (lighter in grayscale) depicts a negative sign. Blank cells denote the exclusion of the variable from the given model.

Figure 3.5: Model inclusion in Bayesian model averaging



Notes: The response variable is the estimate of the discount rate reported in a primary study. The columns denote individual models; variables are sorted by posterior inclusion probability in descending order. The horizontal axis denotes cumulative posterior model probabilities. The estimation is based on the unit information prior recommended by (Eicher *et al.* 2011) and the dilution prior suggested by George (2010), which takes into account collinearity. Blue color (darker in grayscale) depicts variables with a positive estimated sign. Red color (lighter in grayscale) depicts variables with a negative estimated sign. Variables with no color are not included in the given model. The numerical results of the BMA exercise are reported in Table 3.6.

The numerical results of BMA are reported in the left-hand panel of Table 3.6, which shows the posterior mean and standard deviation for each variable together with the posterior inclusion probability. Not counting the intercept, which is included by default in all models, eleven variables have PIPs above 50%: the standard error, the dummy for lab experiments, the dummy for health domain, the dummy for other (exotic) domains, the dummy for negative framing, sample size, the dummy for students in the subject pool, the interaction between student and lab experiment dummies, the dummy for subjects drawn from Asia, the dummy for Africa, and publication year. In the remainder of this subsection we will go through these results in more detail.

Table 3.6: Explaining the heterogeneity in discount rate estimates

	Bayesian model averaging			Frequentist check (OLS)		
Variable:	Post. mean	Post. SD	PIP	Mean	SE	p-value
Constant	-0.244	NA	1.000	-0.253	0.163	0.126
Standard error	0.549	0.021	1.000	0.542	0.035	0.000
Estimation characteristics						
Hyperbolic discounting	0.039	0.062	0.352			
Exponential discounting	0.006	0.030	0.076			
Delay	0.000	0.002	0.041			
Front-end delay	0.014	0.041	0.143			
Lab experiment	0.155	0.101	0.776	0.222	0.091	0.018
Experimental characteristics						
Real reward	-0.005	0.027	0.077			
Matching task	0.017	0.046	0.161			
Health domain	0.345	0.088	0.993	0.356	0.076	0.000
Other domain	0.441	0.070	1.000	0.442	0.153	0.006
Negative framing	-0.148	0.106	0.734	-0.205	0.102	0.049
Neutral framing	0.003	0.031	0.046			
Subject pool characteristics						
Sample size	0.075	0.014	1.000	0.076	0.029	0.012
Students	0.877	0.111	1.000	0.901	0.223	0.000
Students * Lab experiment	-0.753	0.144	1.000	-0.813	0.239	0.001
Males only	0.013	0.052	0.090			
Females only	-0.001	0.023	0.041			
North America	0.012	0.041	0.127			
Asia	0.385	0.103	0.990	0.428	0.117	0.001
Africa	3.170	0.118	1.000	3.174	0.066	0.000
Publication characteristics						
Citations	-0.003	0.011	0.095			
Publication year	0.121	0.026	1.000	0.114	0.051	0.030
Observations	927			927		
Studies	56			56		

Notes: Response variable = annualized estimates of the discount rate. In the first specification from the left we employ Bayesian model averaging (BMA) using the unit information prior recommended by (Eicher *et al.* 2011) and the dilution prior suggested by George (2010), which takes into account collinearity. The second specification, frequentist check (OLS), includes variables recognized by the BMA as having a posterior inclusion probability above 50%. Standard errors in the frequentist check are clustered at the study level. SD = standard deviation, PIP = Posterior inclusion probability, SE = standard error. All variables are described in Table 3.5.

The first important result of the BMA analysis concerns publication bias. Standard errors are robustly correlated with the point estimates of the discount rate even when we control for 21 additional aspects of studies and estimates. The result corroborates our previous findings that the correlation is not spurious and does not result from an omission of factors that influence both the standard error and the point estimate. Moreover, both the posterior mean in BMA and the point estimate in the frequentist check suggest that the correlation is strong.

Results for Estimation Characteristics

An often-discussed factor potentially affecting the heterogeneity in discount rate estimates is the length of the delay over which the decision is made. This factor is inherently embedded as the parameter k in the discounted utility model presented in Equation 3.1. According to the exponentially discounted utility theory, the values of all future outcomes should be discounted at a constant rate (Frederick *et al.* 2002). Our results do not disagree: we find little systematic relationship between reported estimates of the discount rate and the length of the delay. This finding contrasts the results of, among others, Mazur (1984), who presents evidence for hyperbolic discounting, or, more recently Tsukayama & Duckworth (2010), who find that subjects discount rewards more steeply when they find the discounting domain particularly tempting. On the other hand, our results are in line with Andersen *et al.* (2014). A related effect is the importance of the dummy for exponential discounting, of which the constant discount rate is a key property. Our analysis suggests that tasks with exponential setups, i.e., with a constant discount rate between decisions with different delays, do not systematically differ from other studies in terms of the reported discount rates. Moreover, the estimates in our sample do not seem to be significantly different when hyperbolic discounting is applied. We note, however, that our reference category comprises estimates for which the discounting model is not explicitly identified in the primary studies. But even if the reference category includes some instances of exponential and hyperbolic discounting, our results are consistent with very little difference in the reported discount rates between studies specifying the exponential form and those specifying the hyperbolic form.

Two additional results related to estimation characteristics are important. The first result is the low posterior inclusion probability and therefore the absence of the variable *Front-end delay* in most BMA models, which again contrasts many previous findings in the literature that front-end delay tends to decrease estimated discount rates (for example, Coller & Williams 1999), but is consistent with the results of Andersen *et al.* (2014). A second important result is the difference between field and laboratory experiments. This finding suggests that a controlled laboratory environment produces more evidence for impatience than a field study environment.

Results for Experimental Characteristics

Several studies find that individual discount rates are not very correlated across different domains such as money and health—this diversity is called domain independence. Cairns (1992), for example, estimates of discount rates that are different for future health as compared to future wealth states; Chapman & Elstein (1995) demonstrate in two experiments that decision makers use different discount rates for health-related decisions and money-related decisions, with less patience for the health domain. See Loewenstein *et al.* (2003) for more examples of domain independence.

Our results suggest that people tend to be more impatient when the experiment concerns health than when it concerns money. It is difficult to transfer health states over time, so questions about health are, to some extent, similar to questions about money when liquidity constraints are binding (see Bleichrodt *et al.* 2016). When liquidity constraints are present and binding, people cannot increase current consumption at the expense of consumption in the future. A high discount rate follows. In addition, we also find that people tend to be more impatient when making their decisions in more exotic domains than money: holiday preferences, gift certificates, kisses from movie stars. Our results thus strongly corroborate domain independence.

Describing the estimation characteristics in Section 3.5, we referred to the literature suggesting there should be no difference whether real or hypothetical payoffs are used in discounting experiments. Our results confirm that it indeed does not matter whether the decision is made with fictive payoffs only. Real rewards do not systematically affect the estimates of the discount rate. Researchers can thus use hypothetical questions that have advantages in the elicitation of time preferences since hypothetical setting allows us to ask questions involving long time horizons and large payoffs (Wang *et al.* 2016).

We find no substantial effect for some other experimental characteristics. Different experimental tasks do not bring substantially different results: matching does not seem to differ significantly from choice tasks, which suggests that the inference of an individual's discount rate by the matching method does not systematically outperform the interval elicitation provided by choice tasks. In contrast, the estimated discount rates are affected by framing, and negative framing is associated with smaller estimates. The result is consistent with Harris (2012) and Hardisty *et al.* (2013), among others, who stress the role of dread in intertemporal choices: it is itself aversive to wait for an aversive outcome, and for many subjects it is preferable to get it over with. Finally, we find that the stakes of the experiment (maximum possible payoff relative to personal expenditure) are associated with smaller reported discount rates. (Note that the BMA specification featuring this variable is included in Table B.2.3 in the Appendix; the variable is not available for all observations, and thus is not included in the baseline BMA estimation.) The result is consistent with a large literature (for example, Thaler 1981; Benzion *et al.* 1989; Warner & Pleeter 2001; Meyer 2015), and a possible explanation is that non-monetary transaction costs of borrowing or saving that increase the discount rate may be relatively larger for smaller payments.

Results for Subject Pool Characteristics

The long-term debate over the external validity of the experiments performed on student samples is reflected in our analysis by the variable *Students*. Our results suggest that students make more impatient choices in discounting tasks than the general population, which is consistent with Harrison *et al.* (2002) and can be explained by the fact that students tend to be more liquidity-constrained. In contrast, the interaction between student and lab experiment dummies shows a negative coefficient: students that participate in laboratory experiments tend to display relatively little impatience. This finding can be caused by several factors, out of which the standard argument would point to the self-selection of students into subject pools in laboratory experiments. The vast majority of lab experiments are conducted with university students majoring in economics, who have been shown, for example, to be more selfish than the general population (Marwell & Ames 1981). Two types of hypotheses explain why this may be the case: 1) the selection hypothesis, according to which individuals concerned with economic incentives opt for economic studies, and 2) the learning hypothesis, which states that individuals studying economics learn behavioral patterns out of the theories and models they pursue (Carter & Irons 1991). It might be true that not only more "selfish" individuals self-select into study fields such as economics but also that more patient students self-select into the roles of experimental subjects.

Our results provide some evidence that discount rates elicited from subject pools in Asia and Africa significantly differ from those obtained in other parts of the world. The Asian and (especially) African population is, according to our analysis, more impatient than the population of other continents. This result is in line with the results of the large cross-country study on time preferences by Wang *et al.* (2016, p. 17), who observe that *"Africa has the lowest percentage of participants choosing to wait (*33%*)."* The benchmark demographic area—Europe—seems to follow similar patterns of discounting as North America and display lower discount rates. Again, a possible explanation is related to liquidity constraints, which might be larger in Asia and Africa than in the West. Nevertheless, a disclaimer is in order: for Africa we only have two studies in our sample. Next, we also obtain evidence of an impact of the sample size on the discount rate estimates: large experiments seem to produce larger discount rates, though the effect is economically weak. Finally, neither exclusively male nor female subject pools report significantly different results of discount rates in our sample compared to the baseline (mixed) subject pools.

Results for Publication Characteristics

Out of the publication characteristics that we consider, the number of citations does not matter for the estimated discount rates, while publication year is positively associated with the estimates: other things being equal and on average, newer studies show more evidence for impatience. The age of the study can be considered a rough proxy for (unobserved) quality aspects that are not captured by the variables discussed earlier. There are certainly quality aspects that we do not control for, and an obvious solution is the addition of study-level fixed effects. We opt for the fixed-effects estimator in the previous section that focuses on publication bias, but here, it is not feasible: for many variables in which we are interested the within-study variation is very small.

3.5.3 Robustness Checks

In Section B.2 we perform several different sensitivity checks in order to confirm whether our baseline BMA results presented earlier in this section are robust. First, we combine the reduction in model uncertainty resulting from BMA estimation with traditional frequentist estimation: in other words, we use a Bayesian technique for the selection of variables and a frequentist technique for estimation. The best model identified by the BMA exercise includes eleven explanatory variables (plus the intercept). These variables also have a posterior inclusion probability above 0.5 and therefore should, according to the classification by Kass & Raftery (1995), have a non-negligible impact on our response variable. We re-estimate this best BMA model using the standard

OLS technique, clustering standard errors at the study level. The results of this estimation are provided in the right-hand panel of Table 3.6 and are very similar to the baseline BMA results.

Second, we perform a robustness check using an alternative set of BMA priors, employing the BRIC g-prior suggested by Fernandez *et al.* (2001) together with the beta-binomial model prior, which gives each model size (in contrast to each model) equal prior probability (Ley & Steel 2009). We label this estimation according to the g-prior parameter as "BRIC." The results of this robustness check are reported in Table B.2.2 in the appendix and are again similar to those of the baseline specification. In the right-hand panel of the same table we report the results of a fully frequentist technique, FMA. It employs Mallow's weights, which have been shown by Hansen (2007) to be optimal for frequentist model averaging, and the orthogonalization of model space suggested by Amini & Parmeter (2012). FMA has recently been applied in meta-analysis, for example, by Bajzik *et al.* (2020); Havranek *et al.* (2017; 2018a;b;c). Also this robustness check corroborates the results we have discussed previously.

Third, in Table B.2.3 we present three BMA specifications that use a subset of discount rate estimates, a different set of variables, or both. The first specification from the left excludes the standard error. While the exclusion might introduce an omitted-variable bias (the standard error, our proxy for the extent of publication bias, is a key variable in all our previous models), it reduces the danger of endogenous controls. Of the eleven variables with posterior inclusion probability above 50% in the benchmark model, two (health domain and other domain) slip below the 50% threshold, though in the case of health only slightly (to 44%). Nevertheless, there are 5 new variables that achieve a posterior inclusion probability above 50%, including *Real reward*. Our results thus suggest that if we ignored publication bias in the heterogeneity analysis, we would (erroneously, in our opinion given the remaining evidence) conclude that the use of hypothetical rewards biases the results of experiments. The second specification from the left includes a variable reflecting the size of stakes in the experiment, information that is available only for a subset of the discount rate estimates. The estimated effect of the variable is negative, which is consistent with the magnitude effect (Meyer 2015). The third specification excludes discount rate estimates for which the discounting model is not explicitly specified in the paper. Here we lose high posterior inclusion probability for the variable reflecting student samples, but we note that the variable proves to be important in all other specifications.

Finally, in Table B.2.4 we consider two specifications that feature i) an interaction term between *Money domain* and *Non-linearity correction* and ii) a sub-sample of estimates for which the measurement error in the variable *Delay* is reduced. The interaction term is meant to capture the difference between discount rates estimated with and without correcting for non-linearity in utility functions (non-linearity is discussed in Section 3.2). Nevertheless, the interaction attains a very low posterior inclusion probability. Hence we fail to obtain evidence which would suggest that this variable is important for systematically explaining the heterogeneity in the reported discount rates. Regarding the right-hand part of Table B.2.4, we use a sub-sample of estimates for which delay is precisely defined. For 61% estimates of the discount rate in our sample, the corresponding delay is clearly reported in the papers. The remaining estimates are derived from a series of questions with varying horizons, where for "delay" we use the maximum horizon to which a subject is exposed in a given experimental task. Similarly to the baseline BMA result, we fail to obtain the anticipated significant negative coefficient. The insignificance result would likewise hold if we used the mean or median instead of the maximum to approximate the delay variable for discount rate estimates obtained from questions with varying horizons.

3.6 Concluding Remarks

We provide a quantitative synthesis of the literature that uses experiments to identify individual discount rates. We examine 927 estimates of the discount rate reported in 56 primary studies. By employing meta-regression and other methods, we detect selective reporting against null and negative results. The mean reported discount rate is 0.80. Using conservative techniques, we find that the mean drops to about 0.33 after we correct for publication bias—that is, people are more patient on average than what is indicated by a naive summary of the conclusions of the experiments. This result is in line with Imai *et al.* (2021a), who report evidence of modest selective reporting in the literature estimating the present bias parameter. In contrast, Imai *et al.* (2021b) find little evidence of publication bias in laboratory economics experiments.

The estimates of the discount rate vary a great deal. We explain this heterogeneity by using Bayesian model averaging, a method accounting for model uncertainty inherent in meta-analysis. We corroborate the presence of selective reporting in the literature by showing that the standard error is an important factor in the heterogeneity of discount rate estimates. We corroborate the domain independence hypothesis stressed by the previous literature (Cairns 1992; Chapman & Elstein 1995; Loewenstein *et al.* 2003) since discount rates for different questions (for example, health on one hand and money on the other) differ systematically. Other important results include the systematic difference between lab and field experiments and the importance of framing and the composition of the subject pool.

The results of our study can be used in various settings. The discount rate has implications for decisions regarding savings, education, smoking, exercise, and other contexts of day-to-day behavior (e.g., Chabris *et al.* 2008a; Meier & Sprenger 2010). Accurate measures of discounting parameters can provide helpful guidance in welfare analyses on the potential impacts of policies and provide useful diagnostics for effective policy targeting (Andreoni *et al.* 2015); moreover, they can be applicable to modeling political campaigns, advertisement, and R&D investment (Deck & Jahedi 2015b). Other examples of applications are discussed by Deck & Jahedi (2015a), who examine discounting in strategic settings, such as auctions or experimental contests, in which it is often critical to accurately predict the behavior of counterparts.

Climate change policies, in which the individual pure rate of time preference or the social discount rate is needed to evaluate the long-term effects, can serve as an example of a welfare analysis application of our results. The pure rate of time preference together with the growth rate of per capita consumption and the elasticity of marginal utility of consumption create the basis for the calculation of the Ramsey discount rate consisting of *time* and *growth* discounting elements (Fearnside 2002; Anthoff *et al.* 2009; Foley *et al.* 2013). Our discount rate synthesis together with the results of Havranek *et al.* (2015a), who provide a meta-analysis of the elasticity of marginal utility of consumption, can be employed to calculate the pure rate of time preference from the Ramsey discount rate.

Our results also have broad implications for future experimental research on discounting. The potential for publication bias is correlated with the occurrence of large positive outliers, which means that estimates of the median discount rate are more robust to the bias than estimates of the average discount rate. Indeed, we find some direct evidence in our data set that median estimates may suffer less from publication bias compared to mean estimates. Papers that estimate individual-specific discounting often report median statistics for this reason (see, for example, Kuhn *et al.* 2017). Lab experiments seem to yield, ceteris paribus, larger estimates of the discount rate compared to field experiments. Because both lab and field experiments have their pros and cons (Al-Ubaydli & List 2015), we need more studies along the lines of Andersen *et al.* (2010) that would evaluate the results of both in a comparable environment. We obtain robust evidence that the estimated discount rates are not systematically affected by the fact whether rewards in the experiment are real or hypothetical. In contrast, discount rates vary a lot across domains: subjects display substantially less patience for goods where intertemporal markets are limited compared to money—health, vacations, kisses from movie stars. In conjunction with the finding that discount rates tend to be larger for groups that are likely to be liquidity-constrained (e.g., students), these results suggest that the experimental subjects' decisions are not fully divorced from outside conditions. If this is the case, current experimental measures may not allow us to properly identify preference parameters, though they are useful for understanding the intertemporal behavior of subjects under various external constraints (Dean & Sautmann 2021). The literature thus awaits novel techniques that will ensure narrow bracketing and enable an even cleaner identification of the underlying discount rates.

Chapter 4

Incentives and Motivation: A Meta-Analysis

Abstract Do financial incentives motivate people to work better? A host of research papers in psychology have long tried to answer this question, together with more recent research papers from behavioural economics. We take 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 39 variables to capture the underlying nature of the effect incentives have on motivation and performance. A range of statistical tests suggest 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 effect determinants. Among these, publication bias pushes this effect upward the most, along with laboratory setting and positive framing in the task. School setting, charitable giving, cross-sectional data, self-obtained rewards, quantitative performance, and students subgroup then pull the effect in the opposite direction.

Draft paper was co-authored with Petr Čala.

4.1 Introduction

Today's society stands on the paradigm that we provide our knowledge, time end effort in exchange for a pay. Some provide a work of better, some of less quality. The burning question of any manager or employer is how to make the latter group perform better at their efforts? Financial incentives were long perceived as the key to success. By increasing one's pay, the institutions were inducing higher quality of work from their employees. But is it true that financial incentives motivate people to work better? If one were to ask an economist how to get people to work better, he would probably suggest trying to increase people's pay. After all, it is common knowledge to economists that people respond to incentives (Mankiw 2014). On the flip side, one might receive a different response when posing the same question to a psychologist. A psychologist may object that an emphasis on the reward for performing a task may diminish the enjoyment one associates with that task, possibly resulting in decreased motivation and performance during it (Deci *et al.* 1999). Deci (1971) observed this phenomenon and studied it empirically, following which this phenomenon became commonly known as the "crowding-out intrinsic motivation" theory. Since then, decades of research have further followed in a similar spirit, providing countless experiments that helped put these findings into a quantitative perspective.

However, the most prominent ideas from the rewards-motivation domain do not present identical findings. The variations of results and their corroborations only in specific contexts are substantial. 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 toward 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 that involve judgment and find a positive monetary incentive effect. Bonner *et al.* (2000) 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 performance. Gneezy & Rustichini (2000) argue that one has to be paid enough to show an increase in one's performance. Last but not least, Ariely *et al.* (2009) discover that excessive rewards may have a detrimental effect on performance.

Large heterogeneity among these results suggests that a synthesis of this topic would bring substantial value to the field. And indeed, the synthesis of the literature on the effect of rewards on motivation was done before (Rummel & Feinberg 1988; Cameron & Pierce 1994; Jenkins *et al.* 1998; Deci *et al.* 1999; Cameron 2001; Cerasoli *et al.* 2014; Van Iddekinge *et al.* 2018, among others). None of the studies, however, tries to isolate the outlooks of either economists or psychologists by looking at the available literature from strictly separated perspectives. And yet an economist would have different outlook and, more importantly, the expectations than a psychologist. The former would expect a stronger effect of incentives while the latter would expect intrinsic motivation to have a greater impact on performance. A comparison of these two perspectives would be surely beneficial. Hence, we aim to synthesize the decades of research on this topic in a quantitative meta-analysis from separate perspectives, both economic and psychological. In this study, we focus on the former and look at how the effect behaves strictly across economic literature. We aggregate individual economic studies together being thus able to observe the underlying relationships and causalities of the rewards-motivation effect, along with potential systematic misbehaviour (Hunter *et al.* 1982).

Looking further at the list of synthetic studies that dealt with the topic before, there are only two studies Cameron & Pierce (1994); Cerasoli *et al.* (2014) that examine whether there appears a phenomenon widespread not only in economics but also in other fields in the available literature—a selective reporting (Doucouliagos & Stanley 2013; Ioannidis *et al.* 2017). Researchers and editors tend to publish only statistically significant results, yet, the insignificant and unpublished data contains a lot of valuable information that has the potential to add further value to general debate. We, therefore, analyze reported estimates from available studies and look for this hidden information. Employing the latest methodology in the field, we perform several linear and non-linear methods to uncover potential selective reporting in the economic literature on rewards and motivation.

Last but not least, we trace the heterogeneity in the reported estimates to the design of the experiments while accounting for model uncertainty. We collect a set of 39 explanatory variables focusing on different angles related to effect characteristics, task nature, reward scheme, motivation characteristics, study design, subject pool characteristics, methodology, and publication characteristics. We employ the Bayesian model averaging (Raftery *et al.* 1997, BMA) and frequentist model averaging (Hansen 2007, FMA) to discover which characteristics affect the reported estimates the most.

Together with a noticeable selective reporting, a result demonstrating a publication bias in the literature, we find a diverse overall impact of rewards on motivation. The incentives seem to work more under some conditions while less under others. Our results highlight the importance of both the individual factors driving the experimental methodology and other characteristics. Most importantly for the field of experimental economics, rewards have a larger effect on performance in the laboratory rather than in a field setting. The same appears when the framing of the task is positive—rewards work better than punishments. Several characteristics work in the opposite direction. For example, using students as experimental subjects reports lower performance compared to other samples. Moreover, quantitative performance measures in tasks affect the subject's performance negatively. This can have large implications for various real-life contexts where school or job-related settings are particularly interesting. Qualitative measures of performance could deliver a much better effect on performance. The characteristics affecting performance in our results are used in the experimental economics methodology widely. We, therefore, prompt for great care when designing experiments dealing with performance tasks and motivation since standard experimental setup may provide biased results.

The remainder of the paper is structured as follows. Section 4.2 explores basic concepts and background of the topic, along with already existing most prominent research. Section 4.3 describes methodology and collection of our dataset. Section 4.4 tests for selective reporting and publication bias via various statistical tests and presents their results. Section 4.5 uses model averaging to explain the influence of literature heterogeneity on our findings, presents its results and estimates the best-practice effect in the literature. Section 4.6 concludes the paper.

4.2 Estimating the effect of rewards on motivation

In this section we focus only on a handful of theories most important to our research, briefly describing the basic concepts that are necessary for the understanding of our meta-analysis. While discussing the results of previous studies we summarize to the reader the fundamental ideas and findings from the topic. We also cover existing meta-analyses already elaborated on the topic of rewards and motivation.

The *Crowding-out intrinsic motivation* theory by Deci (1971) forms a cornerstone of psychological research on the topic of people's motivation when presented with incentives. It states that an emphasis on a sole reward in exchange for performing a task may diminish the enjoyment one associates with this task possibly resulting in decreased motivation and performance during it (Deci *et al.* 1999). Deci (1971) highlighted the so-called *undermining effect*: providing incentives to someone for carrying out a task they already enjoy undermines their original reason for doing it. Dozens of research papers have further followed this theory, providing evidence that helped to put these findings into a quantitative perspective. We refer to some of those in the following paragraphs.

Looking at how this motivation changes while varying the type of rewards (e.g., verbal versus tangible), it is possible to refer to the *Cognitive evaluation theory* (Deci & Ryan 1985). This theory explains the effects of external consequences on internal motivation. 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 in 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) expand to this 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.

Further exploring the effect of rewards on motivation we draw our attention to the *General interest theory*, discussed by Eisenberger *et al.* (1999). in contrast to a Cognitive evaluation theory, it takes both the negative and positive effects of the rewards into consideration. Furthermore, the subject's intrinsic motivation should, according to this theory, 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.

The meta-analysis we present in this paper is certainly not the only one carried out on the topic. The substantial differences in the estimates of the effect of rewards on motivation reported in the experimental literature were already stressed by several former meta-analyses before. According to our knowledge, there exist 11 different meta-analyses already published on this topic until this point: 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). Eight of these studies, 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. The other three studies then look at the relation of rewards, motivation, and cognitive ability of performance.

In their methodologies, the former meta-analyses use *Cohen's d* as an indicator for the *undermining effect* most prevalently. Hedges & Olkin (2014) describe the calculation of *Cohen's d* as obtaining the difference between the means of the treatment and control groups and dividing the result by the pooled within-group standard deviations while adjusting for sample size. The resulting estimate's size then indicates either the enhancement effect in the case of a positive sign or the *undermining effect* in the opposite case. Only two of the former 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 build on Hunter & Schmidt (2004) and extend further by computing the *corrected population correlation*, which involves assuming population-level estimates of the effect. The second study that employs a methodology different from Cohen's d is Van Iddekinge *et al.* (2018) who choose to follow Hunter & Schmidt (2004) as well, computing the corrected population correlation 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.

Former meta-analyses suggest very ambiguous results of the *undermining effect* of rewards on motivation. 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 that 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 where they state that 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) further strengthen this claim by providing a result of d = -0.39, again with the task-contingent rewards. The last of the former studies focusing on the *undermining effect*, namely Eisenberger *et al.* (1999); Cameron (2001); Deci *et al.* (2001), more or less present results that are in line with the six above discussed analyses. Their claims differ mainly in the technical details of the topic. We refer the reader to the original works for further detail. Nevertheless, we can say that a possible *undermining effect* might appear in our work due to the reward scheme we choose to employ, which we further discuss in Section 4.5.

Briefly summarizing the results of the rest of the former meta-analyses, Jenkins *et al.* (1998) report an effect of financial incentives on performance, which is not statistically different from zero. Cerasoli *et al.* (2014) 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 ($\rho = 0.21 - 0.45$). Their results suggest that rewards and intrinsic motivation do not have to work in the opposite direction, which contrasts with one part of the *undermining effect* theory. Van Iddekinge *et al.* (2018) claim that the effects of ability and motivation on performance are additive rather than multiplicative and that the ability-motivation interaction may not fully explain or predict performance.

Even though we have gone only to the shores of reward and motivation theory, we now redirect our focus to this meta-analysis. 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: how does the effect of rewards-motivation behave when we look at it from the perspective of the economic literature and what could be driving it. Thus, our approach will differ significantly from the former meta-analyses in the sample of studies on which we collect our data. We filter out only studies published in economic journals, which should allow us to observe the behaviour of the effect from an economic point of view. Furthermore, we may compare this behaviour to the existing results of the primary meta-analyses. Last but not least, we follow and build on the example of Cameron & Pierce (1994) by searching for selective reporting in the literature.

4.3 The dataset

A crucial basis for a meta-analysis is a well-targeted and sufficiently wide list of studies. We search the literature on financial incentives and motivation via Google Scholar's full-text search engine concentrating on studies exclusively from the field of economics. We filter the search for the top 30 economic journals according to the IDEAS/RePEc aggregate rankings to set the bar for quality while aiming at the same time to reduce the number of studies for feasibility. The final query that produces the most relevant hits to our topic ended up in 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`. We complete the search, along with the final list of journals to which the query was applied, during July 2020, taking into account the volatile nature of the IDEAS/RePEc ranking. The Google Scholar search yields a total of 202 studies, which we categorize and save.

Next, we apply two additional inclusion criteria: First, each study must capture an experiment that observes the relationship between an incentive and its effect on subjects' measurable performance. Second, given the nature of the methods later used in this research, each study must report an effect together with its standard errors. The result is a total of 44 relevant studies, the estimates of which we subsequently code. The overview of studies we include in the meta-analysis is presented in Table 4.1.

Table 4.1: 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 Tao *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)

To accomplish a rigorous approach to data collection we thoroughly read each study and establish a stable coding scheme. The original effect of each study, collected along with its standard error, captures the relationship between incentives and a measurable kind of output (e.g., a change in physical/mental performance, students' Grade Point Average (GPA), among others). Based on these output types we categorize the effect into the *effect variable* that allows us to focus on different kinds of performance while simultaneously observing the effect as a whole.

The original effect in some studies captures a positive influence on performance, while in others it captures a negative one. In the first case, the higher the effect, 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 effect indicates worse performance, such as when measuring the time taken to finish a task. We remedy this problem by using a dummy variable equal to one if this relationship is positive and transform the PCC of estimates with a negative relationship to negative numbers by multiplying it by (-1). This approach allows us to unify the effect's direction, meaning that an increase in the effect size always indicates better performance/outcome and consequently becomes straightforward to compare.

Studying the actual treatment effect of experiments in included studies we need to identify whether the subjects of each experiment actually received any reward or not. To separate the observations that correspond to control groups without any rewards we code a dummy variable *grp_reward*. Out of 1655 in total, we discard from the analysis 87 observations corresponding to control groups (5.55%).[1]

Besides these variables pivotal for our analysis, we systematically code other explanatory variables to include additional information. First, we include standard information of each study sub-sample such as the number of observations, sample size, name of the dependent variable, or whether the data are panel or cross-sectional. Furthermore, we control for the time span over which the particular experiment took place along with its average year; if it is performed in the lab or field environment; or whether crowding out intrinsic motivation theory appears in the study to control for utilization of this theory behind the motivation among researchers.[2] We also searched for the journal impact factor

[1] We do not perform any between-group analysis of differences in treatment and control groups due to large discrepancy in group sizes.

[2] Studies that does not mention this theory are Lazear (2000); Karlan & List (2007); Angrist & Lavy (2009); Kremer *et al.* (2009); Dohmen & Falk (2011); Fehr *et al.* (2013); Alberts *et al.* (2016), among others.

according to RePEc, and the number of citations for each study to control for additional publication information.[3]

Next, we code the basic characteristics of experimental design. By distinguishing whether the experiment has positive or negative framing we control for features of the reward scheme. We also take into account the nature of the task by coding whether it was cognitive or manual, measured quantitatively or qualitatively, and is appealing or not. We control also for the type of motivation subjects received during the task (altruism, trust, reciprocity fairness, monetary).

Lastly, we code subject-pool characteristics: whether it is made of students or a more general sample of the population (students, employees, or both); the gender of the subject pool expressed by the ratio of male to female subjects; the average age of the subjects; and whether the authors drew the subject-pools from a developed or developing country.

With these specifications, we gather 1568 estimates of the original effect of incentives on motivation from the years 2000 to 2019, collecting in our opinion a representative dataset covering the latest literature on our topic. The complete list of the variables and detailed reasoning behind the choice appears in Section 4.5 devoted to heterogeneity analysis. Our dataset yields more than 120,000 data points in total. We follow the reporting guidelines for meta-analysis compiled by Havranek *et al.* (2020).

The effect variables we code measure the relationship between incentives and output such as a change in physical/mental performance, pro-social behaviour, or students' Grade Point Average. This makes the collected estimates to be diverse in nature and size. We, therefore, need a measure that allows us to unify and compare the effects. Previous meta-analyses used mostly Cohen's d (e.g. Cameron & Pierce 1994; Jenkins *et al.* 1998) or Pearson correlation (Cerasoli *et al.* 2014; Van Iddekinge *et al.* 2018). Those measures are, however, inapplicable in our case. We would not be able to calculate Cohen's d for every data point since our dataset contains necessary estimates for control groups in only 13 studies out of the 44 (29.55%). Pearson correlation coefficient on the other hand does not control for confounding variables.

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 and would also control for omitted variables. A Partial correlation coefficient (PCC) is presumably the most fitting choice, appearing as a standard in numerous

[3] Data were collected through July 2020.

meta-analyses (e.g. Doucouliagos & Laroche 2003; Zhou *et al.* 2013; Valíčkova *et al.* 2015; 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 while ignoring the size of the dataset (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:

$$PCC = \frac{t}{\sqrt{t^2 + df}}, \tag{4.1}$$

where t stands for the t-statistic of the reported coefficient and df indicates the number of degrees of freedom in the estimation. The inclusion criteria we employ ensure we can code into the dataset both the original effect and its standard error for every study. This enables us to calculate the t-statistic for *all* collected observations, further establishing the PCC as the optimal choice for measuring the effect. We obtain the corresponding standard errors of the PCC following calculation:

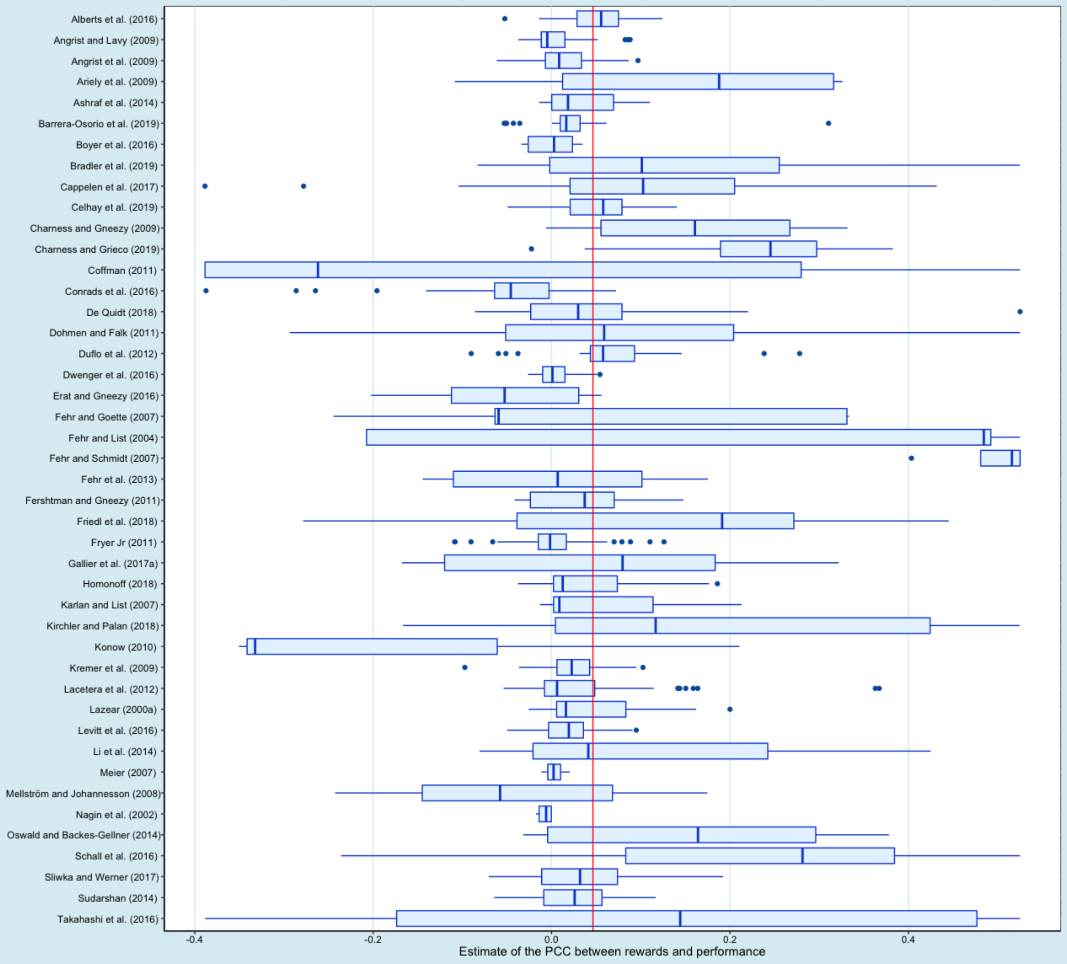
$$SE_{PCC} = \sqrt{\frac{(1 - PCC^2)}{df}}. \tag{4.2}$$

The distribution of PCCs shows several outliers. We address the potential influence of these outliers on our analysis by winsorizing at the 1% level. This level provides the best trade-off value between the necessary external intervention into the data and the results' stability.[4] Final partial correlation coefficients across individual studies after winsorization are shown in Figure 4.1.

To further understand the behaviour of the underlying effect in our data, we present the mean of said effect and the corresponding confidence intervals across various subsets of data in Table 4.2. Doucouliagos (2011) collects 22,141 estimates of Partial correlation coefficients in the field of economics and introduces guidelines for its reporting. He provide three bands of PCC sizes: 1) small with PCC below ±0.07, 2) Medium with PCC $\in [\pm 0.07; \pm 0.32]$, and 3) Large with PCC above ±0.33. The baseline effect in our case shows a mean Partial correlation coefficient of 0.046, which suggests a small incentive-motivation effect according to these guidelines. Looking systematically at the variables having the highest impact on mean statistics, we find a medium effect in some cases but

[4] The results are robust to changes in standard winsorization levels (1%, 2,5%, 5%).

Figure 4.1: Partial correlation coefficient across individual studies



Note: This figure shows a box plot of the partial correlation coefficient estimates across individual studies. Data winsorized at 1% level. PCC = Partial Correlation Coefficient.

never the Large one. We first highlight the increased effect size during game-based and work-based tasks (0.073 & 0.067). The Game variable specifically retains this above-average size even through the weighing procedure (0.085), proposing that the subjects show more effort during the typical controlled experiment in which a game is typically presented as a task. This finding appears to be backed up by the unusually large coefficient tied to the Lab study variable (0.091) belonging to the band of medium coefficients (Doucouliagos 2011). It equals almost three times its field counterpart even across weighted specifications (0.100). Next, the effect observed during the appealing tasks is more than 2.5 times larger than the effect observed during the non-appealing tasks (0.069 & 0.025). Weighting by the inverse of the number of estimates shows even more dispersed results (0.063 & 0.014) indicating that subjects perform better when the task is attractive. On contrary, non-appealing tasks drag the overall mean down. This difference is also highly statistically significant, suggesting that the task nature might substantially determine subjects' motivation. Last but not least the Reciprocity coefficient (0.100) remains relatively high even after the weighing procedure (0.110). This unusual increase may suggest that social influence plays a large role in determining the subjects' motivation during the experimental task. We provide a detailed discussion about the individual variables in Section 4.5.

4.4 Publication bias

Statistically significant results are easier to publish leaving the less significant effects to appear in the literature less often. The latter tend to be 'left in the drawer' unpublished, leading to a general overestimation of the reported effects (Stanley 2005). When an outlier appears in the data, it is feasible for the author to disregard it on an individual level. A problem appears when this 'byproduct of the research' occurs systematically, possibly leading to selective reporting or so-called publication bias.[5] Preference of researchers for statistical significance is becoming more present in recent literature since systematic and quantitative methods are used increasingly often (Rothstein *et al.* 2005).[6] The selective reporting of some estimates (typically those that are intuitive and statistically significant) has been identified as a serious threat to the credibility of empirical economics (Ioannidis *et al.* 2017). Nansen McCloskey & Ziliak (2019) compare selective reporting to the *Lombard effect*, in which speakers increase their vocal effort in the presence of the noise to outweigh it. To uncover the underlying patterns behind the effect of rewards on motivation, we focus in this chapter on detecting publication bias in our literature sample.

Even though several meta-analyses occur before our research in the existing literature, the publication bias is treated in only two of those we mention in Section 4.2. Cerasoli *et al.* (2014) use the so-called 'File drawer analysis' to correct for the publication bias. This technique indicates the number of unretrieved studies, averaging an effect size of zero, that would have to exist in

[5] Selective reporting is commonly called publication bias but this phenomenon is not limited to published papers only.

[6] Selective reporting was documented in various fields in economics, several examples are 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); Brodeur *et al.* (2016; 2020a); Tokunaga & Iwasaki (2017); Ugur *et al.* (2018; 2020); Campos *et al.* (2019); Blanco-Perez & Brodeur (2020); Matousek *et al.* (2022).

Table 4.2: Mean statistics across various subsets of data

		Unweighted			Weighted			
		Mean	95% conf. int.		Mean	95% conf. int.		No. of obs.
	All estimates	0.046	0.039	0.054	0.040	0.032	0.048	1568
Effect chars	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
Task nature	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
	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
Reward scheme	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
	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
	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
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
Study	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
Subject and country	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
Methodology and data	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
	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

Note: This table presents basic summary statistics of the partial correlation coefficients on various subsets of data. Weighted = weighted 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 4.7.

file drawers to reduce the effect size in question to one of two levels ($p > 0.10$ or $p > 0.05$). i.e. making the results of a meta-analysis insignificant.[7] Unfortunately, Rosenthal & Rubin (1988) argues why these results of File drawer analysis are not directly comparable to our analysis. Cameron & Pierce (1994) provide a measure of publication bias comparable to our findings. They use the effect sizes as a function of reward characteristics and display them in the funnel plots. They report little to no evidence of publication bias in their work. Considering the year of publication of Cameron & Pierce (1994), we find it very interesting to look for the publication bias in the rewards-motivation literature using a more recent literature sample together with more novel methods of its detection. In our search for publication bias, we follow the methodology of Zigraiova & Havránek (2016); Gechert *et al.* (2022), and Havránek *et al.* (2020), among others.

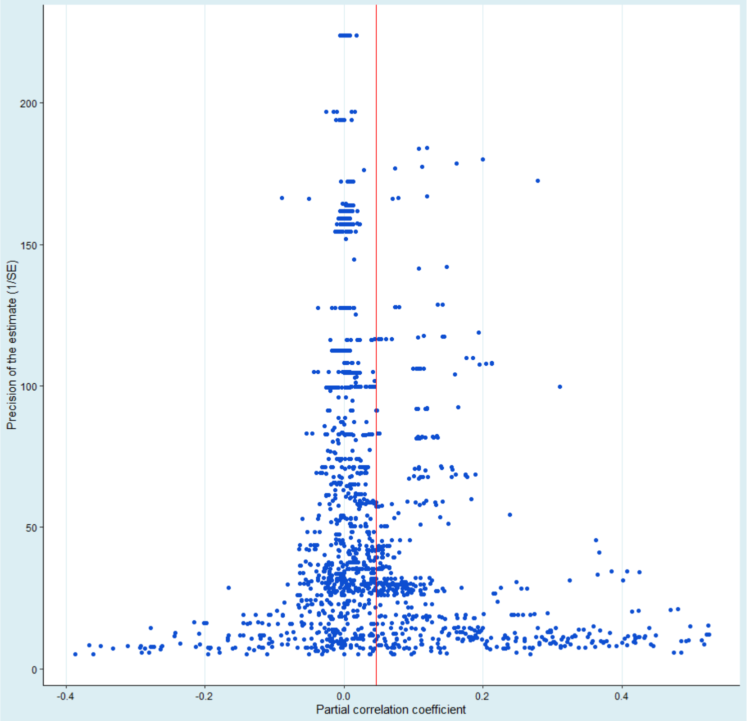
There are several perspectives one can look from at the effect of rewards on motivation. A purely economic point of view would suggest that people react to incentives and therefore rewards would positively affect motivation. People indeed provide performance—work—in exchange for a pay. A positive estimate of the effect of rewards on motivation and performance supports this idea. On contrary, the psychological theory of crowding out the intrinsic motivation by Deci (1971) says that rewards can have an undermining effect on one's intrinsic motivation. Looking at the publication bias from this perspective would suggest negative estimates of the underlying effect. There are both negative and positive estimates present in our dataset with an overall mean value of 0.046.

Selective reporting turns out as a correlation between point estimates and their standard errors. Following the standard approach of meta-analysis literature of Egger *et al.* (1997); Cazachevici *et al.* (2020); Matousek *et al.* (2022), we explore this correlation visually through the funnel plot. Estimates of the effect are plotted against the inverse of the standard error (precision). The higher the precision, the closer the estimates should be to the true underlying effect. The more imprecise the estimates are, the more scattered they shall appear, creating an inverted funnel. In case of a publication bias against negative or positive estimates that are in contrast with popular beliefs, the funnel plot would appear asymmetrical. In other cases, the plot becomes hollow at certain parts due to the omission of low magnitude insignificant effects. In our case, estimates of the effect are represented by partial correlation coefficients plotted against the standard error's precision that is calculated simply as $1/SE_{PCC}$. The right-hand side of the funnel plot in Figure 4.2 appears somewhat denser.

[7] File drawer analyses of Cerasoli *et al.* (2014) indicate that 586 studies reporting null findings would be necessary to reduce their population estimate to $p = 0.05$, suggesting the threat of inadequate search to be low in their case.

Large precise negative estimates are missing in contrast to their positive counterparts, indicating possible preferential reporting of positive effects of rewards on motivation in economic literature.

Figure 4.2: 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.

Nevertheless, conclusions based on a visual inspection are always subjective. To search for potential publication bias more rigorously, we employ a number of both linear and non-linear statistical tests. Shown to be excellent when it comes to publication bias detection (Stanley 2008; Moreno *et al.* 2009), we first proceed to the so-called funnel asymmetry test (Egger *et al.* 1997, FAT-PET). It aims to observe for potential correlation between the estimates and their standard errors through the following regression:

$$PCC_{ij} = \beta_0 + \beta_1 * (SE_{PCC})_{ij} + u_{ij}, \tag{4.3}$$

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 effect corrected for publication bias, β_1 captures the direction and magnitude of this bias, and u_{ij} represents the regression error term. In theory (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. In the following tables capturing the calculation results, we refer to β_0 as the 'Effect beyond bias' and β_1 as the 'Publication bias.'

Table 4.3 shows the results of the funnel asymmetry tests. We cluster the standard errors at the study level and assume exogeneity in the model. The first column shows the results of a simple OLS regression; the second and third columns account for unobserved heterogeneity by employing Fixed and Random effects estimators, respectively; the fourth column presents the results for different sizes of studies by weighing the equation by the inverse of the number of estimations collected from each study; the last column presents a weighted least squares estimator addressing the apparent heteroskedasticity of Equation 4.3 using precision as the weight (Ioannidis *et al.* 2017).

The tests confirm the hypothesis about publication bias, the estimates and standard errors are correlated at a high level of significance across four out of five estimations. Even though with a non-significant value, the publication bias drops when weighed by the number of estimates per study. This drop suggests that a certain studies may drive the publication bias. We can observe this also in Figure 4.2 where estimates are clustered around specific values. Looking further at the results in Table 4.3, FAT-PET suggest that the size of the true corrected value of the underlying effect should be much smaller than a simple average of estimates would indicate.

Table 4.3: Linear tests for publication bias

	OLS	**FE**	**RE**	**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 of estimating $PCC_{ij} = \beta_0 + \beta_1 * (SE_{PCC})_{ij} + u_{ij}$, where PCC_{ij} denotes the i-th partial correlation coefficient, and $(SE_{PCC})_{ij}$ its standard error observed in the j-th study. OLS = Ordinary Least Squares. FE = Fixed Effects. RE = Random Effects. Study = We weigh the estimates by the inverse of the number of observations reported per study. Precision = We weigh the estimates by the inverse of their standard error. Standard errors, clustered at the study level, are included in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

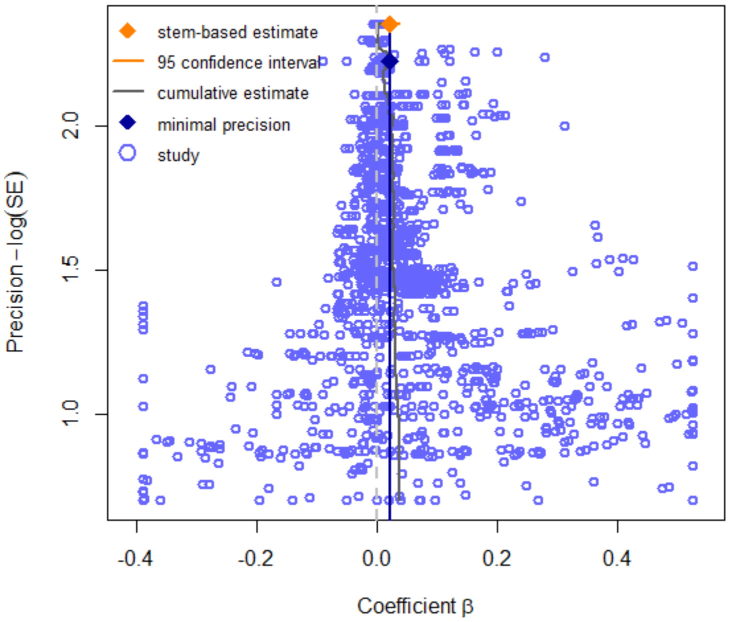
While the previous tests provide a good baseline in search for publication bias, they assume a linear relationship between the PCC and its standard error, which might lead to imprecise estimation if this assumption is not met. One should also be aware that the FAT-PET method tends to underestimate the 'true underlying effect' when it is different from zero (Stanley & Doucouliagos 2014; Bom & Rachinger 2019). Stanley (2008) have shown with the use of Monte Carlo simulations that linear approximations are not precise enough since selective reporting is a more complex function of the standard error. These imperfections, among others, are the reason we employ non-linear techniques to describe publication bias in rewards-motivation literature. We present the results of these tests in Table 4.4.

The first method we employ is the Weighted Average of Adequately Powered estimates (WAAP) technique proposed by Ioannidis *et al.* (2017). They suggest using unrestricted WLS only on estimates of those studies that are adequately powered with power level $> 80\%$. The method tests this condition by comparing the calculated standard errors to a power threshold defined using statistical significance level and adequate power level. As Ioannidis *et al.* (2017) explain 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 dataset, we find a total of 331 estimates, which satisfy the assumption of adequate power. WAAP estimate shows the underlying effect beyond the bias of 0.024, which is almost precisely at the average level of linear tests from Table 4.3 (0.023).

Stanley *et al.* (2010) suggest a very straightforward approach to testing for publication bias, specifically discarding 90% of the data and leaving only the 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 true effect and thus, it should be preferable to leave out most of the data and look only at the most precise part. In our dataset, that makes for a total of 157 observations. The top10 method slightly decreases the effect beyond bias to 0.019.

Furukawa (2019) builds on the Top10 method but suggests using a non-arbitrary portion of the most precise estimates out of the sample - the *stem* of the funnel plot. Furukawa (2019) seeks to minimize the mean square error optimizing thus the trade-off between efficiency and bias. As less precise estimates are included with a growing number of observations, the efficiency but also the bias increase. The threshold where observations are discarded is determined endogenously using observations with the lowest MSE. Figure 4.3 shows the implementation of the Stem method on our data sample.

Figure 4.3: 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. The dark blue diamond indicates the minimal precision above from which the model calculates the stem. The purple circles correspond to individual estimates of the partial correlation coefficient.

We further test for publication bias using the Selection model proposed by Andrews & Kasy (2019). It assumes that the probability of publication changes significantly when thresholds for the t-statistic predefined by the authors are reached. Andrews & Kasy (2019) suggest correcting the bias by utilizing conditional publication probability, which represents the probability of a study being published as a function of a study's results. The Selection model procedure estimates pure zero when utilizing t-distribution at the 5% significance level (t-statistic of 1.96).

Lastly, we employ the Endogenous Kink (EK) meta-regression model of 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) propose fitting a piece-wise linear regression of the collected estimates on their respective standard errors to identify the underlying effect. This method identifies the large portion of publication bias (0.879) at the overall effect, leaving the effect beyond bias generally small (0.013).

Table 4.4: Non-linear tests for publication bias

	WAAP	**Top10**	**Stem**	**Selection**	**Kink**
Publication bias					0.879***
					(0.153)
Effect beyond bias	0.024***	0.019***	0.021***	0.000	0.013***
	(0.003)	(0.004)	(0.007)	(0.001)	(0.002)

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 estimates. Top10 = Top10 Method. Stem = Stam-based method. Selection = Selection model estimate. Bias = Hierarchical bias model. Kink = Endogenous kink. Standard errors, clustered at the study level, are included in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

We present the estimation results of the non-linear methods in Table 4.4. In sum, the non-linear tests further support the results we obtained from the FAT-PET tests. Four of these non-linear methods suggest a minimal yet highly significant effect beyond bias, which would confirm the patterns in the behaviour we observed up to this point. As the first four models do not look for publication bias, we do not report this statistic for these models.

Our dataset consists of results obtained during experiments, each of which naturally contains unique methodological procedures that cause differences in the results of individual studies. Although we aim to control for these differences in many characteristics of the study design, some of these features might be hardly codified, unobserved and correlated not only with the reported effect but also with the reported standard error. We partially address this problem of our results being spurious by relaxing the exogeneity assumption we have been holding and we move on to testing for potential endogeneity of the standard errors.

Following Stanley (2005), Havranek (2015), or Matousek *et al.* (2022) we first use instrumental variables estimation as an alternative to funnel asymmetry testing and the most widely used correction for errors-in-variables bias. We choose the inverse of the square root of the number of observations as an instrument for the standard error since it correlates to the standard error by definition. The results of instrumental regression in Table 4.5 are not statistically significant but in line with our previous findings.

Next, we perform a p*uniform technique developed by van Aert & van Assen (2020). It 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 should exist in the sample. The results in Table 4.5 show evidence of publication bias that is, however, not significant. The mean corrected effect is 0.021 and significant at the 5% level.

To continue our search for publication bias, we conduct a method inspired by Gerber & Malhotra (2008) and Brodeur *et al.* (2020c) called the Caliper test. Typically employed in psychology, it tries to identify a systematic break related to publication bias at a particular psychologically important level (such as 0 for the point estimate or 1.96 for the t-statistic). This approach does not assume any relationship between the effect and its standard error. 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.

The t-statistic distribution in Figure 4.4 shows several peaks in reported t-values notably at 1.96 and -1.96 corresponding to 5% statistical significance levels. We confirm this visual suspicion empirically and examine the frequency of t-values reported around these with calipers 0.05, 0.1, and 0.2 points wide. The results of specified intervals are in Panel A of Table 4.6. We can interpret the values as the difference between the number of observations above and below the given threshold where the default value is the even distribution with 50% on each side. A coefficient of 0.370, for example, means that estimates above the threshold are 37.0 percentage points above the even distribution. The result therefore is 87.0% (50% + 37.0 %) of estimates above the threshold and 13.0% below. The tests suggest a significant discrepancy between the number of observations reported on each side of the threshold. There is a significant preference for results at the 5% significance level present in our sample.

Elliott *et al.* (2022), however, points out that in caliper tests and alike the researcher must arbitrarily specify the values where breaks in the distribution are expected. They derive two new rigorously founded techniques using a conditional chi-squared test that does not require the arbitrary definition of the breaks. Instead, the test slices the data to a specified number of bins and sets, therefore, the "bars" dynamically. The first technique is a histogram-based test for non-increasingness of the p-curve, the second one is a histogram-based test for 2-monotonicity and bounds on the p-curve and the first two derivatives. In their applications, Elliott *et al.* (2022) only focus on p-values below 0.15 and use 15, 30, or 60 bins. We set the target cutoff threshold for the discontinuity

Table 4.5: Relaxing the exogeneity assumption

	Instrumental Variables	**p-uniform***
Publication bias	0.194	2.17
	(2.696)	(0.14)
	{-5.09;5.47}	
Effect beyond bias	0.037	0.021***
	(0.121)	(0.001)
First-stage robust *F-stat*	0.35	
Studies	44	44
Observations	1568	1568

Note: IV = Instrumental Variable Regression; we use the inverse of the square root of the number of observations as an instrument for the standard error. Standard errors, clustered at the study level, are included in parentheses. In curly brackets we show the two-step weak-instrument-robust 95% confidence interval based on Andrews (2018) and Sun (2018). P-uniform* test for publication bias developed by van Aert & van Assen (2020); p-values are reported in parentheses. The test uses the maximum likelihood estimation. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4.6: Tests based on distributions of *t-statistics* and *p-values*

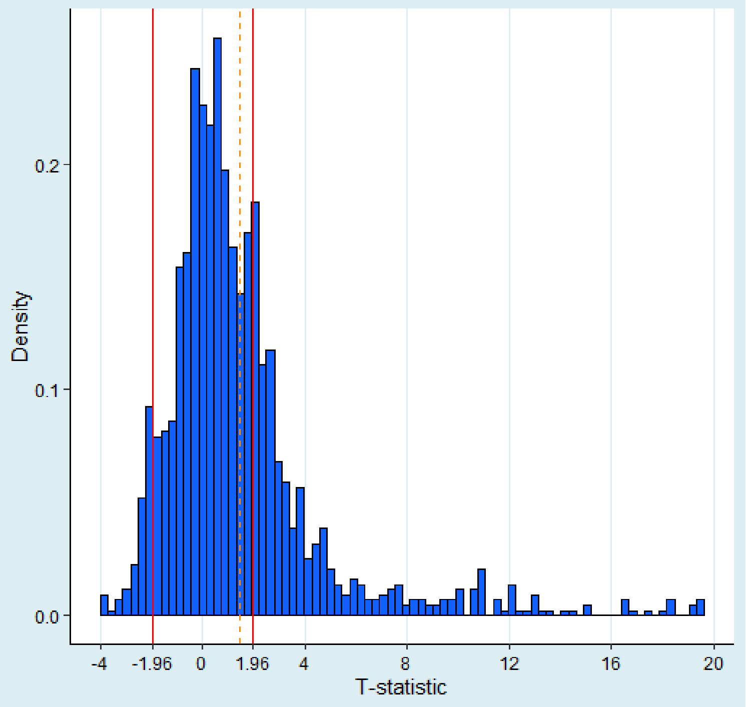
Panel A: Caliper test due to Gerber & Malhotra (2008)		
	Threshold 1.96	Threshold -1.96
Caliper width 5%	0.370***	-0.366***
	(0.038)	(0.053)
N	33	18
n_1/n_2	29 / 4	2 / 16
Caliper width 10%	0.352***	-0.329***
	(0.033)	(0.045)
N	48	28
n_1/n_2	41 / 7	5 / 23
Caliper width 20%	0.303***	-0.310***
	(0.023)	(0.032)
N	99	54
n_1/n_2	79 / 20	10 / 44
Panel B: Test due to Elliott *et al.* (2022)		
	Test for non-increasiveness	Test for monotonicity and bound
p-value	0.09	0.05
Observations ($p \leq 0.15$)	788	788
Total # observations	1655	1655

Note: Panel A shows the results of two sets of Caliper tests around t-statistic thresholds of 1.96 and -1.96. Caliper width 5% equals $t \in < 1.91; 2.01 >$ & $t \in < -2.01; -1.91 >$, caliper width 10% equals $t \in < 1.86; 2.06 >$ & $t \in < -2.06; -1.86 >$, caliper width 20% equals $t \in < 1.76; 2.16 >$ & $t \in < -2.16; -1.76 >$. N = Number of observations found in each of the respective intervals. n_1/n_2 = number of observations above and below the threshold, respectively, rounded to integers. Standard errors, clustered at the study level, are included in parentheses. Panel B reports tests developed by Elliott *et al.* (2022). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

test similarly at 0.15 while using equally distributed 30 bins due to the smaller dataset size. We present the results of the test in Panel B of Table 4.6. We reject the null hypothesis of the absence of the publication bias on 90% interval.

In sum, the vast majority of the tests for publication bias conducted in this section tell the same story: first, there appears to be a positive effect of rewards on motivation, and second, similarly to the rest of the empirical research in economics, the literature estimating the effect of rewards on motivation is affected by publication selection bias as well. We conclude with the average Partial correlation coefficient of the effects of rewards on motivation in our results being 0.023. Compared to the simple mean of PCC for all estimates reaching 0.046, our result is exactly half the size, suggesting that the strength (i.e. statistical significance) of the effect is exaggerated. Our findings, however, may be secretly correlated with hidden drivers we have not had the chance to reveal. Moreover, the effect certainly varies across experimental contexts and also individuals. Next, we thus turn our attention to uncovering the heterogeneity in our dataset.

Figure 4.4: T-statistic distribution



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

4.5 Heterogeneity

As discussed in previous sections, the substantial heterogeneity in the estimates of the effect of rewards on motivation reported in the experimental literature has already been stressed before. The research synthesis of this topic spans from works of Rummel & Feinberg (1988) and Wiersma (1992) to present one of Van Iddekinge *et al.* (2018). Figure 4.1 graphically illustrates this heterogeneity at the study level. To better understand the potential drivers behind the rewards-motivation effect, we now search for the sources of heterogeneity within our data and studies. We intend to quantitatively define the influence of numerous factors on the underlying effect. We try to explain the differences by regressing the partial correlation coefficients on their standard errors together with 32 additional explanatory variables that aim to capture the observable differences in the context of various experiments. Next, we aim to verify whether our previous findings of publication bias prove robust in controlling for heterogeneity. To both these ends, we employ model averaging.

4.5.1 Variables

Table 4.7 lists and explains the explanatory variables that we include in our meta-analysis. We include the description of each variable, its mean, and standard deviation. We divide the variables into the following categories that describe experimental as well as technical features of the studies: Effect Characteristics, Task nature, Reward scheme, Motivation, Study specifications, Subject and country characteristics, Estimation methodology and data, and Publication characteristics.

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 Section 4.2 for more details). On the contrary, Jenkins *et al.* (1998); Cerasoli *et al.* (2014), and Van Iddekinge *et al.* (2018) choose performance as their dependent variable. Cerasoli *et al.* (2014) is of particular interest to us, since it points out that most other studies operate only in laboratory settings, which makes it difficult to generalize their results to the real environment (e.g., in schools, work environment, among others).

Similar, substantial heterogeneity is immediately noticeable when looking at the sample of studies in our dataset. To give only a few examples: Sliwka & Werner (2017) try to observe whether varying wages have an effect on the speed of subjects in counting blocks of numbers during a laboratory experiment; Kremer *et al.* (2009) employ a large-scale experiment in Kenyan schools to see whether money can improve performance of students during academic exams; while Karlan & List (2007) test for charitable giving using mail solicitation to uncover the effects of money on altruistic behaviour. 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. This approach is heavily criticized in the field. For example, Glass *et al.* (1981) argue that conclusions drawn when generalizing different effects are invalid. Distributing the effects into too many categories would be on the other hand misleading for the reader as well as technically infeasible. The analysis would lose its point. Some degree of generalization is necessary for a meta-analysis. One of the latest examples of a meta-study with substantial heterogeneity in the estimates is e.g. DellaVigna & Linos (2022) who compare interventions in research units, versus at scale implemented in Nudge Units in governments. Fortunately, we observe a clear underlying pattern between the studies, which allows us to create a reasonable categorization according to their nature. Namely, we create four categories capturing: students' GPA, charitable giving, an outcome of a game or a simulation, and performance of employees at work. All of the studies collected in our data fit into one of these four categories, making this setup appear suitable. By our approach, we aim to choose a middle ground in the degree of generalization for enabling comparability and provide insights into the inner workings between effects of different nature.

Besides this critical feature, we code one technical variable regarding the effect characteristics—the 'Effect positive.' In short, it is a dummy variable capturing whether a higher effect is desirable or not. As we explained in Section 4.3, this approach lets us unify the effect's direction by transforming the PCC with a negative relationship to negative numbers. Furthermore, it allows us to distinguish whether the tasks measuring quantity/quality of the outcome are more frequent than those measuring a decrease in time, GPA, or similar. An exemplary setup of the positive effect appears in Dohmen & Falk (2011), where more numbers multiplied during a task equals better performance, while a great illustration of the negative effect is the experiment conducted by Fershtman & Gneezy (2011) where the subjects try to run a 60-meter race in the fastest time possible.

Task nature

Following previous meta-analyses, we code task specifics as to the important controls of experimental variations. Specifically, Tang & Hall (1995) focus on the task appeal by codifying an 'Interest level' among their five primary variables. Similarly to the purpose of our 'Appealing' variable, it serves to specify tasks that are of interest to the subjects. Jenkins *et al.* (1998), along with Rummel & Feinberg (1988); Deci *et al.* (1999); Cameron (2001) distinguish between extrinsic and boring tasks. A completely new outlook is proposed by Cerasoli *et al.* (2014), as they argue that most of the previously mentioned analyses study inherently exciting tasks. They further argue that 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 rather than quantitative ones (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. We take inspiration from this theory by distinguishing between quantitative and qualitative performance expressed in the variable 'Quantitative performance.'

Furthermore, we chose to distinguish between cognitive and manual tasks to capture additional nuances of some experiments—captured by the variable 'Cognitive task.' One such specification that a cognitive/manual setup allows is to classify 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 are featured in Takahashi *et al.* (2013). Furthermore, using the cognitive/manual distinction, we can also categorize the work in the 'employee' group by its 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).[8]

[8] With this setup, we initially expected to find either large values of Value Inflation factors (VIF) or correlation between the 'Cognitive task' variable and the 'Effect GPA,' where virtually all students should partake in cognitive tasks. However, we could not confirm any of these suspicions after a closer inspection of the model. We went through the dataset and

Reward scheme

If we look at the primary studies, they mainly distinguish between three major reward categories: if a reward is tangible or verbal—reward type, if one expects the reward or not—reward expectancy, and whether the subjects receive the reward for simply completing the task, completing it well, during a specific time frame, or given other specifications—reward contingency. In Section 4.2 we mention that some authors (Cameron & Pierce 1994; Tang & Hall 1995; Deci *et al.* 1999) suggest a simple task-contingency to have a detrimental effect on motivation, while verbal and unexpected rewards should do the opposite. Such classification, precisely or in part, is chosen by 9 out of the 11 primary meta-analyses. Two studies, however, opt for a different classification (Cerasoli *et al.* 2014; Van Iddekinge *et al.* 2018). Van Iddekinge *et al.* (2018) 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 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.

Given the wide range of the effects we capture, it seems only fitting to choose a setup inspired by Cerasoli *et al.* (2014) when defining the reward scheme for our meta-analysis. However, we decide to design the primary reward variable as follows: first, we denote or calculate the treatment group subjects' average earnings; next, we gather information about the monthly median household expenditure for each of the necessary countries;[9] using this data, we finally divide the logarithm of the treatment group average earning by the logarithm of the median monthly expenditure naming this payoff measure variable as

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 allows us to keep the specification mentioned above in the model.

[9] In some cases, we managed to obtain this statistic only in yearly intervals. In that case, we took the statistic for the year in which the study occurred, and divided it by 12. When the study took place over several years such as in the case of Kremer *et al.* (2009) or Lacetera *et al.* (2012), we use its mid-year and obtained the information for that year.

'Reward scaled.' This approach allows us to quantify the importance of the reward for the subjects.[10]

Looking at further specifications in this category, we also find noticeable heterogeneity in how the subjects receive the reward, which we codify into three variables. First, we target our attention to a fact whether the study rewards its subjects at all instead of punishing them and code this into the variable 'Positive framing.' Examples of positive framing are straightforward: Charness & Grieco (2019) observed how the creativity of subjects changes depending on financial rewards provided in various schemes (number of completed tasks, tournament etc.); Fershtman & Gneezy (2011) used a tournament in a 60-meter run among schoolchildren awarding them the prize for winning. In case of a negative setting, we can mention an experiment of Levitt *et al.* (2016) in which primary school students were offered various rewards that the authors withdrew at the end of the experiment if the students did not reach a threshold in the test; an introduction of a tax on the use of disposable plastic bags when shopping in Homonoff (2018); or motivation of people to charitable giving by Boyer *et al.* (2016).

'All paid' variable then indicates that all subjects participating in the experiment receive 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 behaviour accordingly, as noted in Greene (2018). Often the subjects are guaranteed to receive a certain payment if they voluntarily participate in a laboratory experiment, a show-up fee, such as in the case of Cappelen *et al.* (2017) or Bradler *et al.* (2019). The opposite, i.e. 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 Tao *et al.* 2014). Apart from the above-mentioned scholarship, we also chose this variable to control for lotteries that appear throughout our sample as a reward (Gallier *et al.* 2017). Here, a sizable one-off reward gets compensated with the low probability of receiving it, making the reward schemes more comparable.

The third variable tackling the reward heterogeneity is 'Reward own,' which equals one if the subjects receive the payoff for themselves. A nice example of

[10] We must note at this point that this procedure puts virtually all of the rewards our subjects receive into the category 'tangible,' where the literature predicts a strong 'undermining motivation' effect. According to our perspective, it points out the meaningfulness of experimental settings since economic experiments are based on the premise that subjects are motivated to participate in the experiment by a reward that should be substantial enough to attract them to the experiment.

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 behaviour during the experiment.

Motivation characteristics

A crucial metric in measuring the effect of rewards on motivation is the intrinsic motivation itself. Such motivation, however, may prove to be quite challenging to measure. One of the standard methods, used for example by Cameron & Pierce (1994); Wiersma (1992) or Cameron (2001), measures this motivation as free time spent on a task after the experimenter stops providing the subject with rewards, together with expressed self-interest in the task and willingness to participate in it without any reward.[11] An intriguing variable that might fit into this category would be to distinguish whether extrinsic incentives (i.e., rewards) were present or absent in this setting (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 it in our sample since virtually all studies include a group that received a monetary reward. A middle ground between the type of reward received also appears in some studies. Kirchler & Palan (2018) sometimes gives only a compliment during a food order as an extra reward by the researcher. The subjects are, however, still paid for the service provided, making the distinction unclear and likely impossible.

We opt for a different approach while analyzing subjects' motivation that lies behind their behaviour. We classify the experiments into five possible categories distinguishing the following five driving effects of motivation: *altruism*, *trust*, *reciprocity*, *fairness*, and *pure money*. This setup gives us an idea of whether the performance change is induced by sole monetary incentives such as in for example Gechert & Siebert (2020) or may have a different underlying driving force. To compare our approach to the literature we might choose, for example, a study of Ariely *et al.* (2009) who designs the experiment to observe the influence of monetary incentives on the interaction between the subjects' pro-social behaviour and their internal perspective of the task that are induced by these rewards. Moreover, even though it disregards the standard measure

[11] This reported self-interest is also the primary source of computations in primary meta-analyses that employ Cohen's d we have discussed in detail in Section 4.2.

of intrinsic motivation, the economic aspect of motivation gets highlighted instead.

We also stress an additional advantage of our approach since together with other categorical effect characteristics (such as 'Effect Charity' and 'Altruism' variables) this classification carries some new hidden information.[12] 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 behaviour. 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).

Study specifications

The most important study specification is its general estimation setup—whether the study employs a controlled laboratory experiment or a field experiment. Both Jenkins (1986) and Jenkins *et al.* (1998) suggest that the laboratory setting may yield more substantial effects than its counterpart making it thus a clear candidate for the influential response 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 vary a bit, such as when a blood donation is measured in Lacetera *et al.* (2012) compared to when Kirchler & Palan (2018) observe a change in the size of food obtained from a worker after a compliment.

Other characteristics account for the number of observations per study, the duration of the experiment in days, or the indicator of the year in which the experiment was carried out. This section also includes the variable 'Crowding-out' that 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.

[12] Even though there are similarities between some of these variables, the results of correlation and VIF tests make this setup appear suitable.

Subject and country characteristics

While searching for a standard set of variables to describe characteristics of a subject pool we find that several primary studies classify subjects according to age and education status. Tang & Hall (1995), for example, look at the range of subjects between preschool and college. 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 and Cerasoli *et al.* (2014) distinguish between four categories: *Child*, *Adolescent*, *College*, and *Adult*.

We took inspiration from already existing classifications and categorize subjects into similar groups. Contrary to previous approaches, we decide to merge various categories of students into one variable.[13] The remaining two categories of this variable are called 'Employees' and 'Mix,' which we deem to be the right trade-off between feasibility and comparability to other studies.

Next, we employ two useful descriptive classifications regarding the gender and age of the subjects. However, a substantial number of studies do not report data for our 'Gender' and 'Mid age' variables (such as Lacetera *et al.* (2012); Dwenger *et al.* (2016) or Dohmen & Falk (2011)), which results in a substantial amount of missing data points. Since the model averaging requires a data matrix of a full rank, we set the ratio of male/female for missing observations to 0.5, and for each missing observation in the 'Mid-age' variable we fill in the mean reported mid-age of each of the respective groups (students/mix/employees).[14]

Lastly, we create a variable controlling whether the country where the experiment took place is developed or not. An interesting observation is that a notable portion of the experiments in developing countries consists of measuring students' performance (Kremer *et al.* 2009; Duflo *et al.* 2012; Li Tao *et al.* 2014). No larger correlation or collinearity however appeared between

[13] The data showed a very high VIF suggesting thus a heavy collinearity with the initial categorization that involved separating students into groups spanning from preschool to middle school, high school, and college, . 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), who measure exam performance of Kenyan elementary school students.'

[14] Resulted mean reported mid ages: (students; mix; employees) = (16.5; 20; 24).

the respective variables as a consequence, suggesting that this is not always the case.

Estimation methodology and data

The researchers in our sample use mostly regression analysis to estimate the effects in their data (87%), leading us to differentiation into the seven most common methods we capture in respective variables described in Table 4.7. OLS is undoubtedly the most common method used in the studies. It 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). We classify regression methods appearing with the lowest frequency into the category 'Other method.' Celhay *et al.* (2019) or Angrist & Lavy (2009) who both uses two-stage least squares estimation may serve as representative examples of this category. The remaining cases report a simple mean statistic 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. Furthermore, we control for the type of data in the study by coding whether it is of a cross-sectional or a panel nature.

Publication characteristics

We use two proxies for measuring the quality of an included study. For each study we code the journal impact factor according to RePEc as well as the number of citations from Google Scholar.[15] By this we aim to control for additional publication information.

4.5.2 Model averaging

A lot has been written elsewhere about a typical feature of meta-regression analysis—a model uncertainty—as well as about shortcomings of conventional estimation methods of such models (e.g. Matousek *et al.* 2022; Cazachevici *et al.* 2020; Havranek *et al.* 2017, among others). To address this issue we turn our attention to a Bayesian model averaging (Raftery *et al.* 1997), a formal response to model uncertainty in the Bayesian setting. Bayesian model averaging (BMA) deals with the uncertainty by estimating a sequence of models with all

[15] Up to the point of data collection through July 2020.

Table 4.7: Definition and summary statistics of regression variables

Variable	Description	Mean	SD
PCC	Partial correlation coefficient (response variable)	0.046	0.136
Standard error	The standard error of the partial correlation coefficient	0.045	0.044
Effect characteristics			
Effect GPA	=1 if observed effect captured students' performance	0.344	0.475
Effect Charity	=1 if observed effect captured charitable giving	0.283	0.451
Effect Game	=1 if observed effect captured the outcome of a game	0.279	0.449
Effect positive	=1 if the relationship between rewards and the outcome is positive instead of negative	0.869	0.338
Task nature			
Appealing task	=1 if the task is appealing to the subjects instead of a non-appealing	0.482	0.500
Quan. performance	=1 if the measured performance was quantitative instead of a qualitative	0.702	0.457
Cognitive task	=1 if the task involved cognitive work instead of manual one	0.705	0.456
Reward scheme			
Reward scaled	The logarithm of the average payoff from the experiment divided by the logarithm of the median monthly expenditure in the corresponding country	0.599	0.292
Positive framing	=1 if the study rewards its subjects instead of punishing them	0.831	0.375
All paid	= 1 if all subjects received a reward (or punishment), = 0 if only some received it	0.741	0.438
Reward own	= 1 if the subjects received the reward for themselves instead of someone else received it	0.809	0.393
Motivation characteristics			
Altruism	= 1 if the subjects were motivated by altruism	0.291	0.454
Trust	= 1 if the subjects were motivated by trust	0.015	0.123
Reciprocity	= 1 if the subjects were motivated by reciprocity	0.103	0.304
Fairness	= 1 if the subjects were motivated by fairness	0.151	0.358
Monetary	= 1 if the subjects were motivated purely by money	0.440	0.497
Study specifications			
Lab study	= 1 if the experiment took place in a lab instead of a field	0.233	0.423
N. of obs.	The logarithm of the number of observations used	7.156	1.938
Time span	The logarithm of the number of days over which the experiment was carried out	4.183	2.642
Average Year	The logarithm of the average year of the experiment's time-span	7.606	0.002
Crowding-out	= 1 if crowding-out intrinsic motivation theory appears in the study	0.488	0.500
Subject and country characteristics			
Students	= 1 if the subjects were students	0.610	0.488
Employees	= 1 if the subjects were employees	0.072	0.259
Mix	= 1 if the subjects were a mix of these two	0.318	0.466
Gender	The logarithm of the ratio of male to female subjects (1 = all male, 0 = all female)	0.530	0.232
Mid age	The logarithm of the average year of the subjects	2.934	0.320
Developed country	= 1 if the corresponding country is developed instead of developing one	0.835	0.369
Estimation methodology and data			
OLS	= 1 if the authors use Ordinary Least Squares	0.571	0.495
Logit	= 1 if the authors use Logit regression	0.048	0.213
Probit	= 1 if the authors use Probit regression	0.090	0.286
Tobit	= 1 if the authors use Tobit regression	0.031	0.172
Fixed-effects	= 1 if the authors use Fixed-effects estimation	0.039	0.193
Random-effects	= 1 if the authors use Random-effects estimation	0.028	0.165
Diff-in-diff	= 1 if the authors use Difference-in-differences estimation	0.027	0.163
Other method	= 1 if the authors use a different method	0.037	0.189
Cross-sectional data	= 1 if the data is Cross-sectional instead of a Panel	0.446	0.497

Continued on next page

Table 4.7: Definition and summary statistics of regression variables (continued)

Variable	Description	Mean	SD
Publication characteristics			
Journal impact	The logarithm of the journal impact factor from RePEc	5.490	3.235
Study citations	The logarithm of the number of citations the study received	4.839	1.780

Note: This table presents the summary statistics and descriptions for each of the various study characteristics. Data of PCC and Standard errors are winsorized at 1% level. SD = standard deviation, GPA = grade point average.

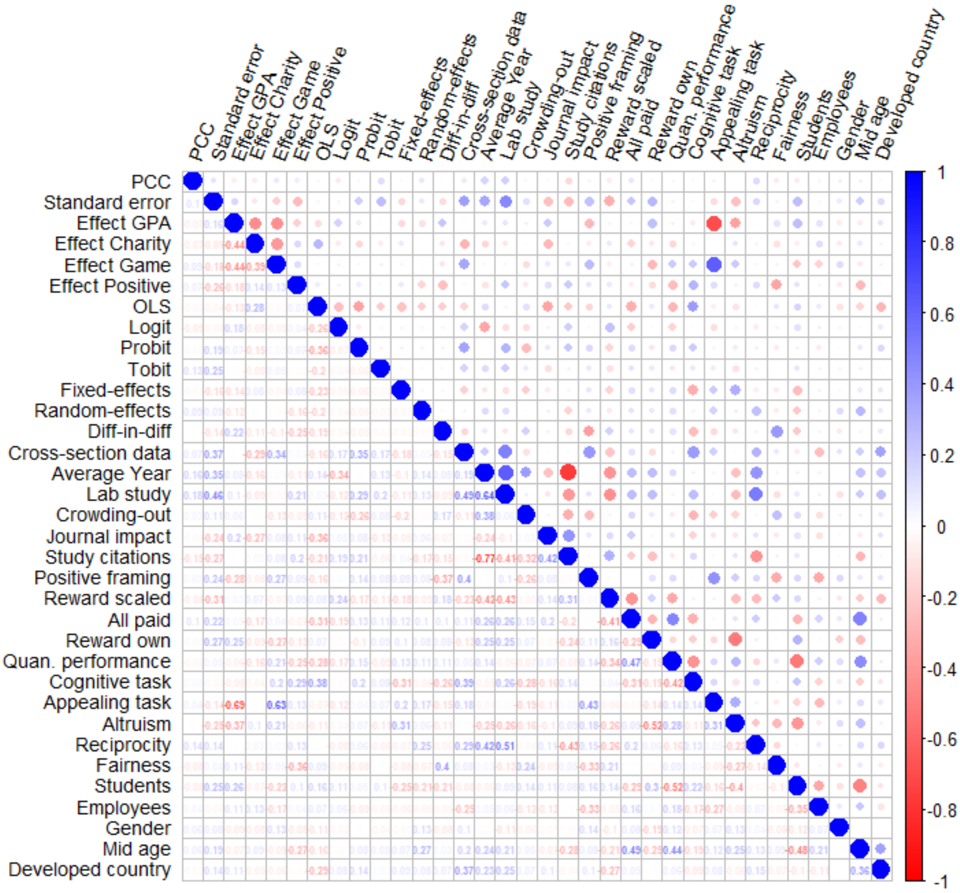
possible combinations of explanatory variables in the dataset and constructing a weighted average over the estimated coefficients across all these models. The weights used for averaging originate in Bayes' theorem and posterior model probabilities. Being an analogy of the information criteria from frequentist econometrics, Posterior model probabilities (PMPs) measure how well the particular model fits the data, conditional on model size. A crucial metric that BMA produces for each variable is the Posterior inclusion probability (PIP). It is the sum of the posterior model probabilities for the models in which the variable is included, and represents therefore the likelihood of the variable to be an important predictor of the effect. PIP can be viewed as a Bayesian analogy to statistical significance.[16]

Like Cazachevici *et al.* (2020), we run the analysis using the *bms* package in R (Feldkircher & Zeugner 2009) using the Markov chain Monte Carlo algorithm. Thanks to it, one can cut a considerable number of models from the analysis without losing information. Aiming to control for potential collinearity in the model we divert from a usual parameter setting of BMA analysis and decide to use a dilution prior (George 2010) rather than a more ordinary uniform model prior. The dilution prior tries to remedy the collinearity by multiplying the model probabilities with the determinant of the correlation matrix of the independent variables. It, therefore, assigns larger weights when the correlation between the variables is small since the determinant will then be close to 1. The weights get therefore smaller when a considerable correlation appears. We opt for the dilution model prior mainly due to the number of similar variables that may bring substantial collinearity into our model.

Before performing the actual BMA procedure, we first check for correlation

[16] More details on BMA, including a formal derivation, can be found in Raftery *et al.* (1997); Hoeting *et al.* (1999); Amini & Parmeter (2011) or Eicher *et al.* (2011).

Figure 4.5: Correlation matrix



Note: This figure displays the correlation table for variables reported in Table 4.7. Blue color (dark in grayscale) indicates positive correlation, while red color (light in grayscale) indicates negative correlation.

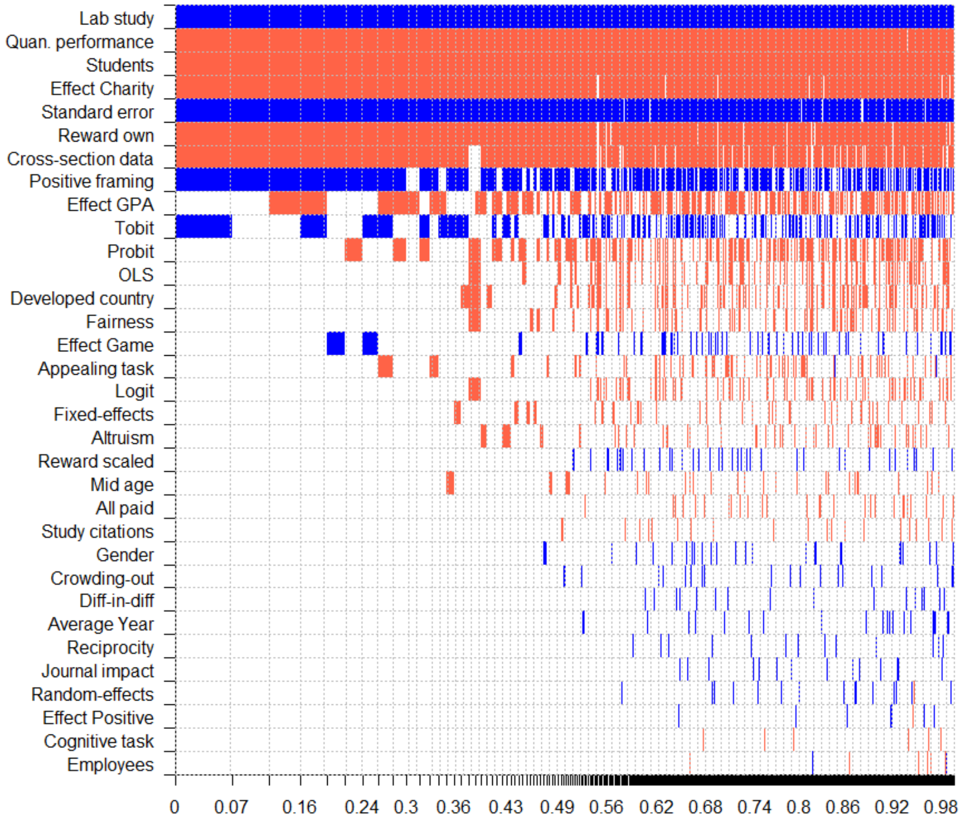
among our model variables and their VIFs. Figure 4.5 shows the correlations between the variables we consider. We set up a model using all the variables from the Table 4.7 and find the following two obstacles. First, a correlation appears between the number of observations and SE (this happens by definition because the number of observations is a factor used for calculating SE). As we prioritize keeping SE in the model we solve this by removing the number of observations. Second, a conflict arises between the dummy variable classifying data as cross-sectional/panel and the variable indicating the time span. As all of the cross-sectional data span only one day, this is only natural, and one variable implicitly carries the other one's information. When we look at the VIF for both of these variables, the cross-sectional/panel dummy displayed lower numbers, indicating a better implicit explanation of other variables in the model. We, therefore, cut the time span variable. Next, to preserve the model's integrity we remove the variable ’Trust’ from our dataset since its low number of observations (33; 1.9%) produced inconsistent results. Last but not least, we set the default variables for each necessary category and remove these to prevent perfect collinearity. We set the defaults as follows: for effect characteristics - effect work, for motivation characteristics - monetary, for subject and country characteristics - mix, for estimation methodology and data - other method. Finally, we move on to the actual estimation.

4.5.3 Results

The results of model averaging are displayed graphically in Figure 4.6, along with the numerical results and results of Frequentist Model Averaging (FMA) as a robustness check in Table 4.8. Besides the direction, size, and effect significance, we also present the Posterior Inclusion Probability for each variable that shows the importance of the variable to the average model and the likelihood of its appearance in the final model. When interpreting the importance of variables, we refer to the classification of 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 on the model. Similar to Havránek *et al.* (2020) and Gechert *et al.* (2022), we use Mallow’s criteria as weights for Frequentist Model Averaging (Hansen 2007) and employ the orthogonalization of the covariate space, as suggested by Amini & Parmeter (2012). We do this since the previously used Markov chain Monte Carlo algorithm for reducing the number of models in the computation is not applicable here.

Besides the intercept that is inherently included in all models, 9 variables have PIPs above 50%: the standard error, the dummy variable capturing student’s performance, the dummy variable capturing charitable giving in the experiment, the dummy variable for quantitatively measured performance, the dummy for positive framing of the experiment, the dummy variable capturing that the subjects received the reward for themselves, the dummy variable for the lab study, the dummy variable for students in the subject pool, and the dummy variable indicating cross-sectional data. We go through these results more closely in the remainder of this section. The first notable result of the BMA analysis, however, concerns publication bias. Our results show a strong effect of the standard error in the regression indicating a robust correlation between standard errors and PCCs even when we control for several additional characteristics of studies. Both the posterior mean in BMA and the point estimate in the FMA suggest that the correlation is strong, validating our previous findings of the presence of publication bias in the literature presented in the Section 4.4.

Figure 4.6: Model inclusion in Bayesian model averaging



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 4.8. For a detailed explanation of the variables, see Table 4.7.

Results for effect characteristics

Regarding the results in effects characteristics, we can further build on the previous paragraphs in Section 4.2 where we discussed to-date findings of the undermining effect of rewards on motivation in previous literature. Our results corroborate the findings of (e.g. Wiersma 1992; Cameron & Pierce 1994; Tang & Hall 1995; Cameron 2001) or Deci *et al.* (2001), who suggest that an undermining effect exists in experiments where the main effect is measured by a school performance task. 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 truly decreases the initial drive the subjects feel towards the activities they are otherwise doing out of their own will or for free. That seems to be true at least in specific contexts. Next, our results show that an undermining effect exists also in experiments where the observed effect is captured by charitable giving. This indicates that compared to a situation without any rewards, people decrease their effort even when the outcome of an experiment is measured by, say, the amount they donate to charity rather than keeping for themselves. This effect is further validated by the strong and negative coefficient of the variable 'reward own' described in Section 4.5.3. Finally, we further find no effect of rewards on motivation in experiments where is the performance measured by a game.

Results for task nature

Our results suggest that, unlike qualitatively measured tasks, the quantitatively measured ones have a strong and negative effect on the relationship between rewards and performance. This result is in line with the undermining effect and the theory suggested by Kruglanski *et al.* (1971) and Evans (1979) that predicts a stronger relationship between intrinsic motivation and qualitative tasks rather than quantitative ones. It is interesting to note that we do not see any relationship between appealing tasks and rewards-performance relationship in our data, such as observed for example by Tang & Hall (1995); Jenkins *et al.* (1998) or Cameron (2001). Subjects seem not to be motivated by interesting assignments. Furthermore, there is no preference between cognitive or manual tasks.

Results for reward scheme

Previous research shows that salience of the incentive is a strong predictor of performance and that more directly salient incentives narrow cognitive focus as well as encourage and intensify behaviour towards a particular goal (Cerasoli *et al.* 2014). Our variable 'Reward scaled,' by which we aim to quantify the

Table 4.8: Model averaging results

Response variable:	Bayesian model averaging (BMA)			Frequentist model averaging (FMA)		
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
Effect positive	0.000	0.001	0.011	-0.003	0.013	0.816
Task nature						
Appealing task	-0.003	0.010	0.123	-0.036	0.014	0.011
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
Reward scheme						
Reward scaled	0.002	0.010	0.062	0.011	0.023	0.628
Positive framing	0.038	0.024	**0.776**	0.036	0.019	0.068
All 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
Motivation characteristics						
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
Study specifications						
Lab study	0.081	0.013	**0.999**	0.100	0.020	0.000
Average Year	0.068	0.550	0.022	-2.449	3.559	0.491
Crowding-out	0.000	0.002	0.026	0.008	0.010	0.445
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
Estimation methodology and data						
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
Cross-sectional data	-0.059	0.021	**0.935**	-0.046	0.017	0.007
Publication characteristics						
Journal impact	0.000	0.000	0.017	0.001	0.002	0.424
Study citations	-0.000	0.001	0.036	-0.004	0.004	0.385

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 4.7.

importance of the reward for experimental subjects, however, does not affect motivation and performance. Our data, therefore, suggest that a scale of the reward is not of great importance when motivating people to better performance.

Positive framing of experimental tasks, on the other hand, increases the motivation of subjects to perform better in those tasks and consequently affects the outcomes of experiments. Our result is in contrast with Hossain & List (2012), who find that framing bonuses as losses improves the productivity of teams in a Chinese factory or with results of Levitt *et al.* (2016), who provide various incentives to influence the level of effort exerted by students in a low stakes testing environment and find suggestive evidence that rewards framed as losses outperform those framed as gains. Zero effects are, however, also present in the literature. List & Samek (2015) provides incentives for students to make healthy food choices and find no effects of framing in their experiment. We tend to lean towards the first literature stream. Our results indicate that a reward truly is more valuable for increasing performance than punishment.

Our results of the reward scheme section further suggest that it does not make a difference whether the rewards are awarded to all subjects. Group rewards do not affect the motivation of an individual. We do not confirm an argument of Greene (2018) that paying all subjects implies their awareness of a guaranteed future payment which then alters their future behaviour. A typical experimental feature—a show-up fee—should not, therefore, distort the results of rewards-motivation experiments.

If the subjects receive the reward for themselves, the effect appears slightly weaker than when someone else gets that reward such as in Mellström & Johannesson (2008). Contrary to our previous result obtained through the variable ’effect charity,’ the reward scheme here points out an altruistic behaviour of subjects and a charitable giving effect.

Results for study specifications

The strongest result of the BMA analysis reports the variable ’lab study’ that has, according to Kass & Raftery (1995), a decisive effect on the model. Its positive coefficient suggests that rewards do have a larger effect on motivation and performance when proposed in a laboratory rather than a field setting. This finding is in line with Jenkins (1986) and Jenkins *et al.* (1998), who predict a more substantial effect of rewards on motivation in a laboratory setting. The debate whether laboratory experiments are more eloquent than field settings is profound in all sub-fields of the experimental economics literature. Controlled experiments provide a natural framework for exploring various research questions in both laboratory and field conditions by enabling researchers to vary the parameters to infer the subject's preferences. However, despite decades of work and dozens of experiments of which we present a summary in this study, no consensus has emerged so far. Similarly to our result regarding the strength of the effect of rewards on motivation on subjects, Matousek *et al.* (2022) also finds a stronger effect in laboratory settings concluding that a controlled laboratory environment produces more evidence for impatience than a field study environment. Neither the average year of the experiment's time span nor the appearance of crowding out intrinsic motivation theory in the given study suggests an effect on motivation.

Results for subject and country characteristics

Our results suggest that almost none of our subject or country characteristics plays a role when rewarding people for better performance. The effects of rewards on performance should prove robust across gender, age groups as well as geographical locations. There is, however, one great exception—a crucial role pays when the experimental subjects are students. The negative coefficient of the 'students' variable then indicates a corroboration of the undermining effect theory of Deci (1971): introducing rewards decreases motivation and subsequently also the performance of students. Even though their participation in experiments is usually incentivized, the intrinsic motivation of students, or more precisely in our setting a motivation of subjects up to 20 years of age, may be a stronger predictor of their performance than extrinsic motivation in the form of, say, money.

Results for the remaining groups of variables and robustness checks

There is no evidence that an estimation methodology of included studies would affect the response variable, none of the estimation methods results in a PIP above 0.5. There is, however, a significant negative effect within the experiments featuring observations from subjects at a given point in time— cross-sectional data. This result may suggest that experiments conducted at multiple points in time, and therefore resulting in data with a panel nature, have a higher positive impact on motivation and performance than one-shot games. Subjects seem to be interested in repeated games with more rounds and exhibit such games also higher performance. We cannot, however, observe from our data whether this is caused more by the latter rounds, which could suggest that subjects learn and improve during the experiment, or by higher motivation stemming from joy out of continuous playing.

Two categories do not provide any variable that affects PCCs. Neither motivation nor publication characteristics seem to matter for the strength of the relationship between rewards and performance. Regardless of subjects being motivated by altruism, trust, reciprocity, or fairness, no motivator in our study shows a positive association with estimates of the PCCs. Both the number of citations and the journal quality proxied by the journal's impact factor do not matter for the estimated response variable.

Table 4.8 shows results of Frequentist model averaging as a robustness check to regular BMA analysis. Not counting the intercept, the results of FMA are in line with the BMA approach, presenting a similar direction/significance of the effect in virtually all of the highlighted variables. The FMA attributes but a small significance to the intercept. Next, our results prove robust to various specifications of the BMA model, namely different g-priors and model priors. We include the alternative BMA specifications, along with their respective results and a graphical representation of PIPs for all variables and different sets of model priors (Section C.1), in the Section C.1.

4.5.4 The best-practice estimates

Finally, we take the BMA model with resulting model coefficients from previous sections and plug in the variable statistics estimating thus the best-practice effect of rewards on motivation. We set the values of most variables conservatively to their sample means since the direction of the underlying effect is quite unclear. We are, however, able to identify several variables where the relationship seems to be clear and therefore we decide to set their values more strictly. We set the standard error equal to zero since having publication bias in the sample is not desirable. Next, we are looking for panel data, rather than cross-sectional, as we can retain more information from this approach. We set the average year to the value of the most recently published study considering this should best reflect current practice in the literature. Similarly, we set the number of citations and journal impact factor at their highest values as that suggests higher credibility of the estimates.

Moreover, from Model averaging results, we identify several scenarios that have interesting results. Those scenarios are represented by the response variables in BMA that have PIP higher than 0.5 and are at the same time of some experimentally conceptual nature. We identify 7 such scenarios: the effect being captured by student's performance (Effect GPA), the effect being captured by charitable giving (Effect charity), the measured performance was quantitative (Quantitative performance), when the study rewards its subjects instead of punishing them (Positive framing), if the subjects received the reward for themselves (Reward own), if the experiment took place in a laboratory (Lab study), when the experimental subjects were students (Students). We provide the best-practice effect of rewards on motivation within these contexts by consecutively setting the coefficients of respective variables equal to one in the basic best-practice estimate equation.

Besides the best-practice estimates, we compare our results to three actual studies from our dataset: i) Lazear (2000) represents a study with the highest amassed number of citations among all other studies, ii) Angrist & Lavy (2009) features the reward scheme with the highest possible payoff for the subjects, iii) and Takahashi *et al.* (2013) is the most representative of the whole dataset in terms of the experiment setup being it a straightforward game, where the motivation/performance is directly and easily comparable with the payoff scheme. We plug the actual values of the respective variables from these studies into the BMA model to obtain the best-practice estimate for each of them with one exception: we set the variable representing publication bias to zero to get unbiased results. Table 4.9 presents the results of the best-practice estimates along with the respective confidence intervals and percentage share of each estimated value to the basic best-practice estimate.[17] The results of most of these estimations are reported with considerably wide confidence intervals. The original underlying reported mean partial correlation coefficient is 0.046.

Finally, we take a closer look at the economic significance of nine variables, which the BMA model assigned a posterior inclusion probability of 0.5 or higher. Table 4.10 displays their ceteris paribus effect on the PCC. We calculate the effect for a change of one standard deviation as well as for a change from its minimum to its maximum value. Furthermore, we present this effect also as a percentage change in the best-practice estimate. Specifically, increasing the standard error by one standard deviation causes the PCC to increase by 24.07% of the best-practice estimate.

Three of the presented variables have a considerable positive effect on the PCC, while the remaining six pull the effect in the opposite direction. The

[17] Calculated using OLS with standard errors clustered at the study level.

Table 4.9: Implied best-practice

Estimate type	**Estimated value**	**95% confidence int.**	**Share of the BE value**
Best-practice estimate	0,079	(0,015 ; 0,144)	100%
Effect GPA	0,068	(0,001 ; 0,135)	86%
Effect charity	0,042	(-0,026 ; 0,110)	53%
Quant. performance	0,062	(0,001 ; 0,123)	78%
Positive framing	0,086	(0,021 ; 0,151)	108%
Reward own	0,070	(0,004 ; 0,137)	88%
Lab experiements	0,141	(0,075 ; 0,208)	178%
Students	0,054	(-0,015 ; 0,123)	68%
Takahashi et al. (2016)	0,071	(-0,034 ; 0,175)	89%
Lazear (2000a)	0,060	(-0,004 ; 0,124)	76%
Angrist et al. (2009)	0,019	(-0,014 ; 0,052)	24%

Note: The table reports overall *best-practice estimate* together with best-practice estimate (BE) according to three different studies. 95% confidence interval bounds are constructed using OLS with study level clustered standard errors.

Table 4.10: Significance of key variables

	One SD change		Maximum change	
	Effect on PCC	% of BP	Effect on PCC	% of BP
Standard error	0,0191	24,07%	0,0854	107,69%
Effect GPA	-0,0080	-10,10%	-0,0170	-21,43%
Effect Charity	-0,0234	-29,48%	-0,0520	-65,55%
Cross-section data	-0,0293	-36,99%	-0,0590	-74,38%
Lab study	0,0343	43,21%	0,0810	102,11%
Positive framing	0,0142	17,96%	0,0380	47,90%
Reward own	-0,0189	-23,81%	-0,0480	-60,51%
Quan. performance	-0,0270	-34,02%	-0,0590	-74,38%
Students	-0,0317	-39,97%	-0,0650	-81,94%

Note: This table presents ceteris paribus effect of key variables on the PCC. Only variables with PIP over 0.5 in the BMA model are included. *One SD change* implies how the PCC changes when we increase a variable value 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 4.7.

substantial influence of the standard error, which serves as a proxy for publication bias, is evident. We detected this bias mainly around the 5% significance level. However, as our analysis is the only one dealing with publication bias, we have no opportunity to compare our results to other papers. Besides publication bias, we point out two other extremes. The strong positive effect of the laboratory setting and the strong negative effect of the student subgroup. The results we have obtained are in line with both sides of the available theory. The resulting overall effect close to zero seems to stem from an interplay of two main factors. On one side, 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, reducing the overall performance change close to zero. This finding is in line with the one presented by Jenkins *et al.* (1998).[18]

4.6 Concluding remarks

We present a quantitative synthesis of the literature that uses experiments to discover the effect of incentives on people's motivation. 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, we aim to provide a synthesis separately from the economists' point of view. We collect 44 studies published strictly in economics journals and extract from them a total of 1568 estimates. We convert these estimates into partial correlation coefficients to unify the varying nature of the effects of rewards on motivation and to be able to compare them with each other. The mean partial correlation coefficient of the reported effect is 0.046. By employing a wide range of statistical tests, we observe a close to zero overall effect of incentives on performance while suggesting a presence of publication bias. We conclude that the mean drops to about 0.023 after we correct for publication bias. Rewards have, therefore, a smaller effect on motivation and performance of people than a simple mean summary statistics of economics literature suggests. Our estimated effect size corresponds to the findings presented 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 address substantial heterogeneity in the estimates of the effect of rewards on motivation. We define 39 different variables and search for an explanation of heterogeneity in our dataset with the Bayesian and Frequentist model averaging. We consider various model specifications, namely in the effect, method, study characteristics, reward scheme, task nature, motivation,

[18] We are aware of the shortcomings of our 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 behaviour in response to the incentives could allow for more detail. Similarly, there are both the groups that partook in a lab experiment and the groups observed at school for test performance coupled together in the student category. 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. Our reward scheme is also very straightforward since we focus on the economic aspect of rewards.

and other attributes. Finding the standard error being an important factor in the heterogeneity of reported estimates, we validate the presence of selective reporting in the literature once more. Conditional on the subjects receiving a reward, our results further show a significant positive relationship between performance and both the positive framing of the experiment and its execution in the laboratory environment. Positively framing the experiments implies a greater motivation and consequently performance in our setting. Conclusions in the existing literature, however, differ. Conclusions about the effect are on positive (Hossain & List 2012), negative (Levitt *et al.* 2016) and even zero (List & Samek 2015) ends of the spectrum. Corroborating the results of Jenkins (1986) and Jenkins *et al.* (1998), the rewards report a greater effect on motivation and performance in the laboratory rather than in the field settings. We advise great care when designing laboratory experiments since stronger effects might appear due to these settings. Several attributes correspond to the undermining effect theory of Deci (1971). These are namely the presence of charitable giving in the experiment, quantitative measure of performance, self-obtained rewards, a student sample used as experimental subjects, and the cross-sectional nature of the data, confirming further results of Kruglanski *et al.* (1971); Evans (1979). Our results prove robust to several model specifications and checks.

As a bottom line of our thesis, we present the best-practice estimates for different contexts. Results emphasize that different contexts incentives work differently in different contexts. To put our findings in contrast to the literature, we calculate the best-practice estimate also for three studies from our dataset. We observe a similar pattern of behaviour across these specifications. Last but not least, we support our results by a computation focused on the economic significance of the key variables from the Bayesian model averaging. We then quantify in both absolute numbers and percentage change the ceteris paribus influence of the publication bias together with the rest of the key variables of the best-practice estimate. We find a noticeable influence of the publication bias on the partial correlation coefficient, confirming thus our results from previous sections.

Lastly, we consider several potential caveats of our paper. Primarily, we are aware of the potential problems tied to the effect generalization (Glass *et al.* 1981), which we employ to transform the effect into a partial correlation coefficient. We partially remedy this flaw by categorizing the effect into four main categories. This approach, however, still does not eliminate the loss of information. Similarly, we simplify several variables which display high correlation numbers in the Bayesian model averaging, 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 a lower emphasis on the nature of the reward. It means that the rewards' effect appears in the model on an equal footing with the rest of the explanatory variables. We could have possibly employed a clearer distinction regarding the impact of different reward types on other variables.

Bibliography

Abbink, K., B. Irlenbusch, P. Pezanis-Christou, B. Rockenbach, A. Sadrieh, & R. Selten (2005): "An experimental test of design alternatives for the British 3G/UMTS auction." *European Economic Review* **49(2)**: pp. 505–530.

Abdellaoui, M., A. E. Attema, & H. Bleichrodt (2010): "Intertemporal Trade-offs for Gains and Losses: An Experimental Measurement of Discounted Utility." *The Economic Journal* **120(545)**: pp. 845–866.

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

van Aert, R. C. M. & M. A. L. M. van Assen (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]* .

Agranov, M. & L. Yariv (2015): "Collusion through Communication in Auctions."

Ainslie, G. (1975): "Specious reward: a behavioral theory of impulsiveness and impulse control." *Psychological bulletin* **82(4)**: pp. 463–496.

Al-Ubaydli, O. & J. A. List (2015): "Do Natural Field Experiments Afford Researchers More or Less Control Than Laboratory Experiments?" *American Economic Review* **105(5)**: pp. 462–466.

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

Altmejd, A., A. Dreber, E. Forsell, J. Huber, T. Imai, M. Johannesson, M. Kirchler, G. Nave, & C. Camerer (2019): "Predicting the replicability of social science lab experiments." *PLOS ONE* **14(12)**: pp. 1–18.

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

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

Andersen, S., G. Harrison, M. Lau, & E. Rutstrom (2010): "Preference Heterogeneity in Experiments: Comparing the Field and Laboratory." *Journal of Economic Behavior & Organization* **73(2)**: pp. 209–224.

Andersen, S., G. W. Harrison, M. I. Lau, & E. E. Rutström (2006): "Elicitation using multiple price list formats." *Experimental Economics* **9(4)**: pp. 383–405.

Andersen, S., G. W. Harrison, M. I. Lau, & E. E. Rutström (2008): "Eliciting risk and time preferences." *Econometrica* **76(3)**: pp. 583–618.

Andersen, S., G. W. Harrison, M. I. Lau, & E. E. Rutstrom (2013): "Discounting Behaviour and the Magnitude Effect: Evidence from a Field Experiment in Denmark." *Economica* **80(320)**: pp. 670–697.

Andersen, S., G. W. Harrison, M. I. Lau, & E. E. Rutström (2014): "Discounting behavior: A reconsideration." *European Economic Review* **71**: pp. 15–33.

Andreoni, J., M. A. Kuhn, & C. Sprenger (2015): "Measuring time preferences: A comparison of experimental methods." *Journal of Economic Behavior & Organization* **116**: pp. 451–464.

Andreoni, J. & C. Sprenger (2012a): "Estimating Time Preferences from Convex Budgets." *American Economic Review* **102(7)**: pp. 3333–3356.

Andreoni, J. & C. Sprenger (2012b): "Risk Preferences Are Not Time Preferences." *American Economic Review* **102(7)**: pp. 3357–3376.

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

ANDREWS, I. & M. KASY (2019): "Identification of and correction for publication bias." *American Economic Review* **109(8)**: pp. 2766–2794.

ANGRIST, J., D. LANG, & P. OREOPOULOS (2009): "Incentives and services for college achievement: Evidence from a randomized trial." *American Economic Journal: Applied Economics* **1(1)**: pp. 136–163.

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

ANTHOFF, D., R. S. J. TOL, & G. W. YOHE (2009): "Risk Aversion, Time Preference, and the Social Cost of Carbon." *Environmental Research Letters* **4(2)**: pp. 240–242.

ARIELY, D., A. BRACHA, & S. MEIER (2009): "Doing good or doing well? Image motivation and monetary incentives in behaving prosocially." *American Economic Review* **99(1)**: pp. 545–555.

ASHRAF, N., O. BANDIERA, & B. K. JACK (2014): "No margin, no mission? A field experiment on incentives for public service delivery." *Journal of public economics* **120**: pp. 1–17.

ASTAKHOV, A., T. HAVRANEK, & J. NOVAK (2019): "Firm Size And Stock Returns: A Quantitative Survey." *Journal of Economic Surveys* **33(5)**: pp. 1463–1492.

ATTEMA, A. E., H. BLEICHRODT, Y. GAO, Z. HUANG, & P. P. WAKKER (2016): "Measuring discounting without measuring utility." *American Economic Review* **106(6)**: pp. 1476–1494.

BABECKY, J. & T. HAVRANEK (2014): "Structural reforms and growth in transition." *The Economics of Transition* **22(1)**: pp. 13–42.

BACHRACH, Y. (2010): "Honor among thieves: collusion in multi-unit auctions." In "Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1," pp. 617–624. International Foundation for Autonomous Agents and Multiagent Systems.

BAJZIK, J., T. HAVRANEK, Z. IRSOVA, & J. SCHWARZ (2020): "Estimating the Armington elasticity: The importance of study design and publication bias." *Journal of International Economics* **127(C)**: p. 103383.

BÁNYAIOVÁ, L. (2012): "Opravdu mobilní operátoĹ™i zakopali válečnou sekeru?" *Antitrust* **4/2012**: p. 4.

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

BAUER, M. & J. CHYTILOVÁ (2010): "The Impact of Education on Subjective Discount Rate in Ugandan Villages." *Economic Development and Cultural Change* **58(4)**: pp. 643–669.

BAUER, M. & J. CHYTILOVÁ (2013): "Women, Children and Patience: Experimental Evidence from Indian Villages." *Review of Development Economics* **17(4)**: pp. 662–675.

BAUER, M., J. CHYTILOVA, & J. MORDUCH (2012): "Behavioral Foundations of Microcredit: Experimental and Survey Evidence from Rural India." *American Economic Review* **102(2)**: pp. 1118–1139.

BENZION, U., A. RAPOPORT, & J. YAGIL (1989): "Discount rates inferred from decisions: An experimental study." *Management science* **35(3)**: pp. 270–284.

BICHLER, M., J. GOEREE, S. MAYER, & P. SHABALIN (2014): "Spectrum auction design: Simple auctions for complex sales." *Telecommunications Policy* **38(7)**: pp. 613–622.

BICHLER, M. & J. K. GOEREE (2017): "Frontiers in spectrum auction design." *International Journal of Industrial Organization* **50**: pp. 372–391.

BICHLER, M., V. GRETSCHKO, & M. JANSSEN (2017): "Bargaining in spectrum auctions: A review of the German auction in 2015." *Telecommunications Policy* **41(5-6)**: pp. 325–340.

BICHLER, M., P. SHABALIN, & J. WOLF (2013): "Do core-selecting Combinatorial Clock Auctions always lead to high efficiency? An experimental analysis of spectrum auction designs." *Experimental Economics* **16(4)**: pp. 511–545.

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

Bleichrodt, H., Y. Gao, & K. I. M. Rohde (2016): "A measurement of decreasing impatience for health and money." *Journal of Risk and Uncertainty* **52(3)**: pp. 213–231.

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

Bonner, S. E., R. Hastie, G. B. Sprinkle, & S. M. Young (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)**: pp. 19–64.

Booij, A. S. & B. M. S. van Praag (2009): "A Simultaneous Approach to the Estimation of Risk Aversion and the Subjective Time Discount Rate." *Journal of Economic Behavior & Organization* **70(1-2)**: pp. 374–388.

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

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

Brodeur, A., N. Cook, & A. Heyes (2020a): "A Proposed Specification Check for p -Hacking." *AEA Papers and Proceedings* **110**: pp. 66–69.

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

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

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

Brown, A., T. Imai, F. Vieider, & C. F. Camerer (2021): "Meta-Analysis of Empirical Estimates of Loss-Aversion." *mimeo*, LMU Munich.

Brown, A. L., Z. E. Chua, & C. F. Camerer (2009): "Learning and Visceral Temptation in Dynamic Saving Experiments." *The Quarterly Journal of Economics* **124(1)**: pp. 197–231.

Brunner, C., J. K. Goeree, C. A. Holt, & J. O. Ledyard (2010): "An Experimental Test of Flexible Combinatorial Spectrum Auction Formats." *American Economic Journal: Microeconomics* **2(May 2006)**: pp. 39–57.

Burks, S., J. Carpenter, L. Gotte, & A. Rustichini (2012): "Which Measures of Time Preference Best Predict Outcomes: Evidence from a Large-Scale Field Experiment." *Journal of Economic Behavior & Organization* **84(1)**: pp. 308–320.

Burtraw, D., J. Goeree, C. A. Holt, E. Myers, K. Palmer, & W. Shobe (2009): "Collusion in auctions for emission permits: An experimental analysis." *Journal of Policy Analysis and Management* **28(4)**: pp. 672–691.

Cairns, J. & M. der Pol (1997): "Constant and decreasing timing aversion for saving lives." *Social Science & Medicine* **45(11)**: pp. 1653–1659.

Cairns, J. A. (1992): "Health, wealth and time preference." *Project Appraisal* **7(1)**: pp. 31–40.

Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen, & H. Wu (2016): "Evaluating replicability of laboratory experiments in economics." *Science* **351(6280)**: pp. 1433–1436.

Camerer, C. F., A. Dreber, F. Holzmeister, T. H. Ho, J. Huber, M. Johannesson, M. Kircher, G. N. G, B. A. Nosek, T. Pfeiffer, A. Altmejd, N. Buttrick, T. Chan, Y. Chen, E. Forsell, A. Gampa, E. Heikensten, L. Hummer, T. Imai, S. Isaksson, D. Manfredi,

J. Rose, E. J. Wagenmakers, & H. Wu (2018): "Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015." *Nature Human Behaviour* **2**: pp. 637–644.

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

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

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

Campbell, J. Y. (1994): "Inspecting the mechanism: An analytical approach to the stochastic growth model." *Journal of Monetary Economics* **33(3)**: pp. 463–506.

Campos, N. F., J. Fidrmuc, & I. Korhonen (2019): "Business cycle synchronisation and currency unions: A review of the econometric evidence using meta-analysis." *International Review of Financial Analysis* **61**: pp. 274–283.

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

Carlsson, F., H. He, P. Martinsson, P. Qin, & M. Sutter (2012): "Household Decision Making in Rural China: Using Experiments to Estimate the Influences of Spouses." *Journal of Economic Behavior & Organization* **84(2)**: pp. 525–536.

Carter, J. R. & M. D. Irons (1991): "Are Economists Different, and If So, Why?" *Journal of Economic Perspectives* **5(2)**: pp. 171–177.

Cassar, A., A. Healy, & C. von Kessler (2017): "Trust, Risk, and Time Preferences After a Natural Disaster: Experimental Evidence from Thailand." *World Development* **94(C)**: pp. 90–105.

Castillo, M., P. J. Ferraro, J. L. Jordan, & R. Petrie (2011): "The Today and Tomorrow of Kids: Time Preferences and Educational Outcomes of Children." *Journal of Public Economics* **95(11)**: pp. 1377–1385.

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

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

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

Chabris, C., D. Laibson, C. Morris, J. Schuldt, & D. Taubinsky (2008a): "Individual Laboratory-Measured Discount Rates Predict Field Behavior." *Technical Report 2–3*, National Bureau of Economic Research, Cambridge, MA.

Chabris, C., D. Laibson, C. Morris, J. Schuldt, & D. Taubinsky (2008b): "Measuring intertemporal preferences using response times." *Technical report*, National Bureau of Economic Research, Cambridge, MA.

Chabris, C. F., D. Laibson, C. L. Morris, J. P. Schuldt, & D. Taubinsky (2009): "The Allocation of Time in Decision-Making." *Journal of the European Economic Association* **7(2-3)**: pp. 628–637.

Chapman, G. B. (1996): "Temporal discounting and utility for health and money." *Journal of Experimental Psychology: Learning, Memory, and Cognition* **22(3)**: p. 771.

Chapman, G. B. & A. S. Elstein (1995): "Valuing the Future." *Medical Decision Making* **15(4)**: pp. 373–386.

Chapman, G. B., R. Nelson, & D. B. Hier (1999): "Familiarity and time preferences: Decision making about treatments for migraine headaches and Crohn's disease." *Journal of Experimental Psychology: Applied* **5(1)**: pp. 17–34.

Chapman, G. B. & J. R. Winquist (1998): "The magnitude effect: Temporal discount rates and restaurant tips." *Psychonomic Bulletin & Review* **5(1)**: pp. 119–123.

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

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

Chesson, H. & W. K. Viscusi (2000): "The heterogeneity of time–risk tradeoffs." *Journal of Behavioral Decision Making* **13(2)**: pp. 251–258.

Cheung, S. L. (2016): "Recent Developments in the Experimental Elicitation of Time Preference." *Journal of Behavioral and Experimental Finance* **11(C)**: pp. 1–8.

Choi, J., G. G. Lim, & K. C. Lee (2010): "An Agent Based Market Design Methodology for Combinatorial Auctions." *Journal of Artificial Societies and Social Simulation (Jasss)* **13(2)**.

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

Cohen, J., K. M. Ericson, D. Laibson, & J. M. White (2020): "Measuring Time Preferences." *Journal of Economic Literature* **58(2)**: pp. 299–347.

Coller, M. & M. B. Williams (1999): "Eliciting Individual Discount Rates." *Experimental Economics* **2(2)**: pp. 107–127.

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

Cramton, P. (2013): "Spectrum Auction Design." *Review of Industrial Organization* **42(2)**: pp. 161–190.

Crawford, V. (1998): "A Survey of Experiments on Communication via Cheap Talk." *Journal of Economic Theory* **78(2)**: pp. 286–298.

Croson, R. & S. Gächter (2010): "The science of experimental economics." *Journal of Economic Behavior & Organization* **73(1)**: pp. 122–131.

CTO (2012a): "Cesky telekomunikacni urad zverejnuje navrh analyzy relevantniho trhu c. 8 (The Czech Telecommunication Office publishes the draft analysis of the relevant market No. 8)." *Technical report*, Cesky telekomunikacni urad (Czech Telecommunications Office), Prague.

CTO (2012b): "Vyhlaseni vyberoveho rizeni za ucelem udeleni prav k vyuzivani radiovych kmitoctu k zajisteni verejne komunikacni site v pasmech 800 MHz, 1800 MHz a 2600 MHz (Announcement of a tender for the purpose of granting rights to use radio frequencies to ensure a." *Technical report*, Cesky telekomunikacni urad (Czech Telecommunications Office), Prague.

CTO (2013a): "Cesky telekomunikacni urad rozhodl o zastaveni aukce (Czech Telecommunications Office decides to stop the auction)." *Technical report*, Cesky telekomunikacni urad (Czech Telecommunications Office), Prague.

CTO (2013b): "Informace o ukonceni aukce a o rozdeleni radiovyych kmitoctu Vitezum aukce ziskanych v ramci aukce (Information on the end of the auction and on the distribution of radio frequencies The winner of the auction obtained within the auction)." *Technical report*, Cesky telekomunikacni urad (Czech Telecommunications Office), Prague.

CTO (2013c): "Informace o ukonceni aukcni faze vyberoveho rizeni a informace o vysledcich aukcni faze (Information on the end of the auction phase of the tender and information on the results of the auction phase)." *Technical report*, Cesky telekomunikacni urad (Czech Telecommunications Office), Prague.

CTO (2013d): "Vyhlaseni vyberoveho rizeni za ucelem udeleni prav k vyuzivani radiovych kmitoctu k zajisteni verejne komunikacni site v pasmech 800 MHz, 1800 MHz a 2600 MHz (Announcement of a tender for the purpose of granting rights to use radio frequencies to ensure a." *Technical report*, Cesky telekomunikacni urad (Czech Telecommunications Office), Prague.

CTO (2017): "ÚOHS provede sektorové šetĹ™ení mobilních operátorĹŻ (Czech Telecommunications Office will carry out a sector inquiry of mobile operators)."

Davis, D. D. & C. A. Holt (1993): *Experimental economics.* Princeton university press.

De Long, J. B. & K. Lang (1992): "Are all economic hypotheses false?" *Journal of Political Economy* **100(6)**: pp. 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)**: pp. 522–559.

De Vries, S. & R. V. Vohra (2003): “Combinatorial auctions: A survey.” *INFORMS Journal on computing* **15(3)**: pp. 284–309.

Dean, M. & A. Sautmann (2021): “Credit Constraints and the Measurement of Time Preferences.” *Review of Economics and Statistics* **103(1)**: pp. 119–135.

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

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

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

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

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

Deck, C. & S. Jahedi (2015a): “An experimental investigation of time discounting in strategic settings.” *Journal of Behavioral and Experimental Economics* **54**: pp. 95–104.

Deck, C. & S. Jahedi (2015b): “Time discounting in strategic contests.” *Journal of Economics & Management Strategy* **24(1)**: pp. 151–164.

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

Depositario, D. P. T., R. M. Nayga, X. Wu, & T. P. Laude (2009): “Should students be used as subjects in experimental auctions?” *Economics Letters* **102(2)**: pp. 122–124.

DOHMEN, T. & A. FALK (2011): "Performance pay and multidimensional sorting: Productivity, preferences, and gender." *American Economic Review* **101(2)**: pp. 556–590.

DOLAN, P. & C. GUDEX (1995): "Time preference, duration and health state valuations." *Health economics* **4(4)**: pp. 289–299.

DOUCOULIAGOS, C. (2011): "How large is large? Preliminary and relative guidelines for interpreting partial correlations in economics." *Working papers*, Deakin University, Department of Economics.

DOUCOULIAGOS, C. & P. LAROCHE (2003): "What do unions do to productivity? A meta-analysis." *Industrial Relations* **42**: pp. 650–691.

DOUCOULIAGOS, C. & T. D. STANLEY (2013): "Are all economic facts greatly exaggerated? Theory competition and selectivity." *Journal of Economic Surveys* **27(2)**: pp. 316–339.

DOUCOULIAGOS, H. & M. PALDAM (2011): "The ineffectiveness of development aid on growth: An update." *European Journal of Political Economy* **27(2)**: pp. 399–404.

DREYFUS, M. K. & W. K. VISCUSI (1995): "Rates of Time Preference and Consumer Valuations of Automobile Safety and Fuel Efficiency." *The Journal of Law and Economics* **38(1)**: pp. 79–105.

DUAN, J., K. K. DAS, L. MERILUOTO, & W. R. REED (2020): "Estimating the effect of spillovers on exports: a meta-analysis." *Review of World Economics* **156(2)**: pp. 219–249.

DUFLO, E., R. HANNA, & S. P. RYAN (2012): "Incentives work: Getting teachers to come to school." *American Economic Review* **102(4)**: pp. 1241–1278.

DUQUETTE, E., N. HIGGINS, & J. HOROWITZ (2012): "Farmer Discount Rates: Experimental Evidence." *American Journal of Agricultural Economics* **94(2)**: pp. 451–456.

DWENGER, N., H. KLEVEN, I. RASUL, & J. RINCKE (2016): "Extrinsic and intrinsic motivations for tax compliance: Evidence from a field experiment in Germany." *American Economic Journal: Economic Policy* **8(3)**: pp. 203–232.

EASTERBROOK, P. J., R. GOPALAN, J. A. BERLIN, & D. R. MATTHEWS (1991): "Publication bias in clinical research." *The Lancet* **337(8746)**: pp. 867–872.

EGGER, M., G. DAVEY SMITH, M. SCHNEIDER, & C. MINDER (1997): "Bias in meta-analysis detected by a simple, graphical test." *British Medical Journal* **315(7109)**: pp. 629–34.

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

EISENBERGER, R., W. D. PIERCE, & J. CAMERON (1999): "Effects of reward on intrinsic motivationâ€"Negative, neutral, and positive: Comment on Deci, Koestner, and Ryan (1999)." *Psychological Bulletin* **125(6)**: pp. 677–691.

ELLIOTT, G., N. KUDRIN, & K. WÜTHRICH (2022): "Detecting p-Hacking." *Econometrica* **90(2)**: pp. 887–906.

ERAT, S. & U. GNEEZY (2016): "Incentives for creativity." *Experimental Economics* **19(2)**: pp. 269–280.

EUROPEAN COMMISSION (2016): "Commission opens formal investigation into mobile telephone network sharing in Czech Republic."

EVANS, G. W. (1979): "Behavioral and Physiological Consequences of Crowding in Humans 1." *Journal of applied social psychology* **9(1)**: pp. 27–46.

FARNIA, F., C. BEAUDRY, F. FARNIA, J.-m. FRAYRET, L. LEBEL, & C. BEAUDRY (2015): "Agent-Based Simulation of Multi-Round Timber Combinatorial Auction." *CIRRELT* **CIRRELT-20(May)**.

FEARNSIDE, P. M. (2002): "Time preference in global warming calculations: A proposal for a unified index." *Ecological Economics* **41(1)**: pp. 21–31.

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

FEHR, E., H. HERZ, & T. WILKENING (2013): "The lure of authority: Motivation and incentive effects of power." *merican Economic Review* **103(4)**: pp. 1325–1359.

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

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

FELDKIRCHER, M. & S. ZEUGNER (2009): "Benchmark Priors Revisited:on Adaptive Shrinkage and the Supermodel Effect in Bayesian Model Averaging." *IMF Working Papers* **09(202)**: p. 1.

FERNANDEZ, C., E. LEY, & M. F. J. STEEL (2001): "Benchmark Priors for Bayesian Model Averaging." *Journal of Econometrics* **100(2)**: pp. 381–427.

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

FIELD, E., R. PANDE, J. PAPP, & N. RIGOL (2013): "Does the Classic Microfinance Model Discourage Entrepreneurship Among the Poor? Experimental Evidence from India." *American Economic Review* **103(6)**: pp. 2196–2226.

FINKE, M. S. & S. J. HUSTON (2013): "Time preference and the importance of saving for retirement." *Journal of Economic Behavior & Organization* **89(C)**: pp. 23–34.

FISCHBACHER, U. (2007): "z-Tree: Zurich toolbox for ready-made economic experiments." *Experimental Economics* **10(2)**: pp. 171–178.

FISHER, I. (1930): "The theory of interest." *New York* **43**.

FOLEY, D. K., A. REZAI, & L. TAYLOR (2013): "The social cost of carbon emissions: Seven propositions." *Economics Letters* **121(1)**: pp. 90–97.

FRANK, R. H., T. GILOVICH, & D. T. REGAN (1993): "Does Studying Economics Inhibit Cooperation?" *Journal of Economic Perspectives* **7(2)**: pp. 159–171.

FREDERICK, S., G. LOEWENSTEIN, & T. O'DONOGHUE (2002): "Time Discounting and Time Preference: A Critical Review." *Journal of Economic Literature* **40(2)**: pp. 351–401.

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

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

FUJII, T. & L. KARP (2008): "Numerical analysis of non-constant pure rate of time preference: A model of climate policy." *Journal of Environmental Economics and Management* **56(1)**: pp. 83–101.

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

FURUKAWA, C. (2021): "Publication Bias under Aggregation Frictions: From Communication Model to New Correction Method." *Working paper*, MIT.

GALLIER, C., C. REIF, & D. RÖMER (2017): "Repeated pro-social behavior in the presence of economic interventions." *Journal of behavioral and experimental economics* **69**: pp. 18–28.

GECHERT, S., T. HAVRANEK, Z. IRSOVA, & D. KOLCUNOVA (2022): "Measuring capital-labor substitution: The importance of method choices and publication bias." *Review of Economic Dynamics* **45**: pp. 55–82.

GECHERT, S. & J. SIEBERT (2020): "Preferences over wealth: Experimental evidence." *Journal of Economic Behavior & Organization* .

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

GEYER-KLINGEBERG, J., M. HANG, & A. W. RATHGEBER (2019): "What drives financial hedging? A meta-regression analysis of corporate hedging determinants." *International Review of Financial Analysis* **61(C)**: pp. 203–221.

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

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

Goeree, J. K. & C. a. Holt (2010): "Hierarchical package bidding: A paper & pencil combinatorial auction." *Games and Economic Behavior* **70(1)**: pp. 146–169.

Goeree, J. K. & T. Offerman (2003): "Competitive Bidding in Auctions with Private and Common Values." *Economic Journal* **113(489)**: pp. 598–613.

Goulder, L. H. & R. N. Stavins (2002): "Discounting: An eye on the future." *Nature* **419(6908)**: pp. 673–674.

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

Greiner, B. (2004): "An Online Recruitment System for Economic Experiments." *Forschung und wissenschaftliches Rechnen, GWDG Bericht 63* **63(13513)**: pp. 79–93.

Groenewegen, J. (1994): "About Double Organized Markets: Issues of Competition and Cooperation. The Dutch Construction Cartel: An Illustration." *Journal of Economic Issues* **28(3)**: pp. 901–908.

Guala, F. (2001): "Building Economic Machines : The FCC Auctions." *Studies in History and Philosophy of Science* **32**: pp. 1–34.

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

Hardisty, D. J., K. F. Thompson, D. H. Krantz, & E. U. Weber (2013): "How to Measure Time Preferences: An Experimental Comparison of Three Methods." *Judgment and Decision Making* **8(3)**: pp. 236–249.

Harris, C. (2012): "Feelings of Dread and Intertemporal Choice." *Journal of Behavioral Decision Making* **25(1)**: pp. 13–28.

Harrison, G. W., M. Igel Lau, E. E. Rutström, & M. B. Sullivan (2005): "Eliciting risk and time preferences using field experiments: Some methodological issues." In "Field experiments in economics," pp. 125–218. Emerald Group Publishing Limited.

Harrison, G. W., M. I. Lau, & E. E. Rutstrom (2010): "Individual Discount Rates and Smoking: Evidence from a Field Experiment in Denmark." *Journal of Health Economics* **29(5)**: pp. 708–717.

Harrison, G. W., M. I. Lau, & M. B. Williams (2002): "Estimating individual discount rates in Denmark: A field experiment." *American economic review* **92(5)**: pp. 1606–1617.

Hausman, J. (2001): "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left." *Journal of Economic Perspectives* **15(4)**: pp. 57–67.

Hausman, J. A. (1979): "Individual Discount Rates and the Purchase and Utilization of Energy-Using Durables." *Bell Journal of Economics* **10(1)**: pp. 33–54.

Havranek, T. (2010): "Rose effect and the euro: is the magic gone?" *Review of World Economics* **146(2)**: pp. 241–261.

Havranek, T. (2015): "Measuring Intertemporal Substitution: The Importance of Method Choices and Selective Reporting." *Journal of the European Economic Association* **13(6)**: pp. 1180–1204.

Havranek, T., D. Herman, & and Zuzana Irsova (2018a): "Does Daylight Saving Save Electricity? A Meta-Analysis." *The Energy Journal* **39(2)**: pp. 35–61.

Havranek, T., R. Horvath, Z. Irsova, & M. Rusnak (2015a): "Cross-country heterogeneity in intertemporal substitution." *Journal of International Economics* **96(1)**: pp. 100–118.

Havranek, T. & Z. Irsova (2010): "Meta-Analysis of Intra-Industry FDI Spillovers: Updated Evidence." *Czech Journal of Economics and Finance* **60(2)**: pp. 151–174.

Havranek, T. & Z. Irsova (2017): "Do Borders Really Slash Trade? A Meta-Analysis."

Havranek, T., Z. Irsova, K. Janda, & D. Zilberman (2015b): "Selective reporting and the social cost of carbon." *Energy Economics* **51**: pp. 394–406.

Havránek, T., Z. Irsova, L. Laslopova, & O. Zeynalova (2020): "Skilled and Unskilled Labor Are Less Substitutable than Commonly Thought." *IES Working Paper Series* **2020(29)**.

Havranek, T., Z. Irsova, & T. Vlach (2018b): "Measuring the Income Elasticity of Water Demand: The Importance of Publication and Endogeneity Biases." *Land Economics* **94(2)**: p. 32.

Havranek, T., Z. Irsova, & O. Zeynalova (2018c): "Tuition Fees and University Enrolment: A Meta-Regression Analysis." *Oxford Bulletin of Economics and Statistics* **80(6)**: pp. 1145–1184.

Havranek, T. & O. Kokes (2015): "Income elasticity of gasoline demand: A meta-analysis." *Energy Economics* **47(C)**: pp. 77–86.

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

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

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

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

Holt, C. A. (2005): "Markets, Games, and Strategic Behavior: Recipes for Interactive Learning."

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

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

HOSSAIN, T. & J. A. LIST (2012): "The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations." *Management Science* **58(12)**: pp. 2151–2167.

HU, A., T. OFFERMAN, & S. ONDERSTAL (2011): "Fighting collusion in auctions: an experimental investigation." *International Journal of Industrial Organization* **29(1)**: pp. 84–96.

HUNTER, J. E., T. BALLARD, J. HUNTER, H. JOHN EDWARD, F. L. SCHMIDT, & G. B. JACKSON (1982): *Meta-analysis: Cumulating research findings across studies*, volume 4. SAGE Publications, Incorporated.

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

IFCHER, J. & H. ZARGHAMEE (2011): "Happiness and Time Preference: The Effect of Positive Affect in a Random-Assignment Experiment." *American Economic Review* **101(7)**: pp. 3109–3129.

IMAI, T., T. A. RUTTER, & C. F. CAMERER (2021a): "Meta-Analysis of Present-Bias Estimation using Convex Time Budgets." *The Economic Journal* **131(636)**: pp. 1788–1814.

IMAI, T., K. ZEMLIANOVA, N. KOTECHA, & C. F. CAMERER (2021b): "How Common are False Positives in Laboratory Economics Experiments? Evidence from the P-Curve Method." *mimeo*, LMU Munich.

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

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

IRSOVA, Z. & T. HAVRANEK (2010): "Measuring Bank Efficiency: A Meta-Regression Analysis." *Prague Economic Papers* **2010(4)**: pp. 307–328.

IRSOVA, Z. & T. HAVRANEK (2013): "Determinants of Horizontal Spillovers from FDI: Evidence from a Large Meta-Analysis." *World Development* **42(C)**: pp. 1–15.

Irsova, Z., T. Havranek, L. Laslopova, & O. Zeynalova (2022): "Publication and Attenuation Biases in Measuring Skill Substitution." *The Review of Economics and Statistics* **(Forthcoming)**.

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

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

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

Johnson, M. W. & W. K. Bickel (2002): "Within-subject comparison of real and hypothetical money rewards in delay discounting." *Journal of the Experimental Analysis of Behavior* **77(2)**: pp. 129–146.

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

Kass, R. E. & A. E. Raftery (1995): "Bayes factors." *Journal of the American Statistical Association* **90(430)**: pp. 773–795.

Kirby, K. N. & N. N. Marakovic (1995): "Modeling Myopic Decisions: Evidence for Hyperbolic Delay-Discounting within Subjects and Amounts." *Organizational Behavior and Human Decision Processes* **64(1)**: pp. 22–30.

Kirby, K. N. & N. N. Marakovic (1996): "Delay-Discounting Probabilistic Rewards: Rates Decrease as Amounts Increase." *Psychonomic Bulletin and Review* **3(1)**: pp. 100–104.

Kirby, K. N., N. M. Petry, & W. K. Bickel (1999): "Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls." *Journal of Experimental Psychology: General* **128(1)**: pp. 78–87.

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

Klemperer, P. (2004): *Auctions: Theory and Practice*. Princeton: Princeton University Press.

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

Kovacs, K. F. & D. M. Larson (2008): "Identifying Individual Discount Rates and Valuing Public Open Space with Stated-Preference Models." *Land Economics* **84(2)**: pp. 209–224.

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

Krishna, V. (2009): *Auction Theory*. Academic Press.

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

Kuhn, M. A., P. Kuhn, & M. C. Villeval (2017): "Decision-environment effects on intertemporal financial choices: How relevant are resource-depletion models?" *Journal of Economic Behavior & Organization* **137(C)**: pp. 72–89.

Kuhnberger, A., M. Schulte-Mecklenbeck, & J. Perner (2002): "Framing Decisions: Hypothetical and Real." *Organizational Behavior and Human Decision Processes* **89(2)**: pp. 1162–1175.

Kwasnica, A. M. & K. Sherstyuk (2013): "Multiunit Auctions." *Journal of Economic Surveys* **27(3)**: pp. 461–490.

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

Laibson, D. (1997): "Golden Eggs and Hyperbolic Discounting." *The Quarterly Journal of Economics* **112(2)**: pp. 443–478.

Laury, S. K., M. M. McInnes, & J. Todd Swarthout (2012): "Avoiding the curves: Direct elicitation of time preferences." *Journal of Risk and Uncertainty* **44(3)**: pp. 181–217.

Lawrance, E. C. (1991): "Poverty and the Rate of Time Preference: Evidence from Panel Data." *Journal of Political Economy* **99(1)**: pp. 54–77.

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

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

LEY, E. & M. F. STEEL (2009): "On the effect of prior assumptions in Bayesian model averaging with applications to growth regression." *Journal of Applied Econometrics* **24(4)**: pp. 651–674.

LI TAO, L. H., L. ZHANG, & S. ROZELLE (2014): "Encouraging classroom peer interactions: Evidence from Chinese migrant schools." *Journal of Public Economics* **111**: pp. 29–45.

LIST, J. A. & A. S. SAMEK (2015): "The behavioralist as nutritionist: Leveraging behavioral economics to improve child food choice and consumption." *Journal of Health Economics* **39**: pp. 135–146.

LOCEY, M. L., B. A. JONES, & H. RACHLIN (2011): "Real and Hypothetical Rewards." *Judgment and Decision Making* **6(6)**: pp. 552–564.

LOEWENSTEIN, G. (1987): "Anticipation and the Valuation of Delayed Consumption." *The Economic Journal* **97(387)**: p. 666.

LOEWENSTEIN, G., D. READ, & R. F. BAUMEISTER (2003): *Time and decision : economic and psychological perspectives on intertemporal choice.* Russell Sage Foundation, New York: NY.

LOPOMO, G., R. C. MARSHALL, & L. M. MARX (2005): "Inefficiency of Collusion at English Auctions." *Contributions to Theoretical Economics* **5(1)**.

LUCKING-REILEY, D., D. BRYAN, N. PRASAD, & D. REEVES (2007): "Pennies from eBay: The determinants of price in online auctions." *Journal of Industrial Economics* **55(2)**: pp. 223–233.

MADIGAN, D., J. YORK, & D. ALLARD (1995): "Bayesian Graphical Models for Discrete Data." *International Statistical Review / Revue Internationale de Statistique* **63(2)**: p. 215.

MALESZA, M. (2019): "The Effects of Potentially Real and Hypothetical Rewards on Effort Discounting in a Student Sample." *Personality and Individual Differences* **151(C)**: p. 108807.

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

MARWELL, G. & R. E. AMES (1981): "Economists free ride, does anyone else?. Experiments on the provision of public goods, IV." *Journal of Public Economics* **15(3)**: pp. 295–310.

MATOUŠEK, J. (2014): *An Experimental Test of Design Alternatives for Spectrum Auctions with Communication Channels.* Diploma thesis, Charles University in Prague.

MATOUSEK, J., T. HAVRANEK, & Z. IRSOVA (2022): "Individual discount rates: a meta-analysis of experimental evidence." *Experimental Economics* **25(1)**: pp. 318–358.

MAZUR, J. E. (1984): "Tests of an equivalence rule for fixed and variable reinforcer delays." *Journal of Experimental Psychology: Animal Behavior Processes* **10(4)**: pp. 426–436.

MCCLURE, S. M., K. M. ERICSON, D. I. LAIBSON, G. LOEWENSTEIN, & J. D. COHEN (2007): "Time Discounting for Primary Rewards." *Journal of Neuroscience* **27(21)**: pp. 5796–5804.

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

MEIER, S. & C. SPRENGER (2010): "Present-biased preferences and credit card borrowing." *American Economic Journal: Applied Economics* **2(1)**: pp. 193–210.

MEIER, S. & C. D. SPRENGER (2013): "Discounting financial literacy: Time preferences and participation in financial education programs." *Journal of Economic Behavior & Organization* **95**: pp. 159–174.

MEIER, S. & C. D. SPRENGER (2015): "Temporal Stability of Time Preferences." *The Review of Economics and Statistics* **97(2)**: pp. 273–286.

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

MEYER, A. G. (2015): "The impacts of elicitation mechanism and reward size on estimated rates of time preference." *Journal of Behavioral and Experimental Economics* **58(C)**: pp. 132–148.

MIRALLES, A. (2010): "Self-enforced collusion through comparative cheap talk in simultaneous auctions with entry." *Economic Theory* **42(3)**: pp. 523–538.

MOCHON, A. & Y. SAEZ (2017): "A review of radio spectrum combinatorial clock auctions." *Telecommunications Policy* **41(5-6)**: pp. 303–324.

MORENO, S. G., A. J. SUTTON, A. E. ADES, T. D. STANLEY, K. R. ABRAMS, J. L. PETERS, & N. J. COOPER (2009): "Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study." *BMC medical research methodology* **9(1)**: pp. 1–17.

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

NANSEN MCCLOSKEY, D. & S. T. ZILIAK (2019): "What quantitative methods should we teach to graduate students? A comment on Swann's â€śIs precise econometrics an illusion?â€ť." *The Journal of Economic Education* **50(4)**: pp. 356–361.

NELSON, J. & J. MORAN (2020): "Effects of Alcohol Taxation on Prices: A Systematic Review and Meta-Analysis of Pass-Through Rates." *The B.E. Journal of Economic Analysis & Policy* **20(1)**: pp. 1–21.

NEWELL, R. G. & J. SIIKAMÄKI (2015): "Individual time preferences and energy efficiency." *American Economic Review* **105(5)**: pp. 196–200.

NOBEL, A., S. LIZIN, R. BROUWER, S. B. BRUNS, D. I. STERN, & R. MALINA (2020): "Are biodiversity losses valued differently when they are caused by human activities? A meta-analysis of the non-use valuation literature." *Environmental Research Letters* **15(7)**: pp. 1–23.

NOUSSAIR, C. N. & G. SERES (2020): "The effect of collusion on efficiency in experimental auctions." *Games and Economic Behavior* **119**: pp. 267–287.

OLIVOLA, C. Y. & S. W. WANG (2016): "Patience Auctions: The Impact of Time vs. Money Bidding on Elicited Discount Rates." *Experimental Economics* **19(4)**: pp. 864–885.

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

PERCOCO, M. & P. NIJKAMP (2009): "Estimating individual rates of discount: a meta-analysis." *Applied Economics Letters* **16(12)**: pp. 1235–1239.

PHELPS, E. S. & R. A. POLLAK (1968): "On Second-Best National Saving and Game-Equilibrium Growth." *The Review of Economic Studies* **35(2)**: p. 185.

PHILLIPS, O. R., D. J. MENKHAUS, & K. T. COATNEY (2003): "Collusive practices in repeated English auctions: Experimental evidence on bidding rings." *American Economic Review* **93(3)**: pp. 965–979.

RAFTERY, A. E., D. MADIGAN, & J. A. HOETING (1997): "Bayesian Model Averaging for Linear Regression Models." *Journal of the American Statistical Association* **92(437)**: pp. 179–191.

READ, D. & N. L. READ (2004): "Time Discounting Over the Lifespan." *Organizational Behavior and Human Decision Processes* **94(1)**: pp. 22–32.

ROBINSON, M. S. (1985): "Collusion and the Choice of Auction." *The RAND Journal of Economics* **16(1)**: pp. 141–145.

ROSENTHAL, R. & D. B. RUBIN (1988): "[Selection models and the file drawer problem]: comment: assumptions and procedures in the file drawer problem." *Statistical Science* **3(1)**: pp. 120–125.

ROTHSTEIN, H. R., A. J. SUTTON, & M. BORENSTEIN (2005): "Publication bias in meta-analysis." *Publication bias in meta-analysis: Prevention, assessment and adjustments* pp. 1–7.

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

Rusnak, M., T. Havranek, & R. Horvath (2013): "How to Solve the Price Puzzle? A Meta-Analysis." *Journal of Money, Credit and Banking* **45(1)**: pp. 37–70.

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)**: p. 450.

Sahakyan, N. & K. W. Stiegert (2012): "Corruption and Firm Performance." *Eastern European Economics* **50(6)**: pp. pp. 5–27.

Samuelson, L. (2005): "Economic Theory and Experimental Economics." *Journal of Economic Literature* **43(1)**: pp. 65–107.

Samuelson, P. (1937): "Note on Measurement of Utility." *Review of Economic Studies* **4(2)**: pp. 155–161.

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

Skrzypacz, A. & H. Hopenhayn (2004): "Tacit collusion in repeated auctions." *Journal of Economic Theory* **114(1)**: pp. 153–169.

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

Stanley, T. & H. Doucouliagos (2010): "Picture this: A simple Graph that Reveals Much Ado about Research." *Journal of Economic Surveys* **24(1)**: pp. 170–191.

Stanley, T. D. (2005): "Beyond Publication Bias." *Journal of Economic Surveys* **19(3)**: pp. 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)**: pp. 103–127.

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

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

STANLEY, T. D., S. B. JARRELL, & H. DOUCOULIAGOS (2010): "Could It Be Better to Discard 90% of the Data? A Statistical Paradox." *The American Statistician* **64**: pp. 70–77.

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)**: pp. 30–34.

STERN, B. B. & M. R. STAFFORD (2006): "Individual and social determinants of winning bids in online auctions." *Journal of Consumer Behaviour* **5(1)**: pp. 43–55.

SUDARSHAN, A. (2014): "Nudges in the marketplace: Using peer comparisons and incentives to reduce household electricity consumption." *Work. Pap., Energy Policy Inst. Chicago* .

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

SUTTER, M., M. G. KOCHER, D. GLÄTZLE-RÜTZLER, & S. T. TRAUTMANN (2013): "Impatience and Uncertainty: Experimental Decisions Predict Adolescents' Field Behavior." *American Economic Review* **103(1)**: pp. 510–531.

TAKAHASHI, H., J. SHEN, & K. OGAWA (2013): "An experimental examination of compensation schemes and level of effort in differentiated tasks." *Journal of Behavioral and Experimental Economics* **61**: pp. 12–19.

TAKEUCHI, K. (2011): "Non-parametric test of time consistency: Present bias and future bias." *Games and Economic Behavior* **71(2)**: pp. 456–478.

TANAKA, T., C. F. CAMERER, & Q. NGUYEN (2010): "Risk and time preferences: Linking experimental and household survey data from Vietnam." *American Economic Review* **100(1)**: pp. 557–571.

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

Thaler, R. (1981): "Some empirical evidence on dynamic inconsistency." *Economics Letters* **8(3)**: pp. 201–207.

Thaler, R. H. (2016): "Behavioral Economics: Past, Present, and Future." *American Economic Review* **106(7)**: pp. 1577–1600.

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

Tokunaga, M. & I. Iwasaki (2017): "The Determinants of Foreign Direct Investment in Transition Economies: A Meta-analysis." *The World Economy* **40(12)**: pp. 2771–2831.

Tol, R. S. J. (1999): "Time Discounting and Optimal Emission Reduction: An Application of FUND." *Climatic Change* **41(3-4)**: pp. 351–362.

Tsukayama, E. & A. L. Duckworth (2010): "Domain-specific temporal discounting and temptation." *Judgment and Decision Making* **5(2)**: pp. 72–82.

Tversky, A. & D. Kahneman (1974): "Judgment under Uncertainty: Heuristics and Biases." *Science* **185(4157)**: pp. 1124–1131.

Tversky, A. & D. Kahneman (1981): "The framing of decisions and the psychology of choice." *Science* **211(4481)**: pp. 453–458.

Ubfal, D. (2016): "How general are time preferences? Eliciting good-specific discount rates." *Journal of Development Economics* **118(C)**: pp. 150–170.

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

Ugur, M., S. Awaworyi Churchill, & E. Solomon (2018): "Technological innovation and employment in derived labour demand models: a hierarchical meta-regression analysis." *Journal of Economic Surveys* **32**: pp. 50–82.

Valíčkova, P., T. Havránek, R. Horváth, P. Valickova, T. Havranek, & R. Horvath (2015): "Financial Development And Economic Growth: A Meta-Analysis." *Journal of Economic Surveys* **29(3)**: pp. 506–526.

Valley, K. (1995): "Is Talk Really Cheap? Outperforming Equilibrium Models of Communication in Bargaining Games."

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

Voors, M. J., E. E. M. Nillesen, P. Verwimp, E. H. Bulte, R. Lensink, & D. P. V. Soest (2012): "Violent Conflict and Behavior: A Field Experiment in Burundi." *American Economic Review* **102(2)**: pp. 941–964.

Wang, M., M. O. Rieger, & T. Hens (2016): "How time preferences differ: Evidence from 53 countries." *Journal of Economic Psychology* **52**: pp. 115–135.

Warner, J. T. & S. Pleeter (2001): "The Personal Discount Rate: Evidence from Military Downsizing Programs." *American Economic Review* **91(1)**: pp. 33–53.

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

World Bank (2020): "Median Monthly Per Capita Expenditure (Or Income)." *World Bank's global database of household surveys (PovcalNet) March 2020 Update*, Global Poverty Monitoring, Washington, DC: World Bank.

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

Zauberman, G., B. K. Kim, S. A. Malkoc, & J. R. Bettman (2009): "Discounting Time and Time Discounting: Subjective Time Perception and Intertemporal Preferences." *Journal of Marketing Research* **46(4)**: pp. 543–556.

Zeugner, S. & M. Feldkircher (2015): "Bayesian Model Averaging Employing Fixed and Flexible Priors: The BMS Package for R." *Journal of Statistical Software* **68(4)**: pp. 1–37.

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

Zhou, X. & H. Zheng (2010): "Breaking bidder collusion in large-scale spectrum auctions." In "Proceedings of the eleventh ACM international symposium on Mobile ad hoc networking and computing," pp. 121–130. ACM.

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

Zigraiova, D., T. Havranek, Z. Irsova, & J. Novak (2021): "How puzzling is the forward premium puzzle? A meta-analysis." *European Economic Review* **134**: p. 103714.

Appendix A

Appendix to Chapter 2

A.1 Optimal Allocation

Table A.1.1 shows optimal allocation generating maximum possible surplus.

Table A.1.1: Optimal Efficiency Allocation

		Player type			
Goods		**I**	**II**	**III**	**IV**
A	Quantity	1	1	2	2
	Valuation	1886	1727	1900	1865
	Surplus	1886	1727	3800	3730
B	Quantity	15	6	2	1
	Valuation	53	53	47	48
	Surplus	795	318	94	48
C	Quantity	1	11	1	1
	Valuation	138	140	130	138
	Surplus	0.8	1	0	0.8
D	Quantity	1	1	3	3
	Valuation	48	47	46	50
	Surplus	48	47	138	150
	Total activity	27	28	26	25
	Activity disposed	27	28	26	25
	Surplus per player	3727.1	4721.6	5410.6	5285.8
	Optimal surplus		**19145.1**		

A.2 Additional Parameters of the Experiment

Table A.1.2 shows random draws for private value components and activity points assigned for each player. Table A.1.3 shows random draws for private signals on the common value component for each player.

Table A.1.2: Random Draws for PVC and Activity Points

Players	**PVC random draws**				**Activity**
	A	**B**	**C**	**D**	**random draw**
Type I	66	3	-2	-2	-1
Type II	-93	3	0	-3	0
Type III	80	-3	-10	-4	-2
Type IV	45	-2	-2	0	-3

Table A.1.3: Random Draws for Private Signals on CVC

Players	**PVC random draws**			
	A	**B**	**C**	**D**
Type I	1718	51	141	48
Type II	1859	50	145	46
Type III	1770	50	148	54
Type IV	1947	49	141	50

A.3 English Instructions

Hereby we present the experimental instructions to our paper *"Collusion in Multi-object Auctions: Experimental Evidence."* The original instructions were written in Czech language. We present an English translation; original version is available online at the author's webapge `http://ies.fsv.cuni.cz/cs/staff/matousek`.

The instructions were divided into three parts: (I) the introduction in Subsection A.3.1; (II) general instructions common to all treatments in Subsection A.3.2; and (III) a treatment specific supplement for respective treatments in Subsection A.3.3.

A.3.1 Introduction

Welcome to the Laboratory of Experimental Economics. My name is Jindřich Matoušek and my colleague's name is Lubomír Cingl. Thank you for participating in today's experiment.

Please, put all your belongings away so we can have your full attention.

In the course of the experiment, please do not talk to other participants and do not drink water. Please shut down your mobile phones. Violation of these rules will cause immediate exclusion from the experiment without any payment.

You cannot lose any money in this experiment. You will be given 100 CZK for coming on time. This 100 CZK and any money that you earn during the experiment will be paid to you, privately in cash, at the end of the experiment. The average expected payment in today's experiment is 500CZK and the average length of the experiment is 2 hours. The length of the experiment depends on the speed of participants, therefore please be patient.

All amounts in this experiment will be given in Experimental Currency Units (ECU). The exchange rate to Czech Crowns is one CZK for three ECUs.

You can make notes on the enclosed sheet of paper. With the control questions placed at the beginning, we only want to make sure you understand the experiment; you will not be excluded nor discriminated against in any manner because of them.

Please note that you commit yourself to participation in the whole experiment and if you leave before the end, you receive no payment at all. For your participation on the experiment, we need you to sign the consent form. Please take the consent form provided on a separate sheet of paper, read it and when you sign it, raise your hand and the experimenters will collect them. If you are not willing to participate and not sign the consent form, please leave the experiment now and your participation fee of 100 CZK will be paid to you.

If you have any question now or during the experiment, please raise your hand and we will answer it in private.

A.3.2 General Instructions

THE EXPERIMENT

The experiment will involve a series of auctions. Each auction will consist of multiple rounds.

In each auction, you will be competing with others for a set of multiple goods, which will contain various types and quantities. There are several important rules in this auction that encompasses (i) the way you can bid for the goods; (ii) provisional winners in each round of the auction; (iii) your eligibility; and (iv) the possibility of bid withdrawal. These rules are not trivial but crucial for your participation in the auction and also for your payoff from the experiment. Therefore, please, devote to them the utmost attention. Each auction will have an indefinite number of rounds, which depends on the decisions of its participants.

Let us explain the individual rules of the auction more closely.

INDIVIDUAL AUCTION ROUNDS

Each auction will consist of a series of a preliminary inexactly determined number of auction rounds. Each round has a time limit in which you have to submit your bid. After each round, the system will evaluate all submitted bids and show the round summary. This process repeats until the auction ends.

GROUPS AND BIDDERS

At the beginning of each auction, you will be randomly assigned to a group of four bidders (you and 3 others). Within these groups, you will be competing in all rounds of this auction. After this auction ends, you will be randomly assigned to a new group of four bidders.

GOODS FOR SALE

In each group of four players, four types of goods labeled A, B, C, and D, will be auctioned off. Each type of the goods is offered in multiple homogeneous units. You can submit bids for as many units of each type of the goods and for as many types as you want to. You will submit bids by adding the units in your bidding basket.

PRICES AND VALUES OF GOODS

Each type of goods A; B; C; a D offered for sale has a different upset price. The price of each type can increase gradually throughout the auction, in case an offer was made for this type in the previous auction round. Each increase will be implemented at a volume of 20% of the upset price of a respective type of the goods. If an offer was not made, the price of the goods remains the same. The price of goods within each type will be always the same for all units.

Each player will have different valuations for all types of goods. Your total personal value for each type of the goods will be known only to you.

The total valuation of each type of the goods consists of two components: the common value and the private value components of the goods. The total valuation is then the sum of these two components.

Each unit of the goods has its own common value, which is identical for all units of the goods. No bidder has the precise information about this common value. Each bidder receives only her private estimate of the common value determined by a random draw. The estimate of the common value is drawn separately for each player, but always out of the same interval.

Each player is further informed about her own private valuation of the goods, which she receives upon each unit bought in the auction. The private value is typically different for each player and is determined by a random draw from the interval in the range of ±10% of the common value component (which you do not know, but which is the same for all players); that is:

$$\mathbf{PVC} \in [-0.1 \cdot CVC; +0.1 \cdot CVC].$$

Even though your private component can be negative, your total value of each type of the goods is always positive.

The following table summarizes your knowledge of each type of the goods in the auction:

PVC	$\mathbf{PVC} \in [-0.1 \cdot CVC; +0.1 \cdot CVC]$.
CVC	Estimate
Total information	PVC + Estimate of CVC

EXAMPLE:

Private value component PVC	20
Estimate of CVC	300
Total signal	PVC + Estimate of CVC = 320

COMPLEMENTARITIES OF GOODS

All types of goods offered in the auction are complements. It means that a set of multiple goods containing more types (A; B; C; or D) has higher value than each type separately; thus the winning of more than one type of goods at once gives you the advantage of higher profit.

If you win one type of goods (in an arbitrary quantity), your profit is equal to the value of this goods. However, if you win more than one type of goods at once, your profit will rise according to following formula:

$$\textit{valuation} = [1 + 0.1 \cdot (X - 1)] \cdot \textit{ sum of valuations of goods won},$$

where X stands for the number of goods types acquired. Thus:

1 type - value is equal to the valuation of goods;

2 types - value raises by 10% of the valuation of goods;

3 types - value raises by 20% of the valuation of goods;

4 types - value raises by 30% of the valuation of goods.

EXAMPLE:

Total value of the goods of type A is 300, of type B is 100.

If a player wins goods A, her profit is 300.

If a player wins goods B, her profit is 100.

If a player wins goods A and B, her profit is $1.1 \cdot (300+100) = 1,1 \cdot 400 = 440$.

PROVISIONAL WINNERS

The system automatically process all submitted bids when the auction round is finished. In the auction round summary you will be informed if and for how many units of goods are you currently the provisional winner. A situation that more than one player submits the same bid can occur in the course of the auction round. If the sum of such bids in your group exceeds the number of goods sold in the auction, the system determines the winner of given units randomly, since the price is the same for all players. You therefore do not have to win a complete set of goods on which you have submitted your bid.

After the time limit runs out or when each player submits her bid a new auction round occurs.

You will win precisely such goods in the last auction round for which you are currently the provisional winner. Only the final offers out of the last auction round are used for the calculation of auction profits and therefore your real payoff out of the experiment.

THE RULE OF ELIGIBILITY

Each participant in the auction has a certain number of activity points at her disposal, which represents her eligibility to submit bids in the auction. The activity points determine the maximum number of goods on which a player is able to submit bids.

Each unit of the goods costs a certain number of activity points. Your total bid cannot exceed your current level of activity points.

The number of your activity points can decrease during the auction, since it depends on your behavior in previous auction rounds. In each round, you will gain the same number of activity points as you have used in the previous one. If you submit a bid in a given round with a total activity cost lower than your current level of activity at your disposal, your eligibility for subsequent rounds will diminish - your number of activity points will fall.

There is no way of acquiring the activity points back throughout the auction, nor to acquire more of them.

EXAMPLE:

You have 10 points of activity in a given auction round at your disposal. One unit of goods A costs 3 activity points, one unit of goods B costs 1 activity point.

If you submit a bid for 3 units of goods A and for 1 unit of goods B in a given round, you will pay 10 activity points in total $(3 \cdot 3 + 1 \cdot 1)$, by which you will use up your activity for this round. You will have 10 points of activity at your disposal in a subsequent round of the auction.

If you submit a bid for 2 units of goods A and for 2 units of goods B in a given round, you will pay 8 activity points in total $(2 \cdot 3 + 2 \cdot 1)$. You will have 8 points of activity at your disposal in a subsequent round of the auction.

WITHDRAWING WINNING BIDS

A situation could arise during the course of an auction, in which you win in some auction round only a subset of goods on which you have placed your bid. You can therefore win only a subset of goods for a price which exceeds the actual value of the goods.

If such a situation occurs, you have the possibility to withdraw your provisionally winning bid. Bid withdrawal is always available during the auction round summary. You can withdraw your bid for as many goods (both types and units), for which you are currently the provisional winner.

The possibility of bid withdrawal is limited in its volume. In particular, each bidder can use the right withdraw in at most two auction rounds, without any reference to the number of withdrawn goods in each particular round. However, the number of activity points for subsequent rounds will be appropriately decreased during each bid withdrawal by the sum of activity points for all respective withdrawn bids.

FINAL AUCTION ROUND

A final auction round arises when no participant submits an additional bid on any goods. Technically this situation means that all four participants in a group submit a bid for "empty bidding basket." The auction ends with this situation.

If you submit an empty bidding basket in some auction round during the course of the auction, your activity will fall to zero. You will not be able to participate in the auction any further. Submit, therefore, an empty bidding basket only in the situation when you wish to terminate your participation in the auction.

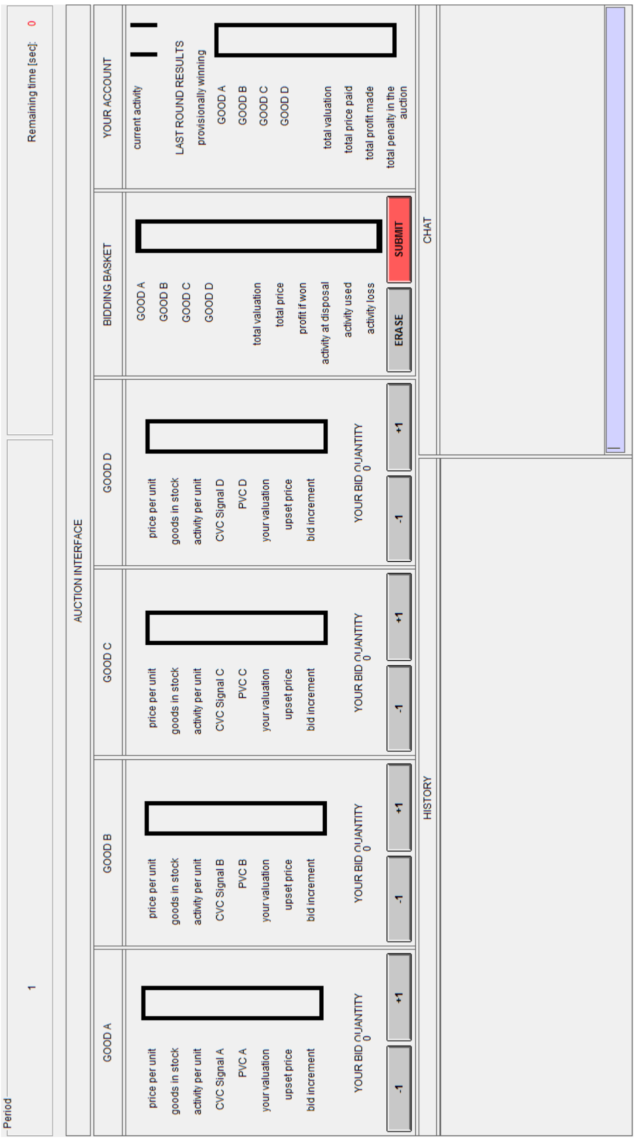


Figure A.1.1: The Auction Interface

HISTORY

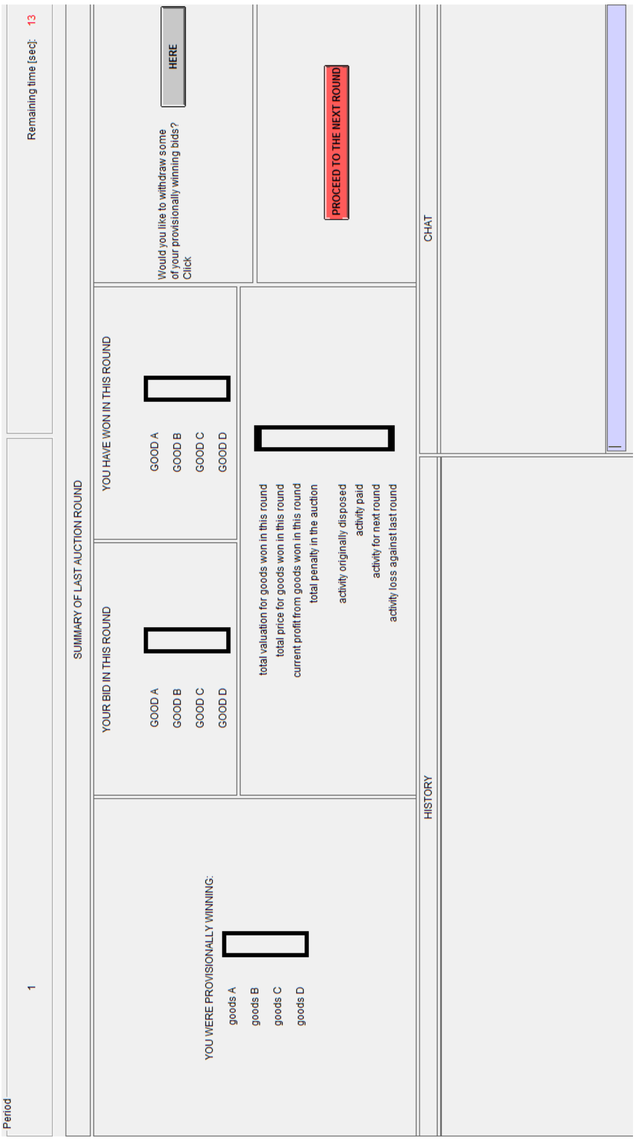
There is a history box present during the whole auction in the bottom left corner of the auction interface. It displays, for each player,the number of individual types of goods in this box for which this player was a provisional winner in a given auction round. The history, due to space constraints, is displayed with abbreviations (1-A; 1-B; 1-C; 1-D; 2-A; 2-B; etc.). The abbreviation "1-A" means "player 1 - goods of type A;" the abbreviation "2-B" means "player 2 - goods of type B" etc.

YOUR PROFIT AND EARNINGS FROM THE EXPERIMENT

At the end of the auction your earnings for this auction are determined. Your profit will be equal to the total value of the goods you won at the end of the auction (that is in the final round of the auction), minus the total cost you paid for them. Thus:

$$\textit{Profit} = \textit{total value of the goods won - price paid for all goods won}$$

Your final earnings will depend on one of the auctions held in this experiment. Which auction will be determined randomly at the end of the experiment.



Period 1

Remaining time [sec]: 13

SUMMARY OF LAST AUCTION ROUND

YOU WERE PROVISIONALLY WINNING:

goods A
goods B
goods C
goods D

YOUR BID IN THIS ROUND

GOOD A
GOOD B
GOOD C
GOOD D

YOU HAVE WON IN THIS ROUND

GOOD A
GOOD B
GOOD C
GOOD D

Would you like to withdraw some of your provisionally winning bids? Click HERE

total valuation for goods won in this round
total price for goods won in this round
current profit from goods won in this round
total penalty in the auction
activity originally disposed
activity paid
activity for next round
activity loss against last round

PROCEED TO THE NEXT ROUND

HISTORY

CHAT

Figure A.1.2: Summary of the Auction Round

SUMMARY

1. The experiment will consist of a series of auctions. The first auction is a trial and will not influence your payoff from the experiment. Each auction will consist of a series of a preliminary inexactly determined number of auction rounds. A final auction round arises when no participant submits an additional bid on any goods.

2. You will submit your bids by adding the units of goods in your bidding basket.

3. The price of a particular type of the goods can rise during the auction if there is positive demand for this type of goods. Your payoff from the experiment will depend on your ability to win the desired goods but also on the luck and abilities of others.

4. Provisional winning bids are announced after each auction round. However, these do not affect the final profit from the auction until they became final winning bids in the last round of the auction.

5. The rule of eligibility says, in principal, that you cannot wait to submit your bid until the end of the final rounds of the auction. If you want to win your desired portion of goods in the auction, you have to submit bids already from the beginning.

6. During the course of the auction, but not at its end, you will have the possibility to renounce your provisionally winning bid. This possibility will however be limited.

7. Individual valuations of goods are determined randomly for each player. It consists of a common and private value components of the goods, where the private component is known individually to all bidders. The common value component is, on the other hand, not known and the players have only a private signal about its value.

8. Your profit out of each auction will be determined only based on the situation from the final auction round and will be equal to the difference of the total value of goods you have won and the total price of your final bidding basket. Only one of the auctions held today will be chosen for your payoff at the end of the experiment.

A.3.3 Treatment-Specific Supplements

The treatment-specific supplements to the instructions were presented to the participants in the following sequence. There were three basic parts of the supplement: (i) notice; (ii) communication window; and (iii) the set of goods as a package. The following table summarizes which parts were presented in which treatment. There was a simple one-sentence introduction "treatment specific supplement introduction" present at the beginning in all treatments.

	Notice	Communication window	Set of goods as a package
SMR Basic	✓	✗	✗
SMR Comm	✓	✓	✗
SMRPB Basic	✓	✗	✓
SMRPB Comm	✓	✓	✓

TREATMENT-SPECIFIC SUPPLEMENT INTRODUCTION

Hereby presented additional rules were not stated in the online questionnaire.

NOTICE

1. Price of the goods gradually rises throughout the auction (in case the offers are made for this type of goods). If you are not able to find an optimal situation with positive profit in any round and you will incur a loss, it is highly improbable that you would find such a situation in subsequent rounds.

2. If you incur a loss out of the auction used for the calculation of your payoff from the experiment, it will appear in that payoff. Potential loss will be adequately subtracted from your payment for timely arrival. We therefore strongly recommend not submitting bids that can incur losses.

3. If you submit a bid in any round, your bidding basket will reset and assigns the goods freshly again according to your new offer. It is not possible to add some goods in your existing bidding basket. You always have to submit an offer for a complete set of desired goods.

4. Your task in the experiment is to gain a positive profit at the end of each auction, not to maintain your full level of activity points.

COMMUNICATION WINDOW

There will be a communication window present in the bottom right corner of the auction interface. You can send any messages to other participants in your group through this window. Such messages will be visible only to the players in your own group. The communication window will also be displayed for two minutes before each auction.

SET OF GOODS AS A PACKAGE

You will be bidding for a set of goods of your preference in each round of today's auctions. The system will handle this set as one compact package. Your bid will be either accepted as a package or refused as a package; you will therefore win the complete set you were bidding for or nothing.

At the end of each auction round, the system processes all bid packages submitted in the current auction round and displays information about the provisionally winning bids of this round. The processing runs based on the package with highest price. Even the players who did not submit an offer with the highest price, but whose offer was, after the processing stage and determination of other provisionally winning bids, still available from the perspective of the quantity, can become the provisional winners of their packages.

Appendix B

Appendix to Chapter 3

B.1 Robustness Checks to Tests of Publication Bias

Table B.1.1: Funnel asymmetry tests with standard errors clustered at the level of authors

	OLS	Fixed effects	Instrument	Precision
Standard error	0.535***	0.875***	0.316	1.031**
(*publication bias*)	(0.0331)	(0.0146)	(0.194)	(0.455)
Constant	0.518***	0.341***	0.633***	0.259***
(*effect beyond bias*)	(0.125)	(0.00762)	(0.180)	(0.0391)
Observations	927	927	927	927
Clusters	31	31	31	31

Notes: The table reports the results of regression $\delta_{ij} = \delta_1 + \gamma_1 \cdot SE(\delta_{ij}) + u_{ij}$, where δ_{ij} denotes the i-th annualized discount rate estimated in the j-th study, and $SE(\delta_{ij})$ denotes its standard error. The table shows estimation by OLS, study-level fixed effects, instrumental variables (where the instrument for the standard error is the inverse of the square root of the number of observations in a study), and precision weighting (where estimates are weighted by the inverse of their standard error). Standard errors, clustered at the level of authors, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.1.2: Funnel asymmetry tests for medians of individual-specific discounting

	OLS	Fixed effects	Instrument	Precision
Standard error	0.535^{***}	0.875^{***}	0.535^{***}	1.012^{**}
(*publication bias*)	(0.0282)	(0.0154)	(0.0282)	(0.453)
Standard error * Median	0.373	-1.093^{***}	0.373	0.417
(*additional bias in median estimates*)	(0.259)	(0.0518)	(0.259)	(0.619)
Constant	0.509^{***}	0.369^{***}	0.509^{***}	0.258^{***}
(*effect beyond bias*)	(0.118)	(0.00817)	(0.118)	(0.0376)
Observations	927	927	927	927

Notes: The table reports the results of regression $\delta_{ij} = \delta_1 + \gamma_1 \cdot SE(\delta_{ij}) + \gamma_2 \cdot SE(\delta_{ij}) \cdot Median_{ij} + u_{ij}$, where δ_{ij} denotes the i-th annualized discount rate estimated in the j-th study, $SE(\delta_{ij})$ denotes its standard error, and $Median$ is a dummy variable that equals 1 if the estimate of the discount rate is a median of individual-specific discounting. The table shows estimation by OLS, study-level fixed effects, instrumental variable (where the instrument for the standard error is the inverse of the square root of the number of observations in a study), and precision weighting (where estimates are weighted by the inverse of their standard error). Standard errors, clustered at the study level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.1.3: Excluding estimates with unidentified discounting type

PANEL A: Linear models				
	OLS	Fixed effects	Instrument	Precision
Standard error	1.112^{***}	0.852^{**}	-0.233	2.814^{***}
(*publication bias*)	(0.210)	(0.359)	(1.598)	(0.684)
Constant	0.384^{***}	0.414^{***}	0.535^{***}	0.194^{***}
(*effect beyond bias*)	(0.0745)	(0.0403)	(0.192)	(0.0302)
Observations	507	507	507	507
PANEL B: Non-linear models				
	WAAP of Ioannidis *et al.* (2017)	Stem-based method of Furukawa (2021)	Selection model of Andrews & Kasy (2019)	Endogenous kink of Bom & Rachinger (2019)
Effect beyond bias	0.305^{***}	0.067^{*}	0.218^{***}	0.145^{***}
	(0.016)	(0.040)	(0.130)	(0.004)
Observations	507	507	507	507

Notes: The table reports the results of regression $\delta_{ij} = \delta_1 + \gamma_1 \cdot SE(\delta_{ij}) + u_{ij}$, where δ_{ij} denotes the i-th annualized discount rate estimated in the j-th study, and $SE(\delta_{ij})$ denotes its standard error. Estimates for which the discounting model is not explicitly stated are omitted from estimations in this table. Panel A shows estimation by OLS, study-level fixed effects, instrumental variables (where the instrument for the standard error is the inverse of the square root of the number of observations in a study), and precision weighting (where estimates are weighted by the inverse of their standard error). Panel B shows the recently developed non-linear estimation techniques; WAAP stands for the Weighted Average of the Adequately Powered estimates. Standard errors, clustered at the study level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.1.4: Funnel asymmetry tests in absolute value

	OLS	Fixed effects	Instrument	Precision
Standard error	0.534^{***}	0.872^{***}	0.534^{***}	1.040^{**}
(*bias in positive estimates*)	(0.0304)	(0.0158)	(0.0304)	(0.456)
Standard error * Negative	-2.104^{***}	-0.610	-2.104^{***}	-2.306^{***}
(*bias in negative estimates*)	(0.371)	(0.730)	(0.371)	(0.743)
Constant	0.523^{***}	0.344^{***}	0.523^{***}	0.260^{***}
(*effect beyond bias*)	(0.114)	(0.00899)	(0.114)	(0.0374)
Observations	927	927	927	927

Notes: The table reports the results of regression $|\delta_{ij}| = \delta_1 + \gamma_1 \cdot SE(\delta_{ij}) + \gamma_2 \cdot SE(\delta_{ij}) \cdot Negative_{ij} + u_{ij}$, where δ_{ij} denotes the i-th annualized discount rate estimated in the j-th study, $SE(\delta_{ij})$ denotes its standard error, and $Negative$ is a dummy variable that equals 1 if the estimate of the discount rate is negative. The table shows estimation by OLS, study-level fixed effects, instrumental variables (where the instrument for the standard error is the inverse of the square root of the number of observations in a study), and precision weighting (where estimates are weighted by the inverse of their standard error). Standard errors, clustered at the study level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

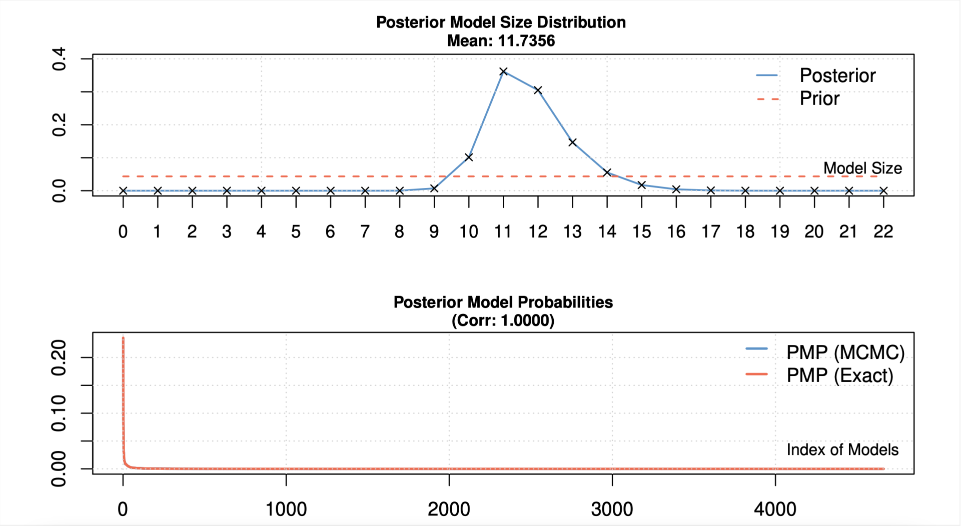
B.2 Robustness Checks and Additional Statistics to BMA

Table B.2.1: Summary of the benchmark BMA estimation

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
11.7356	$2 \cdot 10^6$	$1 \cdot 10^6$	2.350162 mins	402,090
Modelspace	*Models visited*	*Topmodels*	*Corr PMP*	*No. obs.*
$4.19 \cdot 10^6$	9.60%	100%	1.0000	927
Model prior	*g-prior*	*Shrinkage-stats*		
Random/11	UIP	Av = 0.9989		

Notes: We employ the priors recommended by Eicher *et al.* (2011) and George (2010), the unit information prior (the prior has the same weight as one observation in the data) and the dilution prior (accounting for potential collinearity). The results of this BMA exercise are reported in Table 3.6

Figure B.2.1: Model size and convergence for the benchmark BMA model



Notes: The figure depicts the posterior model size distribution and the posterior model probabilities of the BMA exercise reported in Table 3.6.

Table B.2.2: Alternative BMA priors and frequentist model averaging

	Bayesian model averaging (BRIC)			Frequentist model averaging		
Variable:	Post. mean	Post. SD	PIP	Mean	SE	p-value
Constant	-0.244	NA	1.000	-0.393	0.140	0.005
Standard error	0.549	0.021	1.000	0.572	0.024	0.000
Estimation characteristics						
Hyperbolic discounting	0.039	0.061	0.351	0.132	0.062	0.035
Exponential discounting	0.006	0.029	0.074	0.089	0.075	0.235
Delay	0.000	0.002	0.040	-0.002	0.009	0.843
Front-end delay	0.013	0.041	0.141	0.109	0.064	0.089
Lab experiment	0.156	0.101	0.777	0.124	0.075	0.100
Experimental characteristics						
Real reward	-0.005	0.026	0.075	-0.031	0.067	0.648
Matching task	0.017	0.045	0.160	0.017	0.066	0.791
Health domain	0.346	0.088	0.993	0.317	0.091	0.001
Other domain	0.441	0.069	1.000	0.424	0.072	0.000
Negative framing	-0.148	0.106	0.735	-0.139	0.077	0.073
Neutral framing	0.003	0.030	0.045	0.017	0.089	0.851
Stakes						
Subject pool characteristics						
Sample size	0.075	0.014	1.000	0.084	0.017	0.000
Students	0.877	0.111	1.000	0.825	0.132	0.000
Students * Lab experiment	-0.753	0.144	1.000	-0.670	0.160	0.000
Males only	0.013	0.052	0.089	0.098	0.110	0.374
Females only	-0.001	0.022	0.040	0.000	0.012	1.000
North America	0.012	0.040	0.125	0.113	0.066	0.085
Asia	0.385	0.103	0.991	0.384	0.095	0.000
Africa	3.170	0.118	1.000	3.295	0.137	0.000
Publication characteristics						
Citations	-0.003	0.011	0.094	-0.014	0.022	0.527
Publication year	0.121	0.026	1.000	0.104	0.029	0.000
Observations	927			927		
Studies	56			56		

Notes: Response variable = annualized estimates of the discount rate. SD = standard deviation, PIP = Posterior inclusion probability, SE = standard error. The first specification from the left uses Bayesian model averaging with an alternative model prior, the beta-binomial prior advocated by Ley & Steel (2009) and Zellnerâ€™s g prior BRIC according to Fernandez *et al.* (2001). The second specification, frequentist model averaging, applies Mallow's model averaging estimator (Hansen 2007) using the orthogonalization of covariate space suggested by Amini & Parmeter (2012) to reduce the number of estimated models. All variables are described in Table 3.5.

Table B.2.3: Alternative specifications of the baseline BMA model

	Bayesian model averaging (without SE)			Bayesian model averaging (with stakes)			Bayesian model averaging (known model)		
Variable:	P. mean	P. SD	PIP	P. mean	P. SD	PIP	P. mean	P. SD	PIP
Constant	0.790	NA	1.000	0.180	NA	1.000	-0.704	NA	1.000
Standard error				0.567	0.023	1.000	0.856	0.110	1.000
Estimation characteristics									
Hyperbolic discounting	-0.383	0.068	1.000	0.001	0.013	0.039			
Exponential discounting	-0.505	0.084	1.000	0.004	0.023	0.055	0.000	0.012	0.043
Delay	-0.010	0.018	0.306	0.004	0.011	0.122	-0.098	0.017	1.000
Front-end delay	-0.403	0.062	1.000	0.050	0.069	0.398	0.185	0.111	0.808
Lab experiment	0.278	0.148	0.855	0.097	0.122	0.445	0.311	0.073	0.997
Experimental characteristics									
Real reward	0.166	0.140	0.664	-0.041	0.077	0.267	-0.001	0.020	0.051
Matching task	0.335	0.117	0.972	0.007	0.032	0.071	0.002	0.024	0.056
Health domain	0.110	0.144	0.442	0.979	0.173	1.000	0.382	0.095	0.996
Other domain	0.031	0.075	0.201	0.646	0.097	1.000	0.420	0.083	1.000
Negative framing	-0.409	0.092	0.999	-0.033	0.074	0.201	-0.030	0.075	0.179
Neutral framing	0.017	0.076	0.101	0.010	0.059	0.049	0.002	0.035	0.044
Stakes				-0.478	0.094	1.000			
Subject pool characteristics									
Sample size	-0.050	0.027	0.856	0.120	0.018	1.000	0.142	0.029	1.000
Students	0.933	0.193	1.000	0.398	0.296	0.755	-0.007	0.043	0.075
Students * Lab experiment	-0.684	0.254	0.960	-0.395	0.339	0.643	-0.001	0.044	0.066
Males only	0.005	0.042	0.071	0.016	0.061	0.092	0.015	0.063	0.085
Females only	-0.006	0.043	0.073	0.000	0.020	0.029	0.004	0.034	0.050
North America	0.005	0.030	0.093	-0.002	0.018	0.049	0.146	0.111	0.704
Asia	0.306	0.175	0.835	0.073	0.146	0.244	0.351	0.108	0.975
Africa	2.570	0.155	1.000	3.242	0.134	1.000			
Publication characteristics									
Citations	0.003	0.012	0.100	-0.041	0.045	0.511	-0.001	0.009	0.059
Publication year	0.374	0.038	1.000	0.017	0.036	0.232	0.013	0.034	0.173
Observations	927			777			507		
Studies	56			51			32		

Notes: Response variable = annualized estimates of the individual discount rate. P. mean = posterior mean, P. SD = posterior standard deviation, PIP = posterior inclusion probability. We employ Bayesian model averaging (BMA) using unit information prior (Eicher *et al.* 2011) and the dilution prior suggested by George (2010) which accounts for collinearity. In the first specification from the left, we exclude the variable *Standard error*; in the second specification we introduce variable *Stakes* into the model (which reduces the number of observations to 777); in the third specification we use only those observations where the type of discounting can be explicitly identified. All variables are described in Table 3.5.

Table B.2.4: BMA specifications accounting for non-linearity and exact delay

	Bayesian model averaging (money * non-linearity)			Bayesian model averaging (exact delay)		
Variable:	Post. mean	Post. SD	PIP	Post. mean	Post. SD	PIP
Constant	-0.242	NA	1.000	-0.748	NA	1.000
Standard error	0.549	0.021	1.000	0.611	0.027	1.000
Estimation characteristics						
Hyperbolic discounting	0.039	0.068	0.326			
Exponential discounting	0.005	0.028	0.068	0.057	0.094	0.334
Delay	0.000	0.002	0.037	0.002	0.008	0.134
Front-end delay	0.013	0.040	0.132	0.173	0.115	0.775
Lab experiment	0.153	0.102	0.766			
Experimental characteristics						
Real reward	-0.005	0.026	0.072	-0.004	0.053	0.114
Matching task	0.016	0.045	0.152	0.158	0.113	0.747
Health domain	0.346	0.089	0.992	0.304	0.147	0.889
Other domain	0.442	0.071	1.000	0.658	0.100	1.000
Money domain * non-linearity correction	0.001	0.035	0.090			
Negative framing	-0.146	0.107	0.724	-0.007	0.034	0.098
Neutral framing	0.003	0.030	0.042	0.053	0.152	0.165
Subject pool characteristics						
Sample size	0.075	0.014	1.000	0.166	0.032	1.000
Students	0.877	0.111	1.000	1.184	0.178	1.000
Students * Lab experiment	-0.752	0.144	1.000	-0.992	0.142	1.000
Males only	0.012	0.050	0.082	0.081	0.152	0.284
Females only	-0.001	0.021	0.037	0.011	0.062	0.086
North America	0.011	0.039	0.116	0.037	0.076	0.260
Asia	0.382	0.104	0.989	0.288	0.169	0.831
Africa	3.169	0.117	1.000	3.146	0.200	1.000
Publication characteristics						
Citations	-0.003	0.011	0.089	0.000	0.013	0.080
Publication year	0.121	0.027	1.000	0.151	0.041	0.994
Observations	927			568		
Studies	56			28		

Notes: Response variable = annualized estimates of the discount rate. SD = standard deviation, PIP = Posterior inclusion probability, SE = standard error. We employ Bayesian model averaging (BMA) using unit information prior (Eicher *et al.* 2011) and the dilution prior suggested by George (2010), which accounts for collinearity. In the first specification we include variable *Money domain * non-linearity correction*, interaction of *Money domain* with a correction for non-linearity of utility functions; in the second specification we estimate a model on a subsample of estimates for which the exact time horizon is coded (which reduces the number of observations to 568 and eliminates variables *Hyperbolic discounting* and *Lab experiment* due to high collinearity). All variables are described in Table 3.5.

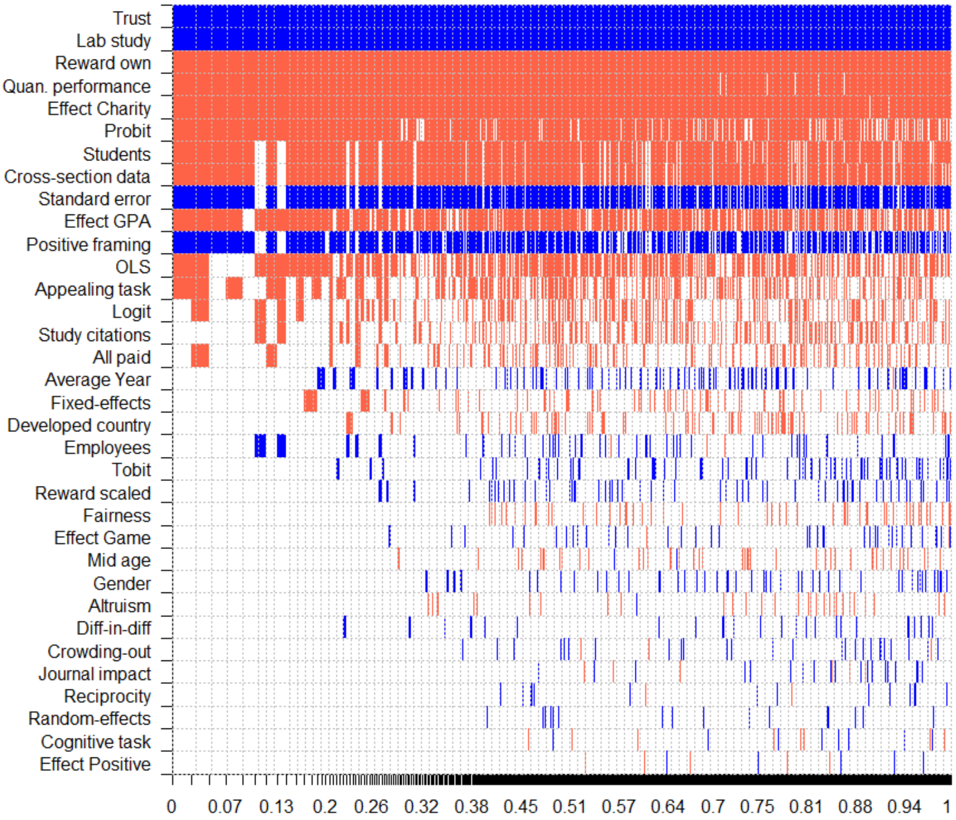
Appendix C

Appendix to Chapter 4

C.1 BMA robustness checks

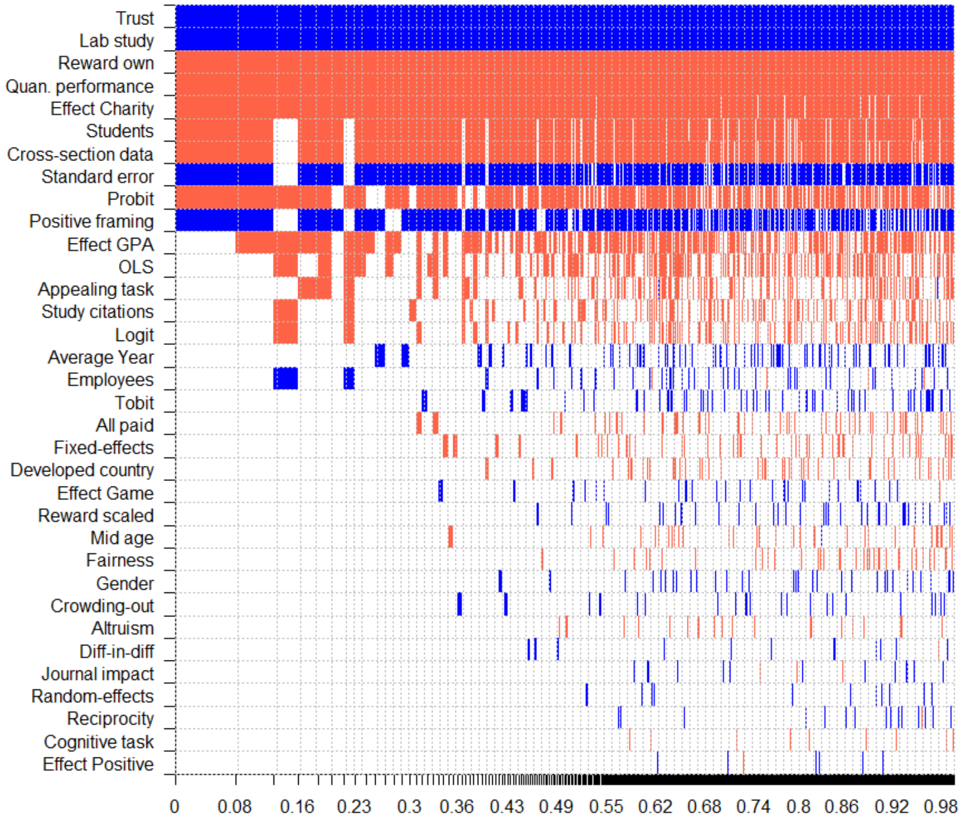
Here we present three additional Bayesian model averaging procedures using different specifications to the one discussed in Section 4.5. We chose these specifications to best capture the potential differences in their approach and present them in the note under each of the three models. With Feldkircher & Zeugner (2009) providing further detail on the theory behind each of the parameters in our setups, our primary goal in these checks is to control for the large number of observations included.

Figure C.1.1: BMA using uniform g-prior and uniform model prior



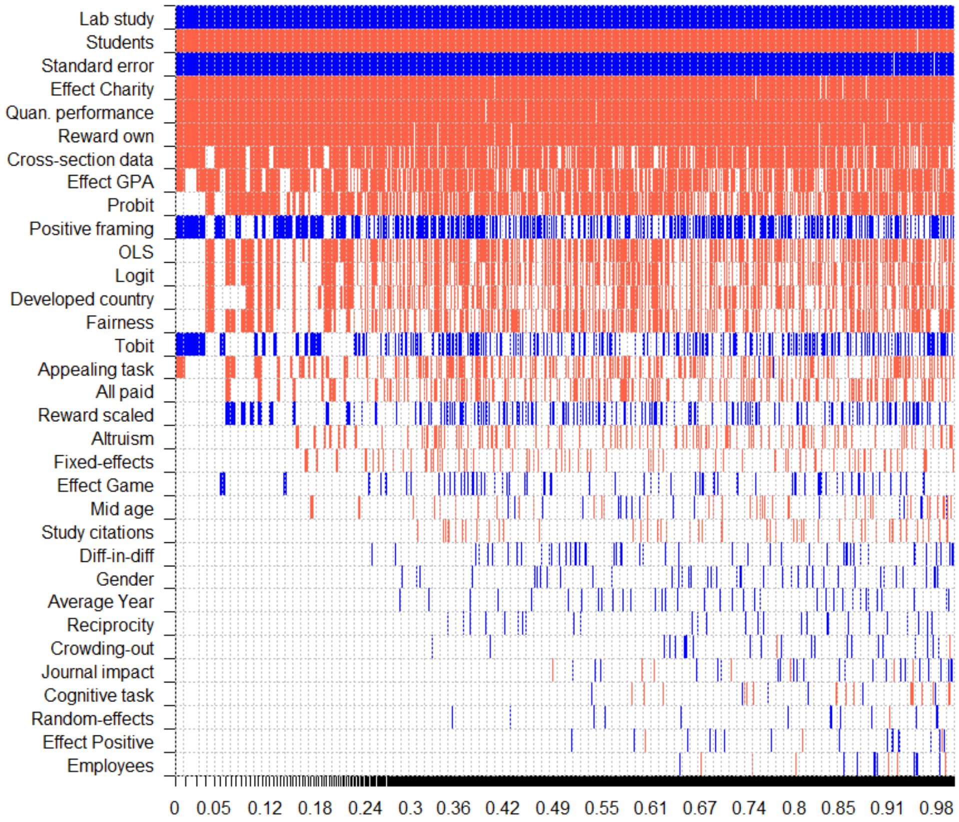
Note: This figure displays the results of the Bayesian model averaging using the uniform g-prior and the uniform model prior setup. BMA = Bayesian model averaging. For further explanation of the procedure and individual variables, see Figure 4.6 and Table 4.7.

Figure C.1.2: BMA using benchmark g-prior and random model prior



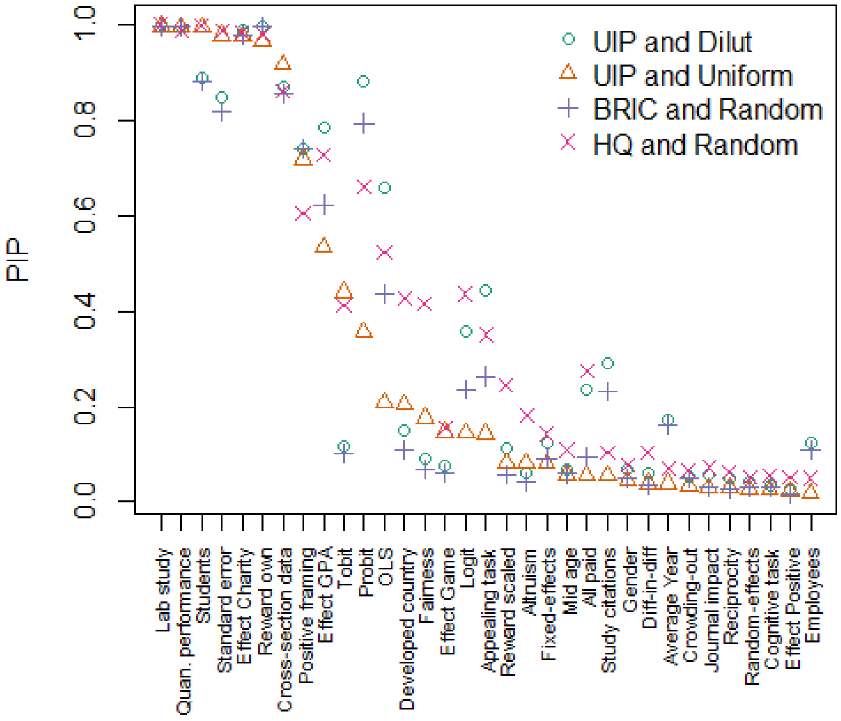
Note: This figure displays the results of the Bayesian model averaging using the benchmark g-prior and the uniform model prior setup. BMA = Bayesian model averaging. For further explanation of the procedure and individual variables, see Figure 4.6 and Table 4.7.

Figure C.1.3: 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 4.6 and Table 4.7.

Figure C.1.4: 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 4.7.

Appendix D

Responses to Referees

First of all, I would like to thank my supervisor prof. PhDr. Tomáš Havránek, Ph.D.and all three referees for the time they devoted to reading my dissertation and for their valuable comments and suggestions that helped me improve the quality of the thesis. I present the responses to referee comments in the following sections. The comments are typeset in italic. My reactions are below a line under each comment in roman.

Since Chapters 2 and 3 have already been published, I do not change the main text of the papers. Fortunately, I managed to answer all the comments on these chapters through this section.

Most of the comments focused on Chapter 4 since that one has not been sent to the review process in any journal. I would like to thank the referees for focusing on this unpublished part of my thesis since their suggestions improve our chances for a well-placed publication of our work substantially. After the discussions with my supervisor, I made several additional changes on top of the ones required by the referees. These changes are presented in the next Section D.1.

D.1 Additional changes to Chapter 4

- The most noticeable change to the Chapter 4 is a change in its title. I made this change since the original title did not accurately express what we are discussing in the paper. First, the incentives in primary studies are not only of a financial nature, the study is about incentives more generally. Second, we study the effect of incentives on motivation and performance rather than work itself. I believe the new title expresses the nature of the study better.

- I have added the statistics together with its p-value for p-uniform Publication bias test in Table 4.5.

- In Section 4.4 I deleted the Hierarchical Bayes method of publication bias estimation. I deleted also its results from Table 4.4. The reason for this is that it is not clear to me if this method fulfils the assumptions. I, therefore, prefer to put more emphasis on techniques that are more econometrically justifiable. Providing other methods for non-linear estimation of publication bias, I decided to remove the method from the analysis.

- I added the reporting guidelines according to Doucouliagos (2011) in Section 4.3:

  Doucouliagos (2011) collects 22,141 estimates of Partial correlation coefficients in the field of economics and introduces guidelines for its reporting. He provide three bands of PCC sizes: 1) small with PCC below ±0.07*, 2) Medium with PCC* $\in [\pm 0.07; \pm 0.32]$*, and 3) Large with PCC above* ±0.33*. The baseline effect in our case shows a mean Partial correlation coefficient of 0.046, which suggests a small incentives-motivation effect according to these guidelines. Looking systematically at the variables that have the highest impact on mean statistics, we find a medium effect in some cases but never the Large one.*

- I explained the transformation of the PCC with effect positive variable:

  We remedy this problem by using a dummy variable equal to one if this relationship is positive and transform the PCC of estimates with a negative relationship to negative numbers by multiplying it by (-1). This approach allows us to unify the effect's direction, meaning that an increase in the effect size always indicates better performance/outcome and consequently becomes straightforward to compare.

- While I was introducing the best-practice estimates for different contexts, I found a minor mistake in summary statistics loaded to the best-practice estimate. I corrected the mistake and re-estimated the results of the best-practice estimate together with the economic significance of key variables. The calculation is correct now and the changes do not change the qualitative implications of the analysis.

D.2 prof. PhDr. Tomáš Havránek, Ph.D.

D.2.1 Chapter 4 - Motivation

comment 1

Most people will find very surprising the result that financial incentives do not work. I think your BMA analysis makes it clear that the incentives work more under some conditions, less under others. Perhaps in your discussion (in the intro and elsewhere) you could focus a bit more on these nuances. So, for example, best-practice estimates could be provided for different contexts.

Yes, you are right and thank you for pointing it out. I have adjusted the discussion and tone of the text throughout the paper to be more in line with our results from the BMA analysis. Moreover, in Subsection 4.5.4 I now provide best-practice estimates for seven additional scenarios:

" Moreover, from Model averaging results, we identify several scenarios that have interesting results. Those scenarios are represented by the response variables in BMA that have PIP higher than 0.5 and are at the same time of some experimentally conceptual nature. We identify 7 such scenarios: the effect being captured by student's performance (Effect GPA), the effect being captured by charitable giving (Effect charity), the measured performance was quantitative (Quantitative performance), when the study rewards its subjects instead of punishing them (Positive framing), if the subjects received the reward for themselves (Reward own), if the experiment took place in a laboratory (Lab study), when the experimental subjects were students (Students). We provide the best-practice effect of rewards on motivation within these contexts by consecutively setting the coefficients of respective variables equal to one in the basic best-practice estimate equation. "

comment 2

Also, perhaps you could acknowledge that the average result (no effect of incentives) is not really that plausible - again, most people would argue that, logically (and from our personal experience), financial incentives matter. Perhaps the empirical studies cannot identify the effect, due to attenuation bias or other problems (the "iron law of econometrics"). At least it's possible.

Truly, the result is not fully plausible. But as you stated in a previous comment, the effect depends on the conditions. And as I hopefully present more clearly now in the BMA results section of the paper (and elsewhere) the incentives simply do not seem to work under some of those conditions, or perhaps do not work much.

Regarding the biases you mention, it is possible and I would say that also very likely to some extent, but not in full. Both types of biases drag the result down to zero. The former is connected to the independent variables while the latter is to the dependent variable in the regression. The attenuation bias is caused by measurement error or noise in independent variables and drags the model coefficients and therefore the overall result towards zero. It includes estimation technique differences such as the difference between estimates obtained through Instrumental variables and OLS estimations. The latter, Hausman (2001) refers to the "Iron law of econometrics" as to the effect of a mismeasured variable when the regression estimate is downward biased in magnitude towards zero—the magnitude of the estimate is usually smaller than expected.

In our setting, this would mean that the effect of rewards on the motivation of subjects is generally even more overestimated. In your forthcoming paper for REStat (Irsova *et al.* 2022), for the detection of the attenuation bias, you exploit the fact that part of the literature uses instrumental variables (IV) to address the bias (and other endogeneity biases), while other studies either use simple OLS or provide natural experiments with exogenous variation in relative labour supply. Unfortunately, we do not have IV estimates in our data sample and therefore cannot use this method of correction for attenuation bias. In our case, we could compare means of PCC for the group of data using OLS estimation (924 observations) with a group that utilizes simple means (215 observations) to control for other endogeneity biases. The reported means of the PCC for these two subgroups are 0.046 and 0.050, respectively. This shows that some bias truly can be present roughly in the magnitude of 13%. Translating this value to our correction implies that the mean PCC corrected for both publication and endogeneity bias could be around 0.026. However, I deal no further with this topic here mainly due to feasibility reasons at this stage of the dissertation, even though I do not rule out the possibility that we will elaborate on it in the next versions of the paper.

comment 3

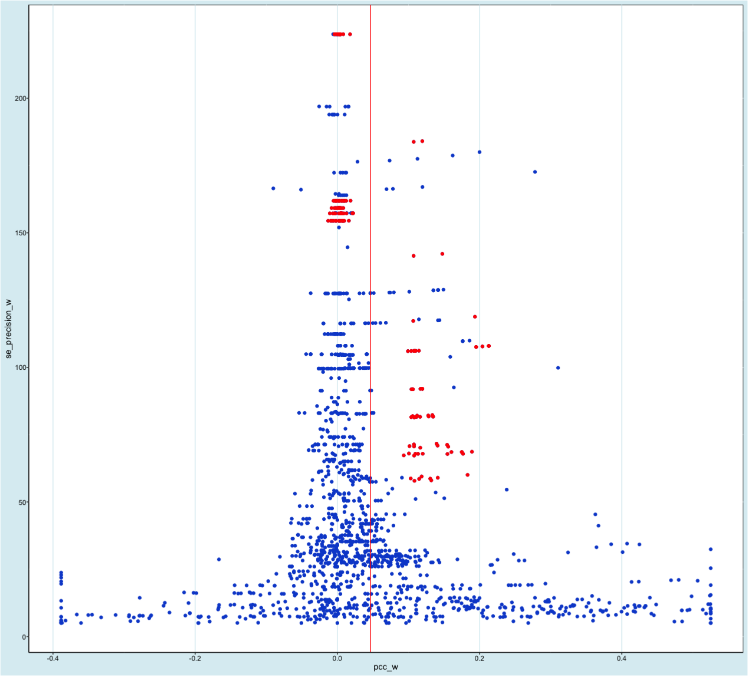
The funnel plot is a bit weird, as there are several outlying studies at the right-hand side, perhaps evaluating a slightly different concept? Otherwise the funnel is exactly as expected due to theory and seems to show relatively little publication bias.

Yes, you are right, the funnel plot is strange at first glance. I must admit that we tried to explain these data points several times but we did not manage to discover any clear pattern that would explain the relationship. The most obvious one can be found when we filter out observations from the top right corner (setting arbitrarily PCC > 0.1 & precision > 50). We get 97 observations in total, out of which 70% are from the study by (Karlan & List 2007) who conduct a natural field experiment using direct mail solicitations from 50,000 donors. The study is therefore characterized by an exceptionally high degree of precision. When we plot the funnel in Figure C.1.1 with data points from this study marked in red, we can see, however, that the greater "outlier" section is not populated by Karlan & List (2007) exclusively. Other studies include e.g. Lazear (2000); Duflo *et al.* (2012); Lacetera *et al.* (2012); Homonoff (2018); Barrera-Osorio *et al.* (2019), among others. Otherwise, the plot is as expected due to theory and only a little publication bias is evident. We ended our endeavour with a conclusion that sometimes things are not such as they appear at a first glance and it is beneficial to verify the visual perception with hard data analysis. Our quantitative analysis in the fourth chapter indeed confirms this conclusion.

comment 4

It is great that you relax the assumption of conditional independence between estimates and standard errors. Could you provide some details on the IV estima-

Figure C.1.1: 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. Red dots denotes observations from study by Karlan & List (2007).

tion? Robust F- stat from the first stage + weak-instrument-robust confidence intervals in the second stage (see our paper on substitution between skilled and unskilled labor under revision for RESTAT). Also one could potentially use the new tests due to Elliott et al. (Econometrica; also see our RESTAT paper for implementation).

Thank you for the positive evaluation of our approach. I now provide requested details on the estimation in the Table 4.5. Moreover, I now provide the tests due to Elliott *et al.* (2022):

″ Elliott et al. *(2022), however, points out that in caliper tests and alike the researcher must arbitrarily specify the values where breaks in the distribution are expected. They derive two new rigorously founded techniques using a conditional chi-squared test that does not require the arbitrary definition of the breaks. Instead, the test slices the data to a specified number of bins and sets, therefore, the ″bars″ dynamically. The first technique is a histogram-based test for non-increasingness of the p-curve, the second one is a histogram-based test for 2-monotonicity and bounds on the p-curve and the first two derivatives. In their applications, Elliott* et al. *(2022) only focus on p-values below 0.15 and use 15, 30, or 60 bins. We set the target cutoff threshold for the discontinuity test similarly at 0.15 while using equally distributed 30 bins due to the smaller dataset size. We present the results of the test in Panel B of Table 4.6. We*

reject the null hypothesis of the absence of the publication bias on 90% interval. "

comment 5

I appreciate the use of the dilution prior that addresses collinearity in BMA.

Thank you for your appreciation.

D.3 Ing. Tomáš Miklánek, M.A., Ph.D.

D.3.1 Chapter 2 - Auctions

comment 1

Section 2.3 introduces hypotheses of the study, but it uses word "parameters" when talking about outcome variables. This is repeated at several consequent places in the text and is quite confusing. However, it is not used in this way in the published version of the article.

Interesting comment, it seems that our text went through significant proofreading in the publishing house. The text is different from what we have sent as our final version, it was rewritten from an active to a passive voice. This is, according to my perspective, not a best practice but, frankly speaking, I haven't noticed it so far. I have reviewed the hypotheses section and compared it to the published version of the paper but did not find what you meant with your comment. All the statements about parameters seem to have preserved their meaning. I also contacted the publishing house for the most recent production version of our article but I received an answer saying that: *"since the files are marked as deleted from our system, they are no longer retrievable."* I admit that the text of the first chapter could have been clearer but I do not change it here since the chapter is already published and it would, therefore, be even more misleading.

comment 2

I might be interested in the relationship between chat communication in the corresponding treatments and outcome variables.

This is a good question since it would be interesting to know this relationship in a robust form. And we have indeed studied it at the beginning of our research efforts. Some clues can be found in my previous research (Matoušek 2014) where I tried to determine the factors that influence the level of the relative efficiency of Simultaneous multi-round auctions. This part, however, eventually fell out of the analysis mainly due to interpretation problems and various other reasons. I can state here that the chat variable representing the presence of a possibility of communication among subjects had a significant and positive effect on the efficiency of the auction roughly around the value of 0.05.

D.3.2 Chapter 4 - Motivation

comment 3

Elaborate more on the first contribution - restriction of the sample to studies published in economic journals.

Thank you for the comment. Other referees also raised it in a more or less similar manner. I try to answer all three comments in one framework.

First, I overall improve the reasoning behind the contribution. I focus more on the use of the latest methods for the detection of selective reporting and tracing the causalities in the heterogeneity of estimates. I also elaborate more on the restriction of the sample to studies published only in economic journals. I focus here on the distinction between economic and psychological expectations and the importance of looking at the literature from separate perspectives. In this paper, however, I still focus strictly on the perspective of an economist. It would be indeed very interesting to add the psychological literature and quantitatively compare both, but I refrain now from this approach due to feasibility. I hope you will not complain when I state here that one can not do all the work at once and that leaving some pieces of the pie for later is, perhaps, not a bad strategy. I provide the improved contribution in the introduction section of the Chapter 4:

" Large heterogeneity among these results suggests that a synthesis of this topic would bring substantial value to the field. And indeed, the synthesis of the literature on the effect of rewards on motivation was done before (Rummel & Feinberg 1988; Cameron & Pierce 1994; Jenkins et al. *1998; Deci* et al. *1999; Cameron 2001; Cerasoli* et al. *2014; Van Iddekinge* et al. *2018, among others). None of the studies, however, tries to isolate the outlooks of either economists or psychologists by looking at the available literature from strictly separated perspectives. And yet an economist would have different outlook and, more importantly, the expectations than a psychologist. The former would expect a stronger effect of incentives while the latter would expect intrinsic motivation to have a greater impact on performance. A comparison of these two perspectives would be surely beneficial. Hence, we aim to synthesize the decades of research on this topic in a quantitative meta-analysis from separate perspectives, both economic and psychological. In this study, we focus on the former and look at how the effect behaves strictly across economic literature. We aggregate individual economic studies together being thus able to observe the underlying relationships and causalities of the rewards-motivation effect, along with potential systematic misbehaviour (Hunter* et al. *1982).*

Looking further at the list of synthetic studies that dealt with the topic be-

fore, there are only two studies Cameron & Pierce (1994); Cerasoli et al. *(2014) that examine whether there appears a phenomenon widespread not only in economics but also in other fields in the available literature—a selective reporting (Doucouliagos & Stanley 2013; Ioannidis* et al. *2017). Researchers and editors tend to publish only statistically significant results, yet, the insignificant and unpublished data contains a lot of valuable information that has the potential to add further value to general debate. We, therefore, analyze reported estimates from available studies and look for this hidden information. Employing the latest methodology in the field, we perform several linear and non-linear methods to uncover potential selective reporting in the economic literature on rewards and motivation.*

Last but not least, we trace the heterogeneity in the reported estimates to the design of the experiments while accounting for model uncertainty. We collect a set of 39 explanatory variables focusing on different angles related to effect characteristics, task nature, reward scheme, motivation characteristics, study design, subject pool characteristics, methodology, and publication characteristics. We employ the Bayesian model averaging (Raftery et al. *1997, BMA) and frequentist model averaging (Hansen 2007, FMA) to discover which characteristics affect the reported estimates the most. "*

D.4 Dr. Heiko Rachinger

D.4.1 Chapter 3 - Discrate

comment 1

First, I was surprised about the notion of domain independence as a low correlation between discount rates for different domains. Should this not be domain dependence since discount rates differ then between the different domains.

I acknowledge that this is misleading and perhaps unintuitive. It is an excellent point which also the editor of Experimental Economics raised and wanted to unite the wording with the literature. We argued with the fact that both domain independence and domain dependence terms occur in the literature but the independence one is more frequent. We then got approval from the editor to the wording independence as a terminology consistent with the literature.

comment 2

Second, including only published papers, while potentially increasing the quality, should have an effect of the publication bias correction. Would you interpret the found publication bias then as a lower bound of the actual one?

I would argue here that a rational researcher would want to have significant results even in the working paper version of her research and so the inflation of the results would be already in non-published papers. Moreover, for example, Havranek (2015) studies this effect between working papers and published papers in a specific domain and finds out that there is publication bias present in both unpublished and published research. I would therefore broadly expect that this relationship could be potentially translated to other contexts which would imply our found publication bias could also appear in unpublished papers. A researcher, however, can never be sure until she performs a thorough analysis of relevant data.

comment 3

Third, do you have any feeling of how well the bootstrap procedure for obtaining estimates of the standard error works. In fact, one could apply this method to papers for which standard errors are available and compare the bootstrapped standard errors with the actual ones.

I recalculate the bootstrapped values for the studies for which standard errors are available as suggested. To sum up the approach: for each study, we use 1,000 iterations for bootstrapping so that the mean of bootstrapped values equals the mean of the estimates reported in the study. From the bootstraps, we then approximate the standard error at the study level.

33 studies report the standard errors. I calculate the mean reported standard error and compare it to bootstrapped standard error at the study level as $abs(mean_{bootstrapped}/mean_{reported}-1)$ and express it in percentages. This difference varies from 9% to extreme 1453% with the median being at 66%. Half of the studies with bootstrapped values have those distorted at about 65% of the basis. The third quartile of the distribution is then at 140% level of distortion. The correlation between the mean reported standard error and bootstrapped standard error at the study level is 0.367.

I acknowledge that some of these distortions are quite large. Nevertheless, even though sometimes not precise, this technique still provides an option when standard errors are not available for some studies. We use it for 23 out of 56 studies in our sample (41%). The price for using all of the studies is reasonable in my perspective Moreover, we include a robustness check on the subsample of the data with only reported standard errors, and the results are in line with the original analysis.

D.4.2 Chapter 4 - Motivation

comment 1

First, the choice of only papers published in the economics literature still needs to be better motivated. In particular, if we expect results to systematically differ between the economics and psychological literature it could be interesting to incorporate both and quantify those differences.

Thank you for the comment. Other referees also raised it in a more or less similar manner. I try to answer all three comments in one framework.

First, I overall improve the reasoning behind the contribution. I focus more on the use of the latest methods for the detection of selective reporting and tracing the causalities in the heterogeneity of estimates. I also elaborate more on the restriction of the sample to studies published only in economic journals. I focus here on the distinction between economic and psychological expectations and the importance of looking at the literature from separate perspectives. In this paper, however, I still focus strictly on the perspective of an economist. It would be indeed very interesting to add the psychological literature and quantitatively compare both, but I refrain now from this approach due to feasibility. I hope you will not complain when I state here that one can not do all the work at once and that leaving some pieces of the pie for later is, perhaps, not a bad strategy. I provide the improved contribution in the introduction section of the Chapter 4:

" Large heterogeneity among these results suggests that a synthesis of this topic would bring substantial value to the field. And indeed, the synthesis of the literature on the effect of rewards on motivation was done before (Rummel & Feinberg 1988; Cameron & Pierce 1994; Jenkins et al. *1998; Deci* et al. *1999; Cameron 2001; Cerasoli* et al. *2014; Van Iddekinge* et al. *2018, among others). None of the studies, however, tries to isolate the outlooks of either economists or psychologists by looking at the available literature from strictly separated perspectives. And yet an economist would have different outlook and, more importantly, the expectations than a psychologist. The former would expect a stronger effect of incentives while the latter would expect intrinsic motivation to have a greater impact on performance. A comparison of these two perspectives would be surely beneficial. Hence, we aim to synthesize the decades of research on this topic in a quantitative meta-analysis from separate perspectives, both economic and psychological. In this study, we focus on the former and look at how the effect behaves strictly across economic literature. We aggregate individual economic studies together being thus able to observe the underlying relationships and causalities of the rewards-motivation effect, along with potential systematic misbehaviour (Hunter* et al. *1982).*

Looking further at the list of synthetic studies that dealt with the topic before, there are only two studies Cameron & Pierce (1994); Cerasoli et al. *(2014) that examine whether there appears a phenomenon widespread not only in economics but also in other fields in the available literature—a selective reporting (Doucouliagos & Stanley 2013; Ioannidis* et al. *2017). Researchers and editors tend to publish only statistically significant results, yet, the insignificant and unpublished data contains a lot of valuable information that has the potential to add further value to general debate. We, therefore, analyze reported estimates from available studies and look for this hidden information. Employing the latest methodology in the field, we perform several linear and non-linear methods to uncover potential selective reporting in the economic literature on rewards and motivation.*

Last but not least, we trace the heterogeneity in the reported estimates to the design of the experiments while accounting for model uncertainty. We collect a set of 39 explanatory variables focusing on different angles related to effect characteristics, task nature, reward scheme, motivation characteristics, study design, subject pool characteristics, methodology, and publication characteristics. We employ the Bayesian model averaging (Raftery et al. 1997, BMA) and frequentist model averaging (Hansen 2007, FMA) to discover which characteristics affect the reported estimates the most. "

comment 2

Second, do you expect that confining the analysis to the top 30 economic journals has an effect on the detected publication bias? In my understanding, in other fields, results in better journals tend to be more inflated than those published in worse journals.

I would again argue here that a rational researcher would want to have significant results even before the publication of her research and so the inflation of the results would be already in working papers. Moreover, Havranek (2015) confirms for a specific domain that there is publication bias present in both unpublished and published research. But according to my knowledge, there is no clear consensus among researchers about this and I would therefore not like to draw any conclusions here. The safest option would again be to gather more studies and perform differential analysis, but this is out of the scope of this Thesis.

comment 3

Third, the choice of the partial correlation coefficient (PCC) relative to Cohen's d or Pearson correlation should be better explained.

This is a valid point. We failed to explain correctly the use of the Partial correlation coefficient. I now provide a better explanation in Section 4.3:

" The effect variables we code measure the relationship between incentives and output such as a change in physical/mental performance, pro-social behaviour, or students' Grade Point Average. This makes the collected estimates to be diverse in nature and size. We, therefore, need a measure that allows us to unify and compare the effects. Previous meta-analyses used mostly Cohen's d (e.g. Cameron & Pierce 1994; Jenkins et al. 1998) or Pearson correlation (Cerasoli et al. 2014; Van Iddekinge et al. 2018). Those measures are, however, inapplicable in our case. We would not be able to calculate Cohen's d for every data point since our dataset contains necessary estimates for control groups in only 13 studies out of the 44 (29.55%). Pearson correlation coefficient on the other hand does not control for confounding variables.

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 and *would also control for omitted variables. A Partial correlation coefficient (PCC) is presumably the most fitting choice, appearing as a standard in numerous meta-analyses (e.g. Doucouliagos & Laroche 2003; Zhou* et al. *2013; Valíčkova* et al. *2015; 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 while ignoring the size of the dataset (Stanley & Doucouliagos 2012)."*

comment 4

Fourth, the publication bias direction seems to depend on the underlying theory, i.e. some papers might be looking for positive significant effects and others for negative ones. Is it possible to code this potential direction and incorporate it? Also does the correlation between estimates and their standard errors appropriately reflect the underlying selective reporting? Further, are selection criteria based on statistical significance of the PCC appropriate (for example, applying the Caliper test at 1.96)?

Regarding the direction of the sought effects and publication bias, we do not have these prior beliefs about the experiment's outcomes in the data. It would be very hard to gather this information and code it into the data since every researcher may have different prior beliefs about those outcomes. It even may be impossible to code it rigorously, a lot of noise would be introduced when potentially trying to imply such information from the text. What we do have is the information about the direction of the effects in the variable Effect positive, i.e. whether a higher value of the effect means better performance (such as when measuring the number of clicks) or the opposite is true (higher finish time in a race). We normalize the data when the Effect positive variable is negative. The Effect positive would, therefore, serve as a poor proxy for theoretical expectations about the outcome. It should not have any effect on publication bias.

As far as the correlation between estimates and their standard errors is concerned, we use methods for publication bias detection that do not assume any correlation between estimates and standard errors. First, we perform a set of caliper tests on arbitrary slices of data. Since the correlation of estimates and their standard errors persists even on these slices of data, the publication bias is then not driven by the selection of underlying methodology only. Second, we perform the p-uniform* technique that focuses on the distribution of p-values at the mean underlying effect size (van Aert & van Assen 2020). Third, we now provide the most novel method of Elliott *et al.* (2022) which uses a dynamic approach to the definition of the breaks in the distribution of p-values. All three techniques indicate publication bias, as we describe in the paper.

To the last point of the appropriateness of selection criteria based on the statistical significance of the PCC within the caliper test, I must note that for both the caliper tests as well as the test by Elliott *et al.* (2022) we use the original t-statistic that the authors reported in the primary studies. So the fact that we are using Partial correlation coefficients does not affect it in our case.

comment 5

Fifth, I am not sure whether including an "effect positive" variable is enough to deal with the effect's different direction. In particular, should this variable not be interacted with all other explanatory variables as well?

The variable Effect positive expresses whether a higher value of the effect means better performance (such as when measuring the number of clicks) or the opposite (higher time in the race means worse performance). It only serves as a normalization variable for the unification of the effect's direction. We multiply the PCC by (-1) when the Effect positive dummy is negative. The Effect positive is by no means a variable that would carry any crucial information and would, therefore, serve as a poor proxy for theoretical expectations about the outcome of an experiment. It should not have any effect on publication bias. Conceptually, it does not make sense to me, in this case, to interact this proxy with all explanatory variables.

D.5 Prof. Dr. Sebastian Gechert

D.5.1 Chapter 2 - Auctions

comment 1

I am afraid, I am a bit puzzled about what is the problem to be solved. Is it that package bidding makes auctions complicated for bidders and thus less efficient or is it that package bidding does not suffice to prevent collusion and thus lead to too low auction prices at the cost of the seller?

Generally, we wanted to compare a Simultaneous multi-round auction (SMR) with its package bidding extension (SMRPB) in the setting of the 2013 Czech spectrum auction. Theoretically, the SMRPB auction should have delivered better results in terms of efficiency and revenues than the SMR auction (Goeree & Holt 2010). Moreover, we added communication to this setup to allow for collusion among bidders. We were interested in which of the auction formats would be more efficient under collusion and whether some of them can potentially break it.

When comparing the standard treatments our results suggest that with large amounts of goods involved package bidding makes auctions indeed complicated for bidders and thus less efficient than the literature originally suggested. We cannot confirm the conclusions of Goeree & Holt (2010) that package bidding is better than a standard ascending auction.

In the communication treatments our experiments show evidence that the package bidding auction may break collusion among cartel members, but for the price of the winner's curse. Since bidders compete for the combinations of goods (packages) that cannot interfere with each other, a lot of bidders may not get their desired packages. Even though collusion is broken and prices are competitive, many goods stay unallocated and the revenue for the seller decreases together with the efficiency of the package bidding auction. Bidders in the standard multiunit ascending auction compete for each good individually and can therefore reach agreements about these units more easily. The collusion is facilitated in SMR and also more goods are sold which leads to higher efficiency and revenues.

comment 2

Also, after the long motivation part of the paper, I am still not sure why one exactly needs the 2x2 design. This should be motivated more clearly right from the start (it is motivated later in the paper).

Thank you for your comment. I see now that this should gave been explained better much sooner than on page 19 in chapter 2.4 Methodology. We consulted the paper with many researchers but unfortunately, none suggested such obvious improvement in the clarity of the paper. Maybe if they had done it the same way you do now, we could have published our research better. Since this research is published and already contains an explanation that should only have been provided earlier, I do not interfere with the text.

comment 3

The author states on p. 9: "The main concern of every auctioneer should be the efficiency of the type of the auction employed, that is, allocating the objects for sale to those who value them the most." I would not agree with this statement. I would suspect, the main concern of an auctioneer is to maximize its brokerage, which may or may not coincide with allocative efficiency.

Thank you for your comment. I acknowledge that not every auctioneer, but I would argue that possibly that one who is in charge of the large public auction should be concerned not only with revenues but also with the efficiency of the auction. Since complicated auction mechanisms are artificially created markets, they suffer from various flaws such as the winner's curse. Excessive prices of auctioned goods could subsequently lead to excessively high prices for customers. It is precisely why we suggest that auctioneers in high-stakes public auctions should use an experimental evaluation of the auction design before it is implemented so that potential flaws are revealed and the real auction is carried out correctly. This was done for example by Abbink *et al.* (2005) or Bichler *et al.* (2014). Such an evaluation would help the auctioneer find the efficiency-maximizing or revenue-maximizing design alternative, depending on their preferences.

comment 4

Finally, a question on external validity of the experiment: do economics students match the real-world sample of professional bidders that probably know each other's stakes well and have a common history in oligopolistic auction markets? In particular, the finding that SMRPB auctions may be overcomplicated and thus reduce efficiency may not carry over to professional bidders.

This is a very good comment and I would like to thank you for it. Your concern is even supported by my research presented in both Chapter 3 where experiments working exclusively with students show less evidence for patience than experiments using mixed population samples and Chapter 4 where usage of a students' samples reports lower performance. When designing the experiment we were aware of potential consequences that stemmed from the quality of the subject pool. We approached this topic by employing a complicated multi-stage hiring procedure that ensured only competent subjects skilled in auction experiments should have participated in our experiment. The procedure is described on page 27 of the Thesis as follows:

" The complexity of the required task to be done in the laboratory was expected to be highly demanding. We were not able to train subjects specifically before the experiment or to carry out the complicated procedures used for example in Abbink et al. *(2005); Brunner* et al. *(2010) or even Bichler* et al. *(2013). This was mainly due to the necessity of high over-recruitment rates in the case of such training and the tightly constrained funding of the research. Therefore, we used a simpler procedure.*

The participants received an invitation five days prior to the experiment and three days prior were asked to fill in an online questionnaire based on the partial instructions available online. This online material consisted of general instructions common to all treatments of the experiment. The instructions were concluded with a 5-question quiz. Those who filled in the questionnaire correctly were preferably invited to the lab. There were no difficulties with the online questionnaires, and the rate of successful completion was over 95%. The whole procedure regarding the instructions in advance and the questionnaire was described in the invitation email for the experiment and was therefore publicly known. "

D.5.2 Chapter 3 - Discrate

comment 1

I suspect that many experimental approaches rule out the possibility of negative discount rates by design. This is different to other forms of publication bias, where researchers strive for statistically significant coefficients or discard non-conformist findings from their regressions. Doesnâ€™t this pose a problem to theories of publication bias that propose a correlation between the point estimate and the standard error as signs of publication selection? Can we employ these *techniques here? Moreover, the only negative estimates, according to figure 3.2 stem from one publication (Loewenstein 1987), which seems to set up a negative framing and thus may simply be an artifact of the framing. Does excluding this study change the assessment of publication bias?*

This is another great comment and I thank you for it. We had a serious discussion on this topic with the two referees from Experimental Economics throughout the revision process and we discuss it in the paper in Section 3.4 about publication bias. To dissect your comment in more depth, it implies that a correlation between estimates and standard errors can arise in the absence of publication bias, which invalidates the identification of all publication bias tests commonly used in economics. In other words, the standard error in meta-regression can easily be endogenous, which has important consequences beyond our application of meta-analysis on individual discount rates.

Trying to explain that these concerns do not drive our results, we employ two techniques. First, we perform a set of caliper tests on arbitrary slices of data. Since the correlation of estimates and their standard errors persists even on these slices of data, the publication bias is then not driven by the selection of methodologies only. Next, we perform the p-uniform* technique that makes no assumptions regarding the correlation between estimates and standard errors; instead, it focuses on the distribution of p-values at the mean underlying effect size (van Aert & van Assen 2020). We also put more weight on our instrumental variable estimation (using the inverse of the square root of the number of observations as an instrument for the standard error) and fixed effects estimation (which only needs to assume that the standard error is exogenous within studies, not between studies). All these techniques indicate publication bias, as we describe in the paper. More can be found in the respective Section 3.4 about publication bias.

D.5.3 Chapter 4 - Motivation

comment 1

This chapter somewhat lacks the clarity and novelty of the previous chapter. Regarding clarity, the chapter would definitely benefit from professional proof reading. Some sentences are not entirely clear in their meaning.

I am aware that the coherence and clarity of chapter 4 were not of a high level for the predefense stage of the dissertation. This chapter was therefore additionally proofread after incorporating all the revisions and the errors were corrected.

comment 2

Regarding novelty, the paper is one among many existing meta-analyses in the field, defining itself as the first one to cover only publications from economic *journals. I am not convinced why this would be an improvement over the existing literature.*

Thank you for the comment. Other referees also raised it in a more or less similar manner. I try to answer all three comments in one framework.

First, I overall improve the reasoning behind the contribution. I focus more on the use of the latest methods for the detection of selective reporting and tracing the causalities in the heterogeneity of estimates. I also elaborate more on the restriction of the sample to studies published only in economic journals. I focus here on the distinction between economic and psychological expectations and the importance of looking at the literature from separate perspectives. In this paper, however, I still focus strictly on the perspective of an economist. It would be indeed very interesting to add the psychological literature and quantitatively compare both, but I refrain now from this approach due to feasibility. I hope you will not complain when I state here that one can not do all the work at once and that leaving some pieces of the pie for later is, perhaps, not a bad strategy. I provide the improved contribution in the introduction section of the Chapter 4:

" Large heterogeneity among these results suggests that a synthesis of this topic would bring substantial value to the field. And indeed, the synthesis of the literature on the effect of rewards on motivation was done before (Rummel & Feinberg 1988; Cameron & Pierce 1994; Jenkins et al. *1998; Deci* et al. *1999; Cameron 2001; Cerasoli* et al. *2014; Van Iddekinge* et al. *2018, among others). None of the studies, however, tries to isolate the outlooks of either economists or psychologists by looking at the available literature from strictly separated perspectives. And yet an economist would have different outlook and, more importantly, the expectations than a psychologist. The former would expect a stronger effect of incentives while the latter would expect intrinsic motivation to have a greater impact on performance. A comparison of these two perspectives would be surely beneficial. Hence, we aim to synthesize the decades of research on this topic in a quantitative meta-analysis from separate perspectives, both economic and psychological. In this study, we focus on the former and look at how the effect behaves strictly across economic literature. We aggregate individual economic studies together being thus able to observe the underlying relationships and causalities of the rewards-motivation effect, along with potential systematic misbehaviour (Hunter* et al. *1982).*

Looking further at the list of synthetic studies that dealt with the topic before, there are only two studies Cameron & Pierce (1994); Cerasoli et al. *(2014) that examine whether there appears a phenomenon widespread not only in economics but also in other fields in the available literature—a selective reporting (Doucouliagos & Stanley 2013; Ioannidis* et al. *2017). Researchers and editors tend to publish only statistically significant results, yet, the insignificant and unpublished data contains a lot of valuable information that has the potential to add further value to general debate. We, therefore, analyze reported estimates from available studies and look for this hidden information. Employing* *the latest methodology in the field, we perform several linear and non-linear methods to uncover potential selective reporting in the economic literature on rewards and motivation.*

Last but not least, we trace the heterogeneity in the reported estimates to the design of the experiments while accounting for model uncertainty. We collect a set of 39 explanatory variables focusing on different angles related to effect characteristics, task nature, reward scheme, motivation characteristics, study design, subject pool characteristics, methodology, and publication characteristics. We employ the Bayesian model averaging (Raftery et al. *1997, BMA) and frequentist model averaging (Hansen 2007, FMA) to discover which characteristics affect the reported estimates the most. "*

comment 3

The author sells the use of the partial correlation coefficient (PCC) as a unique feature of the analysis as compared to other contributions. However, PCCs can only measure statistical significance, not economic significance. That is why they are usually considered second best only when a homogeneous coefficient across studies is unavailable. The danger when using PCCs often is a misinterpretation of the findings. When the author in his conclusion states that "[r]ewards have, therefore, only about a halfway effect on motivation and performance of people than a simple mean summary statistics of economics literature suggests", then he also commits such an error. The corrected mean of the PCC after publication bias only can show that the statistical significance of the effect has shrunk, not the effectiveness of financial incentives on work performance.

This is a true and valuable comment. I forgot this perspective at one point in the paper and thank you for taking it forth. The effect variables across studies are indeed not homogeneous and that is why we need to use Partial correlation coefficients as you correctly point out. I drop the discussion about the PCC being a unique feature of our analysis in the introduction of the paper (see my answer to your comment 2 in chapter 4). I also provide a better explanation as well as tone down the voice in Section 4.3 when I introduce the usage of PCC in the paper, see below. Last but not least I also drop the erroneous statement about the size of the effect the rewards have on motivation in the conclusions of the paper.

" The effect variables we code measure the relationship between incentives and output such as a change in physical/mental performance, pro-social behaviour, or students' Grade Point Average. This makes the collected estimates to be diverse in nature and size. We, therefore, need a measure that allows us to unify and compare the effects. Previous meta-analyses used mostly Cohen's d (e.g. Cameron & Pierce 1994; Jenkins et al. *1998) or Pearson correlation (Cerasoli* et al. *2014; Van Iddekinge* et al. *2018). Those measures are, however, inapplicable in our case. We would not be able to calculate Cohen's d for every data point since our dataset contains necessary estimates for control*

groups in only 13 studies out of the 44 (29.55%). Pearson correlation coefficient on the other hand does not control for confounding variables.

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 and would also control for omitted variables. A Partial correlation coefficient (PCC) is presumably the most fitting choice, appearing as a standard in numerous meta-analyses (e.g. Doucouliagos & Laroche 2003; Zhou et al. *2013; Valíčkova* et al. *2015; 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 while ignoring the size of the dataset (Stanley & Doucouliagos 2012)."*

comment 4

I also wonder whether it is wise to lump together analysis from such different estimates like "students' GPA, charitable giving, an outcome of a game or a simulation, performance of employees at work" (p.102). I am not convinced as to whether the "middle ground in this trade-off between incomparability and excessive generalization" (p.102) is really a trade-off. Excessive generalization leads to incomparability as in this case.

Thank you for your comment. I admit the statement was indeed incorrect. However, there are examples of studies with at least as heterogeneous the nature of the effects as in our work. For example, DellaVigna & Linos (2022) introduces a highly heterogeneous dataset and uses a similar approach as we do. I tried to rewrite the paragraph to express more clearly our approach here.

" With this kind of variety in the data, it seems unfeasible to simply lump all of the effects into one category. This approach is heavily criticized in the field. For example, Glass et al. *(1981) argue that conclusions drawn when generalizing different effects are invalid. Distributing the effects into too many categories would be on the other hand misleading for the reader as well as technically infeasible. The analysis would lose its point. Some degree of generalization is necessary for a meta-analysis. One of the latest examples of a meta-study with substantial heterogeneity in the estimates is e.g. DellaVigna & Linos (2022) who compare interventions in research units, versus at scale implemented in Nudge Units in governments. Fortunately, we observe a clear underlying pattern between the studies, which allows us to create a reasonable categorization according to their nature. Namely, we create four categories capturing: students' GPA, charitable giving, an outcome of a game or a simulation, and performance of employees at work. All of the studies collected in our data fit into one of these four categories, making this setup appear suitable. By our approach, we aim to choose a middle ground in the degree of generalization for enabling comparability and provide insights into the inner workings between effects of different nature. "*