Essays in Applied Microeconomics

Inaugural-Dissertation

zur Erlangung des Grades eines Doktors
der Wirtschafts- und Gesellschaftswissenschaften

durch

die Rechts- und Staatswissenschaftliche Fakultät der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Jonas Radbruch

aus Duisburg

2020

Dekan:	Prof. Dr. Jürgen von Hagen
Erstreferent:	Prof. Dr. Sebastian Kube
Zweitreferent:	Prof. Dr. Matthias Kräkel
Tag der mündlichen Prüfung:	13. March 2020

Acknowledgements

This thesis would not have been possible without the support of many people. First of all, I would like to thank my main supervisor Sebastian Kube, for his support and encouragement throughout the development of this thesis. He always had an open ear for my research ideas and projects, of which some have manifested themselves in chapters of this thesis. He always provided sharp feedback or motivation and support when necessary. I am also indebted to my second supervisor, Matthias Kräkel, who always provided a fresh perspective on the chapters and helped to sharpen their focus.

I would also like to extend my gratitude to many others, who helped me throughout the last few years. Among those, most importantly, I thank Steffen Altmann, who is not only a long-term collaborator, co-author and my official BGSE Mentor, but has also provided a lifetime of advice to me. My gratitude also extends to my fabulous set of co-authors, Sebastian J. Goerg, Andreas Grunewald, Lukas Kießling, Sebastian Schaube and Amelie Schiprowski. Over the years we had many fruitful – if sometimes brutal – discussions, spent endless hours in gym classes, maybe even more in the EconLab, polished texts together or discussed big questions and little details.

I am also thankful for many other colleagues, office mates and members of the faculty, who supported me in writing this dissertation. They provided helpful comments, feedback and support or they taught me the skills needed. This includes Thomas Dohmen, Hans-Martin von Gaudecker, Ingo Isphording, Fabian Kosse, Robert Mahlstedt, Nico Pestel and many others. A special thanks is also due to Holger Gerhardt who sometimes provided even more help than one had asked for. I also acknowledge the remarkable administrative support by the administration of the Institute for Applied Microeconomics, namely Simone Jost and Andrea Reykers, as well as the administration of IZA.

Starting graduate school and finishing it can be a journey. This journey would not have been the same without the enduring support of my friends and fellow PhD students Zvonimir Basic, Dmitry Kuvshinov, Felix Schran, Louis Strang and Kaspar Zimmermann. They managed to make graduate school an enjoyable place.

Finally, I thank my parents for encouraging me to study economics in the first place, and my family for putting up with me all the way to the end. Therefore, the most important “Thank you” of all is reserved for my wife Franziska, who has pro-

vided support if something did not go as expected, celebrated the happy moments with me and endured me over all these years. This thesis would not have been possible without her faith and support.

Contents

List of Figures

List of Tables

Introduction

One of the higher-order goals of economic analysis is the derivation of sound policy recommendations. Prior to designing (optimal) policies, we have to understand the determinants of human behavior. Important aspects of human behavior involve the motivation of individuals, the formation of subjects assessments and allocation of cognitive resources. Firms and organizations, for example, have to understand the behavior of their employees. How do workers react to different work environments and incentive schemes? This knowledge is substantial to decide which work environment to implement, or, given the work environment, which incentive scheme to use. Moreover, supervisors and teachers have to understand the effects of (allowing for) self-selection of peers, as a sound understanding of peer effects is imperative for the exploitation of these. Moreover, we need to understand the formation of subjective assessments, as those are often used to evaluate otherwise unobserved quality. Are interviewers able to rate candidates independently, or do they suffer from biases and heuristics obfuscating their judgment? Similar arguments hold for the evaluation of policy interventions, which try to foster active behavior. Who is affected by those interventions and are there spillovers to other choice domains?

This dissertation consists of four independent chapters, which all study determinants of human behavior. All chapters leverage insights from behavioral economics or methods from experimental economics. The unifying framework is that I use empirical evidence from natural, field or lab experiments, which allow a clean causal identification of different determinants of human behavior.

In chapter one (based on joint work with Sebastian J. Goerg and Sebastian Kube), I study the role of implicit effort costs for effort provision and the effectiveness of incentive schemes. Agents' decisions to exert effort depend on the incentives and the potential costs involved. So far, most of the attention has been on the incentive side. However, our laboratory experiments underline that both the incentive and the cost side can be used separately to shape work performance. In our experiment, subjects work on a real-effort slider task. Between treatments, we vary the incentive scheme used for compensating workers. Additionally, by varying the available outside options, we explore the role of implicit costs of effort in determining workers' performance. We observe that incentive contracts and implicit costs interact in a nontrivial manner. In general, performance decreases as implicit costs increase.

Yet the magnitude of the reaction differs across incentive schemes and across the offered outside options, which, in turn, alters estimated output elasticities. In addition, comparisons between incentive schemes crucially depend on the implicit costs.

In chapter two (based on joint work with Andreas Grunewald and Steffen Altmann), I study the role of scarce cognitive resources as a source of passive behavior and the impact of choice-promoting policies for people with scarce cognitive resources. Passive behavior is ubiquitous even when facing various alternatives to choose from, people commonly fail to take decisions. This chapter provides evidence on the cognitive foundations of such "passive choices" and studies implications for policies that encourage active decision making. In an experiment designed to study passive behavior, we document three main results. First, we demonstrate that scarcity of cognitive resources leads to passive behavior. Second, policies that encourage active choice succeed in reducing passivity and improve decisions in the targeted domain. Third, however, these benefits of choice-promoting policies come at the cost of negative cognitive spillovers to other domains.

In chapter three (based on joint work with Lukas Kiessling and Sebastian Schaube), I study the impact of self-selected peers on performance. In many natural environments, carefully chosen peers influence individual behavior. Using a framed field experiment at secondary schools, we examine how self-selected peers affect performance in contrast to randomly assigned ones.We find that self-selection improves performance by approximately 15% of a standard deviation relative to randomly assigned peers. Our results document peer effects in multiple characteristics and show that self-selection changes these characteristics. However, a decomposition reveals that variations in the peer composition contribute only little to the estimated average treatment effects. Rather, we find that self-selection has a direct effect on performance.

In chapter four (based on joint work with Amelie Schiprowski), I study how the assessment of a candidate is influenced by the other candidates seen by the same interviewer. We leverage novel data on more than 9,000 interviewer assessments made within the admission process of a large study grant program. We find that a candidate's assessment decreases in the measured quality of all other candidates seen by the same interviewer. The influence of the previous candidate, however, exceeds the influence of any other candidate by a factor of about three. The additional effect of the previous candidate appears to be driven by the exaggeration of small differences between current and previous candidate quality. Moreover, it is asymmetric with respect to gender and favors male candidates who follow a female candidate.

Chapter 1

The Effectiveness of Incentive Schemes in the Presence of Implicit Effort Costs*

Joint with Sebastian J. Goerg and Sebastian Kube

1.1 Introduction

What are the determinants of effort provision, and how to incentivize agents to exert high effort? Most studies addressing these questions usually focus on the compensation side, investigating effort responses to fixed and variable wages (Lazear, 2000; Carpenter, 2016), fair wages (Cohn, Fehr, and Goette, 2015), or other contractual details of the incentive scheme (Winter, 2004; Goerg, Kube, and Zultan, 2010; Herweg, Müller, and Weinschenk, 2010). Yet, behavior of agents also depends on additional non-monetary features of the work environment. Examples for such additional influences include task-specific intrinsic motivation (Deci, 1971), recognitions and awards (Kosfeld and Neckermann, 2011; Bradler, Dur, Neckermann, and Non, 2016), personal goals (Koch and Nafziger, 2011; Goerg and Kube, 2012; Corgnet, Gómez-Miñambres, and Hernán-González, 2015), and restrictions on behavior (Falk and Kosfeld, 2006). In this paper, we demonstrate that the opportunity costs of effort, which crucially depend on the work environment, play a central role for effort provision in general and for the effectiveness of incentive schemes in particular.

More generally, effort provision by an agent is determined not only by the incentives provided for a given task, but also by the effort costs an agent faces. Effort costs can be financial expenditures, but more importantly they comprise opportunity costs of foregone alternative activities (see, for example, Holmstrom and Milgrom,

* We are grateful to John Hamman, Lukas Kiessling and Felix Schran for helpful comments on earlier drafts. We thank Anne Mertens for excellent research assistance.

1987). Therefore, incentives to perform in a given task can generally be provided by either setting the incentive scheme or by controlling the outside options of an agent (Holmstrom and Milgrom, 1991). Whereas the incentive side of the problem has been extensively studied, the interaction of outside activities and incentive schemes has been largely ignored. We intend to close this gap with the help of a real-effort experiment in which subjects work on the slider task (Gill and Prowse, 2012) while we manipulate opportunity costs and incentives.

In order to manipulate the opportunity costs, we implement three different work environments resulting in different implicit costs.[1] The first environment, Fix, is a standard lab environment in which subjects have to stay and perform the real-effort slider task for a fixed period of time. In the other two environments, we increase the implicit costs by giving subjects the opportunity to reduce the time they work on the task and allowing them to allocate their time differently. In the environment Inet, subjects can either work on the task or surf the internet; however, they have to stay in the lab for the same time as in Fix. In the environment Free, subjects are free to quit the task and leave the lab early. On the second dimension, we vary the incentive schemes under which the subjects are working. We implement two different piecerate schemes (Piecerate-Low, Piecerate-High) and two non-discretionary bonus schemes. In the first bonus scheme, the necessary output threshold is easy to achieve (Bonus-Easy), and in the second one it is (nearly) impossible to achieve (Bonus-Hard).

We observe higher output in the Fix-environment compared to the Inet- and Free-environments with increased implicit costs. Free results in an even sharper decrease as Inet. This result shows that the increase in implicit effort costs decreases performance. The decrease of output compared to Fix can be observed across all incentive schemes, yet with different magnitudes for each incentive scheme. Free does not result in lower output for both piecerates compared to Inet, but does for both bonus-based incentive schemes. For the latter, the opportunity to leave the lab leads to a stronger decrease in output than the opportunity to use the internet. The different reaction to the introduction of implicit costs across incentive schemes and across implicit effort costs leads to differences in the comparison of incentive schemes, depending on the work environment. In the Fix-environment, all four incentive schemes result in rather similar outputs, although marginal incentives vary substantially. Only the high piecerate leads to a slightly higher output. In the Inet- and Free-environments, subjects are more likely to actually respond to incentives and we observe positive output elasticities for the response to piecerates.

This study contributes to the empirical and experimental literature studying the reaction to incentives (for overviews, see Charness and Kuhn, 2011; Lazear

1. Opportunity costs are the sum of the direct and explicit effort costs a worker bears as well as the implicit effort costs which constitute the foregone utility by not allocating these resources towards an alternative activity.

and Oyer, 2012; Camerer and Weber, 2013). The seminal work of Nalbantian and Schotter (1997), as well as many follow-up studies, examined how incentive systems should be designed to induce high performance without causing negative side effects. The overall finding is that (monetary) incentives change behavior, yet sometimes evoke possible dysfunctional responses (e.g., Asch, 1990; Ordóñez, Schweitzer, Galinsky, and Bazerman, 2009; Gneezy, Meier, and Rey-Biel, 2011; Larkin, 2014). For example, people might show a negative response to the introduction of a very small piecerate (e.g., Gneezy and Rustichini, 2000), performance decreases as incentives become too large (e.g., Ariely, Gneezy, Loewenstein, and Mazar, 2009), or the strength of incentives and performances might generally follow an inverse u-shaped relationship (e.g., Pokorny, 2008). We demonstrate that not only the incentive side of the problem has to be taken into account, but that the opportunity cost side also plays a crucial part which is often neglected.

Methodologically, our paper adds to the literature using real-effort experiments, which are *"considered to be a better match to the field environment."* (Charness and Kuhn, 2011). Just recently, Herbst and Mas (2015) concluded in a meta-study on peer-effects that particularly experiments with real-effort tasks *"simulate realistic work environments"*. Real-effort experiments have been used to study such diverse phenomena as gender effects in competition (Niederle and Vesterlund, 2007), office politics (Carpenter, Matthews, and Schirm, 2010), and sorting into incentive schemes (Dohmen and Falk, 2011). So far, most of the experimental literature using real-effort experiments has considered fixed-time environments or fixed work requirements.[2] By the nature of those experiments, performance changes can only be due to a change in the explicit costs of effort.[3] One recent example of a study that changes the explicit effort cost is by Gächter, Huang, and Sefton (2016), who combine a real-effort task with induced effort costs. In their study, the explicit costs of effort are exogenously varied by inducing different costs for an action. Implicit costs play only a minor role in those experimental procedures, as subjects have to stay in the lab for a fixed time or until a task is completed. Other studies induce implicit costs through outside options, but do not vary them between treatments. Commonly used outside options are leaving the lab (e.g., Abeler, Falk, Goette, and Huffman, 2011; Rosaz, Slonim, and Villeval, 2016), paid pause buttons (e.g., Mohnen, Pokorny, and Sliwka, 2008), surfing the internet (e.g., Corgnet, Gómez-Miñambres, and Hernán-González, 2015), or reading magazines (e.g., Charness, Masclet, and Ville-

2. One notable exception to this is Noussair and Stoop (2015), who use the time spent in the laboratory as a medium for reward. There, higher payoffs lead to shorter time in the lab. Similarly, Danilov and Vogelsang (2016) study time investments as pro-social giving.

3. See Kurzban, Duckworth, Kable, and Myers (2013) for a related discussion in Psychology on salience and the effect of opportunity costs on task performance.

val, 2014).[4] However, those studies offer the outside option to every subject and do not manipulate the option.[5]

Our experiment is complemented by other studies that manipulate outside options (see e.g., Dickinson, 1999; Eckartz, 2014; Corgnet, Hernán-González, and Schniter, 2015; Koch and Nafziger, 2016; Erkal, Gangadharan, and Koh, 2018). The paper closest to ours is by Corgnet, Hernán-González, and Schniter (2015). They study the effect of piecerate and team incentives while varying the access to one real-leisure option, namely internet browsing. Their key finding is that the availability of the real-leisure alternative leads to a sharper decrease in performances under team-based incentives than under piecerate incentives. Their study shows that implicit effort costs might play a role in determining effort. Our study takes this as a starting point to further investigate the role of implicit effort costs. The focus of our study, however, differs from their paper in at least two crucial aspects. First, the focus of our paper is on individual incentives studying two piecerate and two bonus schemes. Second, we manipulate the implicit costs in various ways and demonstrate that the effectiveness of the incentive schemes differs between work environments. Our study therefore investigates, in an unified framework, four commonly used individual incentives schemes in various environments. Our results demonstrate that the effectiveness of incentive schemes crucially depends on the work environment. This helps to explain why in some work environments incentives might not change behavior. This non-responsiveness is unrelated to the monetary incentive side of the problem, but simply due to the absence of implicit effort costs or to low opportunity costs. In addition, our study helps to explain why incentives sometimes might not change behavior in real-effort experiments (e.g., Araujo, Carbone, Conell-Price, Dunietz, Jaroszewicz, et al., 2016). If individuals face (nearly) no costs for their effort or behavior, the corresponding behavior might not be altered by the incentive structure. Managers who are able to control the opportunity costs of effort directly might want to take this into account and consider this part of the work environment more closely. Yet, even if the management is not able to control the costs of effort directly, it should take into account that the behavioral responses to the incentive schemes depend on the given work environment. Thus, our results show the importance of taking implicit as well as explicit costs into account when studying the behavioral response to incentive schemes or implementing them in practice.

4. For more examples of papers using these outside options, see Tables 1 and 2 in the Online Appendix A.

5. Some studies require outside options as only in their presence subjects are able to respond to the treatments, i.e., have a labor-leisure tradeoff (e.g., Kessler and Norton, 2016). In other studies, the usage of the outside option is the dependent variable of the experiment (e.g., Rosaz, Slonim, and Villeval, 2016).

The remainder of the paper is organized as follows. In Section 2, we describe the design of our experiment and provide some behavioral hypotheses. Section 3 presents the results of the experiments. We conclude in Section 4.

1.2 Design

Opportunity costs are the sum of implicit and explicit costs. In our computerized real-effort experiment, we keep the explicit costs fixed while manipulating incentives and implicit costs between treatments. Based on the slider task by Gill and Prowse (2012), subjects had to adjust sliders ranging from 0 to 100 to the middle (50).[6] Each screen had 5 sliders that needed to be adjusted in order to finish a screen. The current number of finished screens was displayed on the screen (see Figure 1.A.1 in the Appendix for a screenshot) and the total number was later used to calculate the payments. The task was constant in all treatments and the effort to move the slider represented the explicit cost part of the opportunity costs. The experiment consisted of 3 stages and implicit effort costs were manipulated in the second stage.[7]

In the first stage, subjects worked on the real-effort task for 5 minutes without any monetary incentives. This stage served two purposes: First, subjects learned the difficulty of the task and could form accurate expectations about the effort costs, and secondly, it provided an ability measure which is not influenced by the subsequent incentive scheme.[8,9] Afterwards subjects received treatment-specific instructions and were informed about the subsequently applied incentives. Independent of the treatment, all instructions stressed that subjects should accomplish as many screens as possible.[10] The dependent variable — output, i.e., number of completed screens — was obtained in the second stage of the experiment. In this stage, subjects had to work on the real-effort task for a maximum of 40 minutes. The exact implementation of this stage depended on the treatment. In the third stage, subjects had to answer a short questionnaire, including sociodemographics, the ten-item version of the Big Five personality measure (Rammstedt and John, 2007), cognitive reflection test (Frederick, 2005), and general risk attitude (Dohmen, Falk, Huffman, Sunde, Schupp, et al., 2011).

Treatments were implemented in the second stage following a full 4 × 3 factorial design. Table 2.1 summarizes the implemented treatments. In the first treatment di-

6. Subjects could only use the computer mouse. Keyboard and mouse wheel were disabled.

7. An English translation of the instructions is provided in Online Appendix F.

8. Strictly speaking, our ability measure is not able to differentiate between ability and intrinsic motivation. We thank an anonymous referee for pointing this out.

9. Another possibility would have been to use an incentivized measure of ability. We choose not to incentivize this, since we didn't want subjects to experience different incentive schemes in the experiment.

10. This was done as to minimize potential differences in crowding out of intrinsic motivation between work environments.

Table 1.1. Treatments and Description of the 4 × 3-Design

Treatment Name	Incentives Scheme	Work Environment
Piecerate-Low Fix	€0.02 per finished screen	No outside option, fixed duration of 40 minutes
Piecerate-Low Inet	€0.02 per finished screen	Internet allowed, fixed duration of 40 minutes
Piecerate-Low Free	€0.02 per finished screen	Free to leave, maximum duration of 40 minutes
Piecerate-High Fix	€0.1 per finished screen	No outside option, fixed duration of 40 minutes
Piecerate-High Inet	€0.1 per finished screen	Internet allowed, fixed duration of 40 minutes
Piecerate-High Free	€0.1 per finished screen	Free to leave, maximum duration of 40 minutes
Bonus-Easy Fix	€5 after 50 finished screens	No outside option, fixed duration of 40 minutes
Bonus-Easy Inet	€5 after 50 finished screens	Internet allowed, fixed duration of 40 minutes
Bonus-Easy Free	€5 after 50 finished screens	Free to leave, maximum duration of 40 minutes
Bonus-Hard Fix	€10 after 100 finished screens	No outside option, fixed duration of 40 minutes
Bonus-Hard Inet	€10 after 100 finished screens	Internet allowed, fixed duration of 40 minutes
Bonus-Hard Free	€10 after 100 finished screens	Free to leave, maximum duration of 40 minutes

mension, we varied the incentives by implementing four different incentive schemes: two different piecerates (Low or High) and two different bonus schemes (Easy or Hard). In the piecerate treatments, subjects received a fixed payment for each successfully completed screen. In Piecerate-Low, subjects received €0.02 per finished screen; in Piecerate-High, €0.1. In the two bonus treatments, subjects received a bonus conditional on reaching a pre-specified target.[11] In Bonus-Easy, subjects received a €5 bonus if they reached the target of 50 screens. This is a relatively easy target that most subjects could, and in fact did, reach. In Bonus-Hard, subjects received a bonus of €10 if they reached the target of 100 screens. This target was deliberately set very high and only one subject managed to reach the target.[12] The size of the bonuses were chosen such that they equated the earnings of a subject in the high piecerate treatment with the same number of completed screens. Thus, a subject with 50 completed screens would earn the same in Bonus-Easy and Piecerate-High and a subject with 100 completed screens would earn the same in Bonus-Hard and Piecerate-High.

11. See Gill, Prowse, and Vlassopoulos (2013) for a previous real-effort slider experiment with fixed targets. The targets in our experiment were chosen based on a pilot session with the real-effort task. The session had the same structure as the treatment Piecerate-High Fix.

12. As the target is set deliberately high, intrinsic motivation is the main driver of effort provision in this treatment (see also the derivation of the hypotheses in 2.1). Another possible explanation could be overconfidence. However, this would require a substantial amount of uncertainty about one's own capabilities – which seems unlikely in our setup, since all subjects should have been able to learn about the difficulty of the task during the first stage of the experiment.

In the second treatment dimension, we manipulated the implicit costs by implementing three different work environments. First, in FIX, we implemented a fixed-time procedure, in which subjects had to stay at the computer for 40 minutes without any leisure alternatives offered.[13] We manipulated the implicit costs by implementing two environments with alternative activities for the subjects. In the INET-environment, subjects were allowed to use a web browser during the working phase of the experiment. Subjects had to remain in the laboratory for the whole time, but could surf the Internet instead of working on the task. This was implemented with a button on the real-effort screen, which would open a web browser and hide the real-effort task. Subjects could not work on the real-effort task and surf the Internet at the same time. However, they could always close the web browser and press a button to return to the real-effort task.[14] This allows us to record how much time subjects spent on the real-effort task and in the internet. In the treatment condition FREE, subjects could adjust their working time between 0 and 40 minutes by stopping to work on the real-effort task whenever they wanted. The screen in the working stage included a leave button. Pressing the button led to the questionnaire and subjects could then leave the cubicle to get their payments. Payments were made based on the number of finished screens at the time the subject stopped working.

The experiments were conducted at the BonnEconLab of the University of Bonn. They were implemented using z-Tree (Fischbacher, 2007) and subjects were recruited via hroot (Bock, Baetge, and Nicklisch, 2014). Upon arrival, subjects were seated in cubicles with curtains and blinds up to the ceiling, which prevented them from observing anything outside their cubicle. We conducted 16 regular sessions with the FIX- and INET-treatments and slightly adjusted the implementation in the FREE-treatments to prevent possible spillovers. In the FREE-treatments, subjects were invited to the lab on a given day, but could show up at any time between 10am and 4pm. This procedure ensured that subjects would not know the duration other subjects spend working on the task. In all treatments subjects received their payments individually in a separate room.

For each treatment, we gathered approximately 48 independent observations. In total, 571 subjects participated, with 58.6% of subjects being female and an average age of 23.58 years.[15] A session lasted on average 75 minutes for FIX and INET and individual sessions in FREE lasted between 20 and 75 minutes. All subjects received a show-up fee of €10 and additional earnings from the real-effort task. Subjects earned on average a total of €12.67, including the show-up fee. Between treatments,

13. The use of mobile phones was forbidden in all treatments.

14. More details on the implementation of INET and FREE can be found in Online Appendix G.

15. Neither age nor ability, as measured in the first stage, differ significantly between the three work environments ($p = 0.41$ and $p = 0.93$, both Kruskal-Wallis test). The gender composition differs slightly between treatments ($p = 0.099$, Kruskal-Wallis test). We use controls for gender, as well as age and ability, in our regression analyses to account for this.

earnings ranged from €10 in the Bonus-Hard-treatments to a maximum of €20.5 in the Piecerate-High Fix treatment.[16]

1.2.1 Behavioral Hypotheses

In a simple theoretical framework, the effort level would be chosen by maximizing

$$u(e) = \bar{w} + b(y) + I\delta(y) - c(e,i),$$

with a production technology $y = f(e)$, a fixed wage $\bar{w}$ (in our experiment the show-up fee), a performance-dependent payment $b(y)$ (either piecerate or bonus), intrinsic motivation $I\delta(y)$, and some costs depending on the explicit costs of effort e and the implicit costs i.[17] Implicit effort costs in our setup represent the foregone utility of not allocating the effort or time to other activities. Following the approaches by Murdock (2002) and James (2005), δ represents the agent's intrinsic motivation for the work (if she is intrinsically motivated) and I is an indicator function which is $I = 1$ if the agent is intrinsically motivated or $I = 0$ if not. In what follows, we present the intuition underlying our behavioral predictions and discuss the framework in more detail in the Online Appendix B.

With our work environment manipulation, which changes the implicit costs, we increase the marginal effort costs in Inet and Free compared to Fix. Furthermore, it seems reasonable to assume that subjects in Free have more outside opportunities than in Inet. This would imply an additional increase of the marginal effort costs in Free compared to Inet. Output should decrease as the marginal costs of effort increases. Thus, we expect the highest output in Fix (since implicit costs are low) and the lowest output in Free (since implicit costs are high).

Hypothesis 1. *We expect higher output in Fix than in Inet and Free. We also expect output in Inet to be higher than in Free.*

Let us now consider the differences between the piecerate treatments. Subjects provide effort as long as the marginal benefits from the piecerate payment and the intrinsic motivation to perform the task are higher than the marginal costs of effort. This point is reached sooner in Piecerate-Low than in Piecerate-High, due to the lower marginal benefits in Piecerate-Low, leading to higher outputs in the latter one. This holds for the comparison of all piecerate treatments within a work environment.

16. Table 1.B.1 in the Appendix presents implicit hourly wages per treatment.

17. For example, this includes versions of $c(e,i)$ like in Koch and Nafziger (2016), where $c(e,i) = i \cdot \tilde{c}(e)$, and where the parameter i differs between work environments, i.e., effort costs increase when alternative actions are present. Therefore $i_{Fix} \leq i_{Inet} \leq i_{Free}$. This could also incorporate a version of $c(e,i)$, where implicit costs are modeled as utility of leisure, but leisure is negatively related with effort, i.e., time, as in (Corgnet, Hernán-González, and Schniter, 2015)

Hypothesis 2. *We expect higher output in Piecerate-High than in Piecerate-Low.*

For Bonus-Easy we would expect only few outputs above 50 as $\frac{\partial b(y)}{\partial y} = 0$ for any output above 50. Additional output would only be driven by workers for whom the marginal intrinsic motivation would still be higher than the marginal costs. In Bonus-Hard, we would expect very low output in general, since subjects should realize very early on in the experiment that they will not reach the target of 100 and thus marginal (monetary) benefits equal zero for all feasible outputs. Consequently, output would again only be driven by workers for whom the marginal intrinsic motivation is higher than the marginal costs.

The differences between Bonus-Easy and the two piecerate treatments are ultimately an empirical question, because predictions about performance differences would require additional assumptions about the exact form of the cost of effort function and the intrinsic motivation. However, since reaching the target of 100 screens in Bonus-Hard is not feasible, we can predict that in both piecerate treatments output should be higher than in Bonus-Hard. This is due to the fact that monetary incentives are basically absent in Bonus-Hard and therefore incentives are higher in the two piecerate treatments.

Hypothesis 3. *We expect higher output in Bonus-Easy than in Bonus-Hard. The output in Bonus-Easy should be 50 screens or slightly above. Furthermore, we expect higher output in both piecerate treatments than in Bonus-Hard.*

1.3 Results

Table 1.2 provides summary statistics for the output in all treatments. In the following, we will first demonstrate that implicit costs have a significant impact on work output and discuss their influence within an incentive scheme. Thereafter, we will demonstrate that implicit costs influence the comparisons between incentive contracts. Finally, we will take a closer look at the usage of the offered outside option and the influence of non-cognitive traits on behavior. If not stated otherwise, reported p-values are two-sided and based on t-tests and regressions.[18]

1.3.1 The Impact of Implicit Costs on Output

We start by looking at the general effect of implicit costs for all incentive schemes. Based on the raw means reported in Table 1.2, output in the Fix-treatments is on average 15.7% higher than in the Inet-treatments and 35.6% higher than in the Free-treatments. The average output in the Inet-treatments is 17.3% higher than in the Free-treatments. The predicted output of our three work environments is presented in Figure 1.1. The figure is based on a least squares regression with controls

18. Additionally, Tables 1.B.4 and 1.B.5 in the Appendix report p-values of non-parametric tests.

Table 1.2. Summary Statistics of Outputs

		Over All	By Incentive Scheme			
		Incentives	Piecerate-Low	Piecerate-High	Bonus-Easy	Bonus-Hard
Fix	Mean	58.79	57.56	59.40	60.09	58.21
	SD	14.44	14.59	13.59	12.49	16.91
	N	188	48	47	45	48
Inet	Mean	50.82	41.87	53.69	53.19	54.43
	SD	19.90	23.47	16.50	15.50	21.06
	N	190	47	48	48	47
Free	Mean	43.34	36.21	52.35	45.94	38.69
	SD	24.42	27.18	21.02	20.41	25.79
	N	193	48	49	48	48

Notes: SD: standard deviation, N: number of independent observations

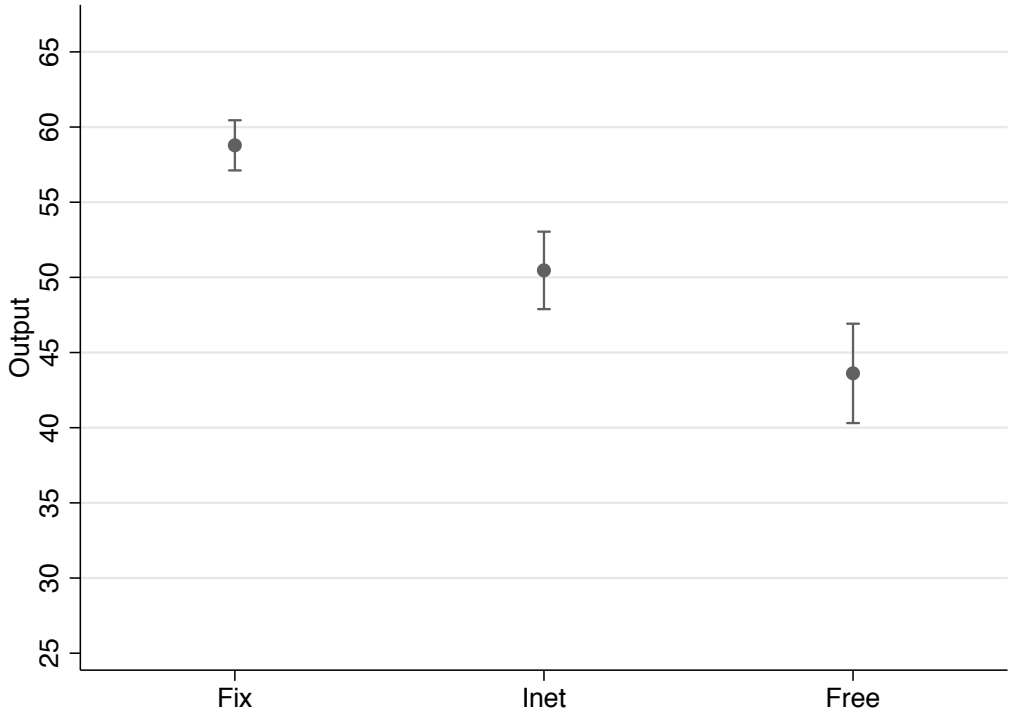


Figure 1.1. Predicted Output with 95% CIs Across Work Environments

Notes: Estimates are based on linear regression controlling for subjects' ability, gender, and age. Plot shows the margins with confidence intervals. For results and coefficients of the corresponding regressions, see Table 1.B.2 in the Appendix.

for ability, gender, and age.[19] In general, Figure 1.1 shows that output decreases significantly with higher implicit costs (all pairwise comparisons $p < 0.01$). Already based on this general look at the data we can conclude that implicit costs in general influence the output negatively, which is in line with our predictions.

However, the impact of implicit costs is not limited to the average outputs; Table 1.2 and Figure 1.1 reveal that implicit costs increase the variance of the output, too. The variance differs significantly between work environments and increases with opportunity costs (all $p < 0.01$, using two-sided Variance-ratio tests). The lowest variance is observed in Fix, increases in Inet, and is highest in Free.[20] Our treatments increase the implicit costs by manipulating outside options and the time spent on the outside option reduces the output. However, not all subjects utilize the outside options to the same extent, which increases the variance. In fact, the largest part of the observed variance is explained by the total time spent working on the task (see Section 1.3.4).[21]

Result 1. *In line with Hypothesis 1, subjects' output decreases significantly as implicit effort costs increase. At the same time, the variance of output increases with implicit costs.*

In the following, we turn to differences within each incentive scheme. Table 1.3 estimates the treatment effects for each incentive scheme using Fix as the benchmark category. In Piecerate-Low, we observe a decline of effort between Fix and both Inet and Free. Output decreases when subjects face increased implicit effort costs, compared to Fix. With added controls, average output significantly decreases by 12.55 screens in Inet and by 17.28 screens in Free (both coefficients with $p < 0.01$). Average output in Piecerate-Low Free is lower than in Piecerate-Low Inet, but this difference turns out to be insignificant (p=0.28).

A similar pattern emerges for Piecerate-High. Compared to Fix, output decreases by 6.15 screens in Inet ($p < 0.05$) and by 8.23 screens in Free ($p < 0.01$). However, the two coefficients are smaller than their counterparts in Piecerate-Low. Again, output is the lowest in Free, but Fix and Inet do not differ significantly from each other ($p = 0.55$). Our results are therefore in line with the first part of Hypothesis 1, but not with the second part that Free induces higher implicit costs than Inet.[22]

19. The corresponding regression table is presented in Table 1.B.2 in the Appendix. Furthermore, Figure 1.B.1 in the Appendix displays the boxplots for the output level for each incentive scheme.

20. This pattern generally holds for each incentive scheme individually.

21. A more detailed analysis of the time worked on the task can be found in Online Appendix D. Table 6 in Online Appendix D demonstrates that the time subjects work on the task changes analogously to the changes in output presented here.

22. In the Online Appendix C, we parameterize our theoretical framework and estimate a structural model. Table 3 in the Online Appendix C presents the results of this exercise. The estimation results support the notion that the presence of outside options change the implicit effort costs. De-

Table 1.3. Regression of Output on Work Environments

	Piecerate-Low		Piecerate-High		Bonus-Easy		Bonus-Hard	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Inet	-15.69***	-12.55***	-5.72*	-6.15**	-6.90**	-7.41***	-3.78	-3.85
	(4.02)	(3.37)	(3.10)	(2.65)	(2.91)	(2.47)	(3.92)	(3.43)
Free	-21.35***	-17.28***	-7.06*	-8.23***	-14.15***	-14.87***	-19.52***	-17.23***
	(4.45)	(4.25)	(3.60)	(3.01)	(3.48)	(3.37)	(4.45)	(4.42)
Constant	57.56***	1.30	59.40***	35.10***	60.09***	47.44***	58.21***	28.62**
	(2.11)	(12.28)	(1.98)	(6.99)	(1.86)	(13.04)	(2.44)	(11.14)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
N	143	142	144	142	141	141	143	143
R^2	.14	.35	.031	.31	.11	.23	.14	.26
(Inet and Free = 0)	0.00	0.00	0.07	0.01	0.00	0.00	0.00	0.00
(Inet = Free)	0.28	0.33	0.73	0.55	0.05	0.04	0.00	0.00

Notes: Table presents least squares regression using performance as dependent variable. * p < 0.1, ** p < 0.05, *** p < 0.01. Robust Standard errors in parentheses. For some observations controls are missing, due to some subjects who refused to answer some of the sociodemographic questions. Controls: ability, age, gender.

Result 2. *Implicit costs significantly reduce the performance in the incentive schemes Piecerate-High and Piecerate-Low. Differences between Inet and Free exist, but do not turn out to be significant.*

Similar to the two piecerate treatments, and in line with our predictions, we observe a decline of output for both bonus-based incentive schemes with increased implicit costs. For Bonus-Easy, the output decreases by 7.41 screens in Inet ($p < 0.01$) and 14.87 screens in Free ($p < 0.01$) compared to Fix. The output in Bonus-Easy Inet is 15.6 % higher compared to Free and is also significantly different ($p = 0.04$). As Figure 1.B.1 in the Appendix shows, the output distribution in Bonus-Easy Free collapses around 50 screens. In line with Hypothesis 3, the majority of subjects stop working once they reached the threshold for the bonus.[23] Only few subjects worked more than necessary and some subjects stopped early, performing poorly. Interestingly, this sharp decline in effort provision beyond 50 cannot be observed in Bonus-Easy Inet. However, those differences in outputs do not translate into significant differences in the number of subjects who earned the bonus. In Bonus-Easy Fix, 88.89% of subjects reached the target of 50 screens; in Inet, 85.42%; and in Free, 77.08% (all pairwise comparisons $p > 0.108$ or above, two-sided Fisher's exact test).

pending on the exact parametrization, the marginal implicit effort costs are estimated between 8.4 and 17.8 cents for Inet and between 10.1 and 18.3 cents for Free.

23. Again, the different outputs result from different durations spent working on the task. Refer to Online Appendix–D for an additional analysis.

In Bonus-Hard, we observe the same pattern in output levels. In Bonus-Hard Inet, the average output does not differ significantly from the average output in Fix. In Bonus-Hard Free, average output is 17.23 screens lower than in Fix ($p < 0.01$) and 13.38 screens lower than in Inet ($p < 0.01$). The output distribution in Bonus-Hard Free is shifted downwards and has a longer lower tail (again, compare Figure 1.B.1 in the Appendix). This is mostly driven by subjects who stop working and leave the lab early. Only one subject in the Bonus-Hard-treatments was able to reach the target of 100 screens.[24]

Our results show that for both bonus based incentive schemes output decreases in Free compared to Inet. This is in line with our hypothesis that implicit costs are increasing in Free resulting in lower output. Interestingly, more subjects work very short times on the task and produce very low outputs in Free. In fact, in Bonus-Easy, the number of subjects who work less than 5 minutes increases from one in Inet to seven in Free. Similarly, in Bonus-Hard this number increases from one to seven. Thus, more subjects exert very low levels of effort.[25]

Result 3. *In Bonus-Easy, the two treatments with increased implicit costs, Inet and Free, result in significantly lower output than Fix. In Bonus-Hard subjects produce significantly lower outputs in Free compared to Fix. Unlike the two piecerate treatments, outputs differ significantly between Inet and Free in both bonus treatments.*

1.3.2 Implicit Costs and the Comparison between Incentive Schemes

So far we have demonstrated that implicit costs can influence the output even if marginal monetary incentives are fixed. In a next step, we will demonstrate that implicit costs influence the comparison of incentive schemes. The descriptive statistics are provided in Table 1.2, while Figure 1.2 presents the estimated output in all treatments after controlling for ability, gender, and age. The figure is based on the estimation results presented in Table 1.B.3 of the Appendix.[26]

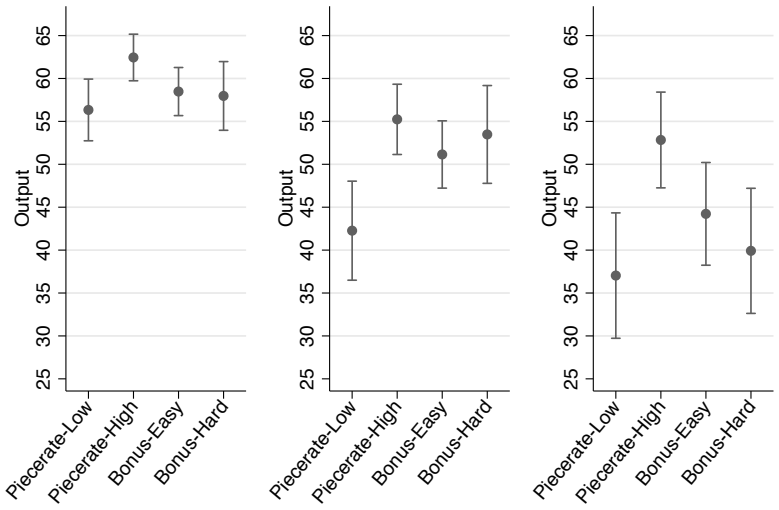
In the Fix-treatments, the highest output is observed in Piecerate-High and the lowest output in Piecerate-Low.[27] Output in Piecerate-High is significantly

24. We use the estimated parameters of the structural model in the Online Appendix C to derive predictions about the two discretionary bonus treatments (see Table 4 in Online Appendix C). Again, the results of the structural model support the analyses presented in this section.

25. These findings are also reflected in subjects' outputs. In Bonus-Easy Inet (Bonus-Hard Inet) two (one) subjects have an output below ten; this number increases to seven (nine) in Bonus-Easy Free (Bonus-Hard Free). One possible explanation is that increasing costs are more likely to trigger the theoretical corner solution of the two bonus treatments.

26. We estimate one regression per work environment controlling for ability, age and gender. All reported p-values in this part of the analysis are based on these regressions.

27. It is worth pointing out that without additional controls we fail to identify any significant differences between the treatments in Fix.



(a) Fix **(b)** Inet **(c)** Free

Figure 1.2. Predicted Output with 95% CIs Across Incentive Schemes

Notes: Estimates are based on linear regressions controlling for subjects' ability, gender, and age. Plot shows the margins with confidence intervals. For results and coefficients of the corresponding regressions, see Table 1.B.3 in the Appendix.

higher than in Piecerate-Low ($p = 0.012$), Bonus-Easy ($p = 0.048$), and Bonus-Hard ($p = 0.078$).[28] All other comparisons are insignificant ($p \geq 0.35$). Comparing output across incentive schemes for Inet, we observe again the highest output in Piecerate-High and the lowest in Piecerate-Low. Output in Piecerate-Low turns out to be significantly lower than in the other three incentive schemes (all $p < 0.001$). Yet, all other comparisons remain insignificant ($p \geq 0.17$). If we compare the incentive schemes within the working environment Free, we again observe the highest output in Piecerate-High and the lowest in Piecerate-Low. In Free, output in Piecerate-High is again significantly higher than in Piecerate-Low ($p = 0.001$), Bonus-Easy ($p = 0.04$), and Bonus-Hard ($p = 0.005$). Outputs in Piecerate-Low, Bonus-Easy, and Bonus-Hard do not differ significantly ($p > .12$ for all pairwise comparisons).

Comparing these reported differences across work environments provides additional insights into the impact of implicit costs. Increasing the implicit costs, we observe stronger negative reactions for Piecerate-Low than for Piecerate-High. This influences the comparison between the two incentive schemes. In Fix, average output in Piecerate-High is only 11% higher than in Piecerate-Low; in Inet it

28. All p-values in this part of the analysis are based on the regressions in Table 1.B.3 in the Appendix using the specifications with control variables. Figure 1.2 is based on the same regression table.

is 31% higher; and in Free, it is even 43% higher.[29] In both Inet and Free, these changes are significantly larger than in Fix (both $p < 0.05$).

Output in the two bonus schemes responds to increased implicit costs. However, implicit costs do not significantly influence the comparison between Bonus-Easy and Bonus-Hard. In all three environments, Fix, Inet, and Free, we observe no significant differences between the outputs in Bonus-Easy and Bonus-Hard (all $p \geq 0.38$). However, comparing bonus schemes with piecerates shows that the outputs respond differently to changes in implicit costs (see Figure 1.2).

Average output under Piecerate-Low decreases most strongly with the introduction of implicit costs (moving from Fix to Inet), while in the bonus schemes the response tends to be stronger if implicit costs increase further (moving from Inet to Free). This influences the comparison between the piecerate schemes and the bonus schemes. Without implicit costs, output does not differ significantly between Piecerate-Low, Bonus-Easy, and Bonus-Hard (all pairwise comparisons $p \geq 0.37$). However, the steep decline in Piecerate-Low Inet results in significantly lower output compared to Bonus-Easy Inet ($p < 0.01$) and Bonus-Hard Inet ($p < 0.01$). Yet, after the output declines more steeply in Bonus-Easy Free and Bonus-Hard Free, outputs are no longer significantly different between the two bonus schemes and Piecerate-Low Free (both $p \geq .13$).

With the introduction of implicit costs, average output also decreases under Piecerate-High, but not as strongly as under Piecerate-Low. At the same time, increased implicit costs increase the variance. Thus, we observe quite the opposite picture when comparing Piecerate-High with Bonus-Easy and Bonus-Hard. While in Fix output is significantly higher in Piecerate-High than in Bonus-Easy and Bonus-Hard (both $p < 0.05$), they no longer differ significantly in Inet. Only after average output in the two bonus schemes declines steeply in Free, is output in Piecerate-High significantly higher than output in Bonus-Easy ($p < 0.05$) and Bonus-Hard ($p < 0.01$). To summarize, we find partial support for Hypothesis 3, but output in Bonus-Hard tends to be higher than expected.

Result 4. *Of all incentive schemes, Piecerate-High responds least to changes in the implicit cost; in contrast we observe a strong response in Piecerate-Low. Therefore, the difference in outputs between the two piecerate schemes increases with implicit costs. The comparison between the bonus- and piecerate-schemes depends on the exact setting.*

1.3.3 Elasticity of Output

A different way to investigate the reaction to changed incentives is to calculate the elasticity of the output in all three work environments with regard to the piecerates. In Piecerate-High, marginal incentives are higher for each additionally produced

29. Percentages are based on the predictive margins presented in Figure 1.2, which are based on the regressions in Table 1.B.3.

Table 1.4. Elasticities

	Fix		Inet		Free	
	(1)	(2)	(3)	(4)	(5)	(6)
ln(Piecerate)	0.0307	0.0841*	0.2577***	0.2722***	0.5272***	0.5231***
	(0.0386)	(0.0438)	(0.0973)	(0.0896)	(0.1378)	(0.1352)
Constant	4.1269***	3.5550***	4.5126***	3.8518***	5.0825***	3.1893***
	(0.1136)	(0.1401)	(0.2518)	(0.6359)	(0.3539)	(0.5967)
Controls	No	Yes	No	Yes	No	Yes
N	95	95	93	92	95	93
R^2	.0067	.34	.075	.25	.14	.23

Notes: Table presents least squares regression using the logarithm of performance as dependent variable. * p < 0.1, ** p < 0.05, *** p < 0.01. Robust Standard errors in parentheses. Observations with a performance of zero are dropped from the estimation. Controls: ability, age, gender.

screen than in Piecerate-Low. Thus, from a pure incentive theory perspective, we would on average expect higher outputs in Piecerate-High than in Piecerate-Low, i.e., a positive output elasticity. Table 1.4 gives the resulting elasticities when regressing the logarithm of the piecerate on the logarithm of the output. For Fix, we observe an elasticity close to zero that turns significant only after adding controls. For both, Inet and Free, we observe significantly larger and positive elasticities compared to Fix ($p = 0.0287$ and $p < 0.01$, two-sided, Wald test). Increasing the piecerate by 1% would increase the outputs by 0.25% in Inet and by 0.52% in Free. However, the difference between these two elasticities falls short of reaching conventional levels of significance ($p = 0.107$).

These results show that implicit costs induced by different work environments matter. While subjects' output only responds marginally to increased incentives in Fix, we are able to observe significant reactions in outputs to increased incentives in both work environments with higher implicit costs. While the response in Inet and Free is positive, it is still inelastic.

Result 5. *The elasticity of the output increases with implicit costs. In Fix, the elasticity of the output differs only weakly significantly from zero. By contrast, in both Inet and Free we observe a positive and significant response of the output to increased incentives.*

1.3.4 Supplementary Analyses

We observe a high variance in performance across all incentive schemes and work environments, especially in those with an outside option, i.e., Inet and Free. Therefore, in the following we take a closer look at individual characteristics and personality traits and their impact in the different work environments. This will help us to understand to what extent the observed variance is driven by those characteristics.

Table 1.5. Determinants of Output

	(1)	(2)	(3)	(4)	(5)
Fix x Ability	6.1241***			6.3441***	6.3441***
	(0.4908)			(0.5681)	(0.5686)
Inet x Ability	5.1249***			5.1377***	4.2718***
	(0.8575)			(0.8900)	(0.5149)
Free x Ability	4.1733***			4.4898***	3.9106***
	(1.0097)			(1.0956)	(0.4158)
Fix x CRT score		1.1676		-0.1180	-0.1180
		(0.8405)		(0.6849)	(0.6855)
Inet x CRT score		2.6692**		1.7756	0.8935
		(1.2194)		(1.1584)	(0.5818)
Free x CRT score		-0.8770		-2.1454	0.4436
		(1.5251)		(1.6693)	(0.6526)
Fix x Conscientiousness			-0.4496	0.0290	0.0290
			(0.6417)	(0.4674)	(0.4679)
Inet x Conscientiousness			1.3723*	1.1101	-0.3025
			(0.8124)	(0.7738)	(0.4219)
Free x Conscientiousness			0.6093	0.0464	-0.2763
			(0.9522)	(1.0703)	(0.4275)
Total time					1.5185***
					(0.0325)
Treatment FE	Yes	Yes	Yes	Yes	Yes
Gender and Age	No	No	No	Yes	Yes
Big 5 (w/o Cons), Risk	No	No	No	Yes	Yes
N	569	571	571	568	568
R^2	.29	.15	.15	.32	.8

Notes: Table presents least squares regression using output as dependent variable. * p < 0.1, ** p < 0.05, *** p < 0.01. Robust Standard errors in parentheses. Big 5 (w/o Cons.) controls for the other Big 5 traits and risk for general risk attitudes all interacted with a variable indicating the work environment.

We therefore regress output on characteristics interacted with an indicator for each environment in Table 1.5. All reported models include treatment fixed effects.

In all implicit cost settings, the output and the ability measures have a significant positive relationship, but the strength differs (see Model 1). For each finished screen in the ability stage, subjects are estimated to complete slightly more than 6 screens in the main experiment in the Fix-treatments, 5 screens in the Inet-treatments, and 4 screens in the Free-treatments. The influence of an agent's ability on output

differs significantly between FIX and FREE (p-value = 0.082, two-sided, Wald-test), but not for any other comparison (all p-values> .31, two-sided, Wald-test).[30]

Furthermore, we can explore the relationship between personal characteristics and output. Along with gender and age, we also elicited general risk attitudes, personality traits, and cognitive ability. We elicited the general risk attitudes (Dohmen et al., 2011) as the two BONUS-treatments involve the risk of investing effort without reaching the target. To elicit personality traits, we administered the 10-item version of the Big 5 (Rammstedt and John, 2007). Of the personality traits, we are particularly interested in the effect of conscientiousness, which has been linked to increased job performance (e.g., Barrick and Mount, 1991).[31] In fact, research on non-cognitive skills suggests that conscientiousness predicts educational attainment and labor market outcomes as strongly as cognitive ability (Heckman and Kautz, 2012). As an additional measure, we implemented the CRT, a cognitive reflection test by Frederick (2005). A recent paper by Corgnet, Hernán-González, and Mateo (2015) shows for a setting similar to our INET-environment that higher cognitive reflection reduces leisure activities. For strategic interactions, Gill and Prowse (2017) find neither a correlation between response times and personality nor between response times and cognitive ability.

We find a positive relationship between the CRT score and output (Model 2) in INET, but not in the other environments. Similarly, conscientiousness (Model 3) is positively associated with output in INET, but not in the other environments. In Model 4, we simultaneously control for all measures. As several of the measures are correlated with each other, only the influence of ability remains significant. In a last step, we additionally include the total time (in minutes) worked on the task (Model 5). This increases the explained variance dramatically as the R^2 improves from .32 in Model 4 to .80 in Model 5. Obviously, the time subjects worked on the task explains the largest part of the variance in our data.[32]

In fact, the reported differences in outputs result from a substantial fraction of subjects using the outside options when available: 36.84% in INET and 47.67% in FREE. In both work environments, the usage of the outside option reduces the time subjects spent working on the task. In INET and FREE, the average working time is significantly below 40 minutes (both $p < 0.01$). Subjects work on the task on average 35.12 minutes in INET and only 29.18 minutes in FREE ($p < 0.01$). Thus,

30. Recall that ability did not differ significantly between treatments.

31. The American Psychology Association defines conscientiousness as "the tendency to be organized, responsible, and hardworking".

32. In the Online Appendix E, we also repeat the analysis from above and use questionnaire data to explore the influence of personality measures on the time subjects spent working on the task. The direction of the point estimates is in line with our results in Table 1.5.

our treatments influence more the extensive margin (time spent working on task) than the intensive margin (speed while working on task).[33]

We conclude our last result:

Result 6. *Conscientiousness and CRT are significantly correlated with higher output in* INET*. The time subjects spent working on the task explains a large part of the observed variance in output. Subjects time on the task decreases significantly as implicit effort costs increase.*

1.4 Discussion and Conclusion

In this paper, we investigate how work environments with different implicit costs influence the effectiveness of linear and non-linear incentive schemes. We exogenously vary the implicit effort costs between work environments by offering real-leisure alternatives and comparing the performance of subjects in two piecerates and two bonus schemes. We observe that incentive contracts and opportunity costs interact in a non-trivial manner. Generally, as implicit costs increase, the average output decreases and the variance of output increases. Yet, the responses are not equally strong for all incentive schemes. We observe stronger negative reactions for PIECERATE-LOW than for PIECERATE-HIGH. These unequal reactions lead to increasing differences among those two incentive schemes: in FIX, average output under the high piecerate is 11% higher than under the low piecerate; in INET, it is 31% higher; and in FREE, even 43% higher. Likewise, an increase in implicit costs increases the output elasticity of piecerates. With respect to non-linear incentive schemes, our results suggest that the effect of bonus schemes depends on the opportunities of workers to allocate their time. Our results in BONUS-EASY suggest that achievable targets induce behavior such that targets are closely matched, but not exceeded, in those work environments with substantial implicit effort costs. For targets like BONUS-HARD, implicit costs increase the number of workers who drop out of the task once they realize that the target is hard to achieve. However, in the FIX-environment with low implicit costs, we observe, for both bonus schemes, an effort that is far from any incentivized points – either beyond the target or far before the target is reached. This behavior might be more in line with subjects who consider this a fixed wage setting than a bonus setting. Although this might be unexpected, it is similar to the fixed bonus treatments reported in DellaVigna and Pope (2018), where experts also fail to forecast the effort provision beyond an incentivized point. Moreover, monetary incentives differ strongly between PIECERATE-HIGH and BONUS-HARD; yet, only in FREE we do observe a large and highly significant difference between the two incentive schemes. Thus, our results in general show the dependency of the effectiveness

33. In the Online Appendix D, we also show that, once we account for the time subjects work on the task, output is similar across all environments.

of incentive schemes with respect to the work environment, i.e., the implicit effort costs. For example, workers in bonus schemes might be less sensitive to incentives in environments similar to Fix and Inet. Behavior in those environments might not be well predicted by standard incentive theory. However, as the implicit costs of effort increase, the behavior aligns more and more with predictions made by incentive theory.

In addition to providing new insights into the interplay of piecerates and bonus schemes with implicit effort costs, our paper also confirms and qualifies previous findings in the literature. In general, increasing the implicit costs of effort, while keeping the incentives (wages or piecerates) fixed, leads to smaller output (Corgnet, Hernán-González, and Schniter, 2015; Koch and Nafziger, 2016). Yet, a superficial look might suggest that our results are not fully in line with Corgnet, Hernán-González, and Schniter (2015). While Corgnet, Hernán-González, and Schniter (2015) find no significant impact of the option to surf the Internet under a piecerate contract, we observe significant differences between Fix and Inet under two piecerate contracts. However, Corgnet, Hernán-González, and Schniter (2015) also report a 10% smaller output if the option to surf the Internet is available. In our most comparable treatments, Piecerate-High Fix and Piecerate-High Inet, we observe the same drop of output by 10%.[34] Furthermore, their results, over time, demonstrate significant differences in the later part of their experiment. Figure 1.B.2 in the Appendix demonstrates that the comparison of output in Piecerate-High Fix and Piecerate-High Inet in our paper follows the same dynamics. For Piecerate-High, we initially observe no significant difference, but over time a pronounced difference develops between the outputs in Fix and Inet. Thus, we confirm the finding by Corgnet, Hernán-González, and Schniter (2015) that implicit cost effects under high piecerates are dynamic and need some time to develop, although in our setting these effects are strong enough to result in overall significant differences. Furthermore, we demonstrate that the elasticity of the output increases with implicit effort costs. Moreover, our results show that implicit effort costs and the exact nature of those might be even more important for bonus-based incentive contracts. Due to the existence – and depending on the exact size – of implicit costs, subjects might either explicitly target the bonus or abstain from working completely. Our paper also contributes to the ongoing discussion of the real-effort slider task and real-effort experiments in general. Araujo et al. (2016) implement the slider task in a fixed laboratory environment with three different piecerate schemes and conclude that it demonstrates no meaningful response to explicit monetary incentives. We show that this is more a problem of the fixed laboratory environment than the slider task itself. Our estimated output elasticity of 0.0307 in Fix is very similar to the elastic-

34. Comparability is based on the available outside option and implicit hourly wages. The implicit wage is calculated by using the performance-dependent pay component (payoff without show-up fee) and scaling it up/down to an hourly wage.

ity of 0.025 estimated by Araujo et al. (2016). Yet, once implicit costs are increased, subjects respond in a meaningful and significant way to the linear incentives. Thus, effort in any (real-effort) task should not be evaluated independently of the work environment it is implemented in.

Our results have implications for the use and design of incentive schemes within organizations. The management has many means to affect worker behavior and every aspect of an organization can be used as a parameter to obtain desired outcomes (Roberts, 2007). In addition to monetary and non-monetary incentives, organizations should recognize that they might want to adjust implicit costs as a relevant parameter, too. For example, to increase the output in our Piecerate-Low Inet setting, one could either implement a higher piecerate (Piecerate-High Inet) or keep the piecerate fixed and reduce the implicit costs (Piecerate-Low Fix). Using our experimental results, the first approach would, on average, increase output by roughly 29% with additional costs of € 4.56 per worker, and the second approach would increase the output by roughly 40% with additional costs of only € 0.34. Even if the management cannot change the work environment, it is important to take implicit costs into account when implementing and evaluating traditional incentive schemes. Our analysis of output elasticities would suggest that there are only minor benefits from increased piecerates in environments similar to our Fix-treatments, but larger gains from changes in the piecerate in environments similar to our Free-treatments. Similarly to managers in firms, unemployment agencies want to incentivize job seekers to find a job.[35] Job seekers have to seek their jobs in an environment where leisure costs are potentially high and leisure alternatives are easily available and always present. Unemployment agencies therefore could use a Fix environment, for example by requiring job seekers to spend a fixed amount of time in a room with access to material needed for applications but no leisure alternatives. Beyond the analysis of incentive schemes, our results and implications are also interesting in light of the recent discussion of workplace flexibility and home offices. Given our results it is not surprising that, after the boom of telecommuting in the last decade, companies like IBM are now adopting more restrictive approaches to home office and telecommuting and either demand full presence or at least required presence times.[36] Other firms, for example call centers, incentivize their flexible workers to work specific hours, using contracts which yield bonuses for making calls for a given time in the evening hours.[37] Our paper demonstrates that reducing implicit costs and temptations like surfing the internet leads to higher productivity. Yet, some caution is warranted as workers might realize that the work environment is an active choice by

35. Instead of relying on incentive contracts, unemployment agencies rely, for example, on binding job search requirements (e.g., Arni and Schiprowski, 2017).

36. See https://www.bloomberg.com/news/articles/2017-07-10/the-rise-and-fall-of-working-from-home.

37. One example is Infas in Bonn. Information is provided on their website: www.infas.de

the management and introduce reciprocal motives. As such, the active choice of inflexible work environments, which reduce implicit effort costs, might signal distrust and reduce motivation and output of the worker (Alder, Noel, and Ambrose, 2006; Corgnet, Hernán-González, and McCarter, 2015; Koch and Nafziger, 2016). More generally, controlling the own work environment can influence workers motivation (Deci, Connell, and Ryan, 1989; Deci and Ryan, 1995) and change the performance of individuals (Kiessling, Radbruch, and Schaube, 2018).

Future work on implicit costs in work environments should extend to non-monotone tasks that require creativity, communication, and innovation. Apple, Unilever, and Facebook are just a few examples of firms that use architecture to design work environments encouraging communication and serendipitous encounters through coffee places and meeting points.[38] While these new work environments are intended to increase innovation and creativity, they also increase the implicit costs of effort. Investigating the net effect in such environments seems to be an important next step.

38. Accessible introductions to this topic are provided by Wagner and Watch (2017) and Waber, Magnolfi, and Lindsay. (2014).

Appendix 1.A Screenshot

Figure 1.A.1. Screenshot of Real-Effort Screen

Appendix 1.B Additional Figures and Tables

Table 1.B.1. Implicit (hourly) Wage in Euro

	Piecerate-Low	Piecerate-High	Bonus-Easy	Bonus-Hard
Fix	1.73	8.91	6.67	0.00
Inet	1.26	8.05	6.41	0.32[a]
Free	1.62	8.56	6.95	0.00

Notes: Implicit wage is calculated by using the performance-dependent pay component (payoff without show-up fee) and scaling it up to an hourly wage. Surf time is working time, i.e., in Inet and Fix working time is fixed to 40 minutes. In Free, subjects can work less than 40 minutes.
[a] One subject achieved the target of 100. Without this subject, the implicit wage is 0.00.

Table 1.B.2. Regression of Output on Work Environments

	Over All Incentives	
	(1)	(2)
Inet	-7.97***	-8.32***
	(2.06)	(1.88)
Free	-15.45***	-15.17***
	(2.05)	(1.87)
Constant	58.79***	26.40***
	(1.46)	(5.28)
Controls	No	Yes
N	571	568
R^2	.091	.24
(Inet and Free = 0)	0.00	0.00
(Inet = Free)	0.00	0.00

Notes: Table presents least squares regression using output as dependent variable. * p < 0.1, ** p < 0.05, *** p < 0.01. Robust Standard errors in parentheses. For some observations controls are missing, due to some subjects who refused to answer some of the sociodemographic questions. Controls: ability, age, gender.

Table 1.B.3. Performance Differences within Work-Environments

	Fix		Inet		Free	
	(1)	(2)	(3)	(4)	(5)	(6)
Piecerate-Low	-1.84	-6.11**	-11.82***	-12.97***	-16.14***	-15.80***
	(2.89)	(2.40)	(4.17)	(3.63)	(4.94)	(4.67)
Bonus-Easy	0.68	-3.97**	-0.50	-4.09	-6.41	-8.61**
	(2.72)	(2.00)	(3.27)	(2.97)	(4.21)	(4.17)
Bonus-Hard	-1.20	-4.48*	0.74	-1.76	-13.66***	-12.92***
	(3.14)	(2.53)	(3.89)	(3.57)	(4.78)	(4.58)
Constant	59.40***	23.86***	53.69***	24.65***	52.35***	25.95**
	(1.98)	(5.13)	(2.38)	(9.31)	(3.00)	(10.48)
Controls	No	Yes	No	Yes	No	Yes
N	188	188	190	189	193	191
R^2	.0047	.4	.067	.26	.068	.15
(PR-L vs Bonus-Easy)	0.37	0.35	0.01	0.01	0.05	0.14
(PR-L vs Bonus-Hard)	0.84	0.54	0.01	0.01	0.65	0.58
(Bonus-Easy vs Bonus-Hard)	0.54	0.84	0.74	0.51	0.13	0.38
Joint test of all vs. PR-H	0.81	0.06	0.02	0.00	0.00	0.00

Notes: Table presents least squares regression using Performance as dependent variable. * p < 0.1, ** p < 0.05, *** p < 0.01. Robust Standard errors in parentheses. For some observations controls are missing, due to subjects not answering the sociodemographic questions. Controls: ability, age, gender.

Table 1.B.4. Mann-Whitney Test for Comparisons between Work Environment

	Overall Fix	Piecerate-Low Fix	Piecerate-High Fix	Bonus-Easy Fix	Bonus-Hard Fix
Inet	0.0000	0.0010	0.0580	0.0068	0.3578
Free	0.0000	0.0001	0.1614	0.0001	0.0002
(Inet = Free)	0.0052	0.3422	0.9511	0.0372	0.0052

Notes: Table shows p-values of a Mann-Whitney u-test, which tests whether two independent samples were selected from populations with same distributions. Table is built analog to the tests of the regression coefficients in the corresponding regression tables.

Table 1.B.5. Mann-Whitney U-test for Comparison of Incentive Schemes

	Fix	Inet	Free
	Piecerate-High		
Piecerate-Low	0.5866	0.0311	0.0050
Bonus-Easy	0.9191	0.5447	0.2055
Bonus-Hard	0.9644	0.3960	0.0176
(PR-L vs Bonus-Easy)	0.5231	0.0749	0.2594
(PR-L vs Bonus-Hard)	0.9644	0.3960	0.0176
(Bonus-Easy vs Bonus-Hard)	0.9571	0.1727	0.4343

Notes: Table shows p-values of a Mann-Whitney u-test, which tests whether two independent samples were selected from populations with same distributions. Table is built analog to the tests of the regression coefficients in the corresponding regression tables.

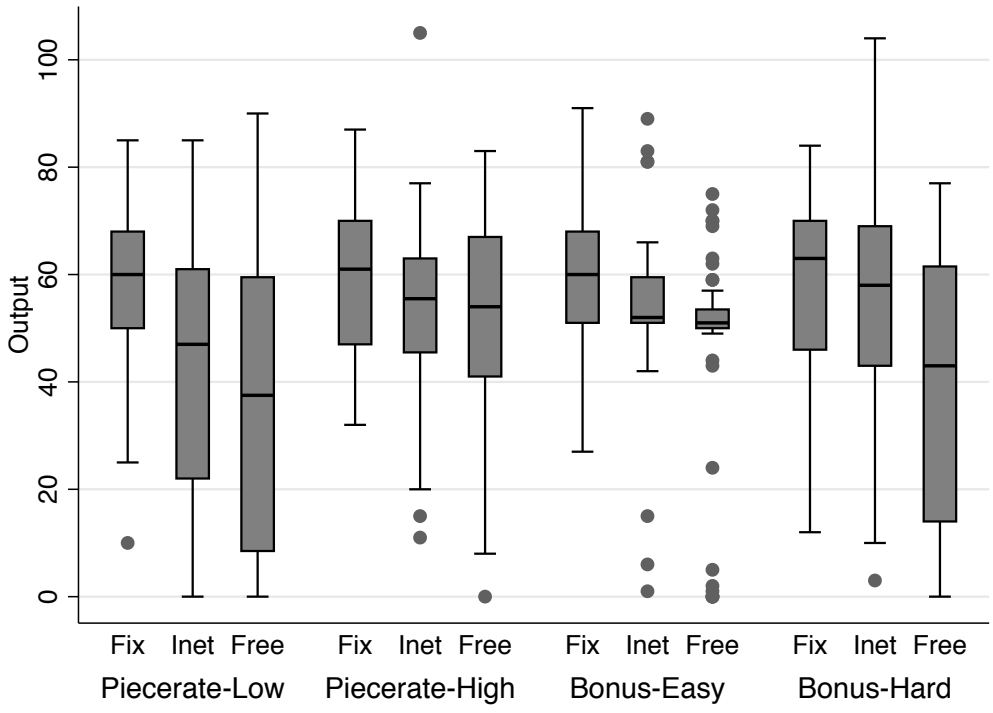


Figure 1.B.1. Boxplot of Outputs in the Different Treatments

Notes: Bold lines give the median outputs, boxes the 25th and 75th quartiles, and whiskers the 1.5xIQR. Circles present outliers, i.e., single observations outside of the whiskers.

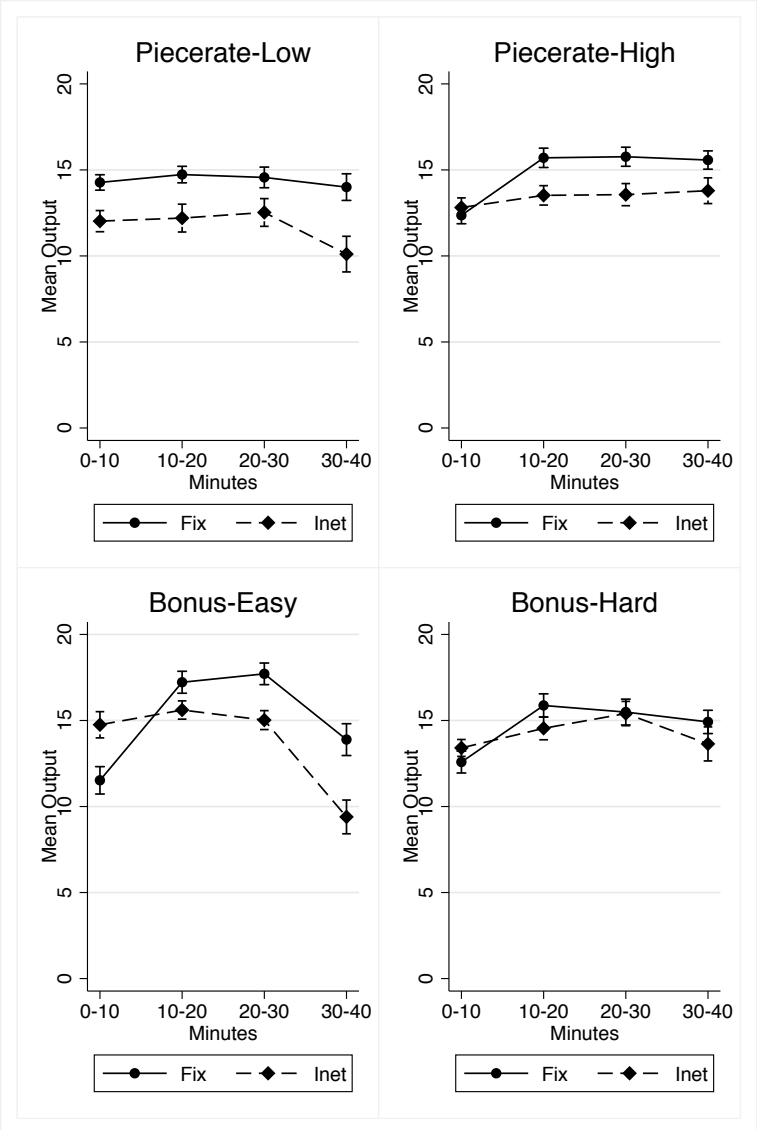


Figure 1.B.2. Mean Output over Time

Appendix 1.C Examples of Experiments with Outside Options

Table 1.C.1 lists real-effort experiments which have treatments with and without outside options. Table 1.C.2 lists real-effort experiments with outside options. These papers do not manipulate the presence of the outside option.

Table 1.C.1. Economic Experiments with Manipulation of Outside Options

Author	Manipulate Outside Option	Outside Option(s)	Real-Effort	Task	Incentive Scheme	Research Topic
Corgnet, Hernán-González, and Schniter (2015)	Yes	None, Internet surfing	Yes	Arithmetic summation task	Fixed wage + piecerate; fixed wage + team-incentive	Reaction to incentive schemes with internet as real-leisure option and without.
Dickinson (1999)	Yes	None, Stop working and Leave	Yes	Typing text	Various fixed wage + piecerate combinations	Individual wage elasticities with work intensity and leisure option.
Eckartz (2014)	Yes	None, Paid pause button	Yes	Letter Puzzle, arithmetic summation task	Fixed-wage; piecerate; tournament	Reaction to incentive schemes across task enjoyability and access to outside option.
Erkal, Gangadharan, and Koh (2018)	Yes	Paid pause button, Stop working and leave, second effort task	Yes	encryption task	tournament	Reaction to tournament incentives with and without outside options.

Table 1.C.1. Economic Experiments with Manipulation of Outside Options

Author	Manipulate Outside Option	Outside Option(s)	Real-effort	Task	Incentive Scheme	Research Topic
Koch and Nafziger (2016)	Yes	None, Internet surfing	Yes	Counting zeros in tables	Principal-Agent: Agent receives fixed wage; Principal piecerate	Gift-exchange game when agents have access to internet or not.

Table 1.C.2. Economic experiments with outside option

Author	Manipulate Outside Option	Outside Option(s)	Real-Effort	Task	Incentive Scheme	Research Topic
Abeler et al. (2011)	No	Stop working and Leave	Yes	Counting zeros in tables	Lottery between fixed payment and piecerate	Compare quitting behavior across wages.

Table 1.C.2. Economic experiments with outside option

Author	Manipulate Outside Option	Outside Option(s)	Real-effort	Task	Incentive Scheme	Research Topic
Berger, Harbring, and Sliwka (2013)	No	Paid pause button	Yes	Counting sevens in tables	Bonus based on rating by supervisor	Effort of subjects working under performance appraisals based bonuses with and without forced distribution.
Blumkin, Ruffle, and Ganun (2012)	No	With consumption goods paid pause button	Yes	Multiplication task	Consumption goods	Experimental test of the equivalence of wage and consumption taxes.
Charness, Masclet, and Villeval (2014)	No	Reading magazines	Yes	decoding one digit numbers to letter	Flat wage	Compare performance across treatments with and without feedback about relative performance (and possibilities of sabotage).
Corgnet, Martin, Ndodjang, and Sutan (2015)	No	Internet surfing	Yes	Arithmetic summation task	Flat or chosen by principal	Test of the effect of influence activities of agents on performance.

Table 1.C.2. Economic experiments with outside option

Author	Manipulate Outside Option	Outside Option(s)	Real-effort	Task	Incentive Scheme	Research Topic
Corgnet, Gómez-Miñambres, and Hernán-González (2015)	No	Internet surfing	Yes	Arithmetic summation task	Piecerate split between principal and agent	Interplay of goal-setting and monetary incentives.
Corgnet, Hernán-González, and Rassenti (2015)	No	Internet surfing	Yes	Arithmetic summation task	Flat wage, piecerate	Effect of firing threats on performance.
Corgnet, Hernan-Gonzalez, et al. (2015)	No	Internet surfing	Yes	Arithmetic summation task	Principal Agent contract, share of agents production (with and without noise)	Compare production in noisy environment to production in environment without noise
Corgnet, Hernan-Gonzalez, and Rassenti (2015)	No	Internet surfing	Yes	Arithmetic summation task	Team incentives or piecerate	Comparison of team incentives (with monitoring) and individual incentives

Table 1.C.2. Economic experiments with outside option

Author	Manipulate Outside Option	Outside Option(s)	Real-effort	Task	Incentive Scheme	Research Topic
Eriksson, Poulsen, and Villeval (2009)	No	Reading Magazines	Yes	Arithmetic summation task	Piecerate; tournament	Impact of incentives and relative performance feedback on performance.
(Falk and Huffman, 2007)	No	Stop working and Leave	Yes	Counting zeros in tables		Subjects had to fulfill work requirement, but could leave as soon as they were done.
Hayashi, Nakamura, and Gamage (2013)	No	Preselected YouTube videos	Yes	Alphabetizing words	Flat payment if leisure option is chosen or piecerate	Study the reaction to different tax regimes.
Hammermann and Mohnen (2014)	No	Reading Magazines	Yes	Solving mathematical equations	Tournament with (non-) monetary prize	Compare subjects' performance in tournaments with monetary and non-monetary prizes
Kajackaite (2015)	No	Paid pause button	Yes	Decoding letters	Piecerate + possibility of piecerate for NRA	Compare subjects performance in treatments with private piecerate and additional piecerate for NRA (possibly unknown, subjects can stay ignorant)

Table 1.C.2. Economic experiments with outside option

Author	Manipulate Outside Option	Outside Option(s)	Real-effort	Task	Incentive Scheme	Research Topic
Kessler and Norton (2016)	No	Internet surfing	Yes	Typing strings	Piecerate	Performance of subjects, with wage decreases due to tax or wage cuts.
Mohnen, Pokorny, and Sliwka (2008)	No	Paid pause button	Yes	Counting sevens in tables	Team-incentives	Effect of information about the performance of the other team member on performance.
Rosaz, Slonim, and Villeval (2016)	No	Stop working and Leave	Yes	Arithmetic math task	Fixed wage + piecerate	Compare quitting behavior across treatments, varying presence of a peer.

Appendix 1.D Conceptual Framework

In a simple theoretical framework, the effort level would be chosen by solving the following maximization problem.

$$\max_{e \geq 0} u(e) = \bar{w} + b(y) + I\delta(y) - c(e,i),$$

The production technology $y = f(e)$ translates effort to output, which we assume to be a continuously differentiable function with $f' > 0$ and $f'' < 0$. The fixed wage, i.e., a lump-sum payment, is represented by $\bar{w}$. The intrinsic motivation is represented by $I\delta(y)$, which indicates the agent's intrinsic motivation for the work. I is an indicator function which is $I = 1$ if the agent is intrinsically motivated and $I = 0$ if not.[39]

Our incentive schemes define $b(y)$, the payment. It simplifies to

$$b(y) = pr \times y$$

for the two piecerate treatments with pr denoting the piecerate (either €0.02 or €0.1). In the two bonus treatments, g denotes the target (either 50 or 100), and it can be written as

$$b(y) = \begin{cases} g \times 0.1, & \text{if } y \geq g \\ 0, & \text{if } y < g \end{cases}$$

The effort costs are represented by $c(e,i)$, which includes the explicit as well as implicit effort costs. The parameter i increases the marginal effort costs depending on the outside options available to the agent.[40,41] We assume that $c'_e(e, i_{Fix}) < c'_e(e, i_{Inet}) \leq c'_e(e, i_{Free}) \;\; \forall e \in [0, E]$, i.e., that marginal effort costs are higher in both environments which provide outside options or alternative activities compared to the environment where subjects have to stay in front of the computer. Additionally, we assume the regularity conditions $\frac{\partial c(e,i)}{\partial e} > 0$ and $\frac{\partial^2 c(e,i)}{\partial e^2} > 0$ on the interval $[0, E]$ and for simplicity $c(0,i) = 0 \;\; \forall i \in \{i_{Fix}, i_{Inet}, i_{Free}\}$. Furthermore, we assume that there is an effort level $E > 0$ at which effort costs increase to infinity, for example due to physical or time constraints, i.e., $\lim_{e \to E} \frac{\partial c(e,i)}{\partial e} = \infty$.[42]

39. We assume $\delta'(y) \geq 0$.

40. This includes, for example, versions of $c(e,i)$ like in Koch and Nafziger (2016), where $c(e,i) = i \times \tilde{c}(e)$ and the parameter i differs between work environments, i.e., effort costs increase when alternative actions are present. Therefore $i_{Fix} \leq i_{Inet} \leq i_{Free}$. This could also incorporate a version of $c(e,i)$, where implicit costs are modeled as utility of leisure, but leisure is negatively related with effort, i.e., time, as in Corgnet, Hernán-González, and Schniter (2015).

41. Implicit effort costs in this setup represent the forgone utility of not allocating the effort or time to other activities.

42. This is the similar to arguing that there is a maximal effort level subjects can exert in the experiment.

Piecerate Incentives:
We first discuss the two piecerate incentive schemes. Under those incentive schemes, the maximization problem leads to the following first-order condition, stating that agents supply effort as long as the marginal benefit of effort is higher than the marginal cost of effort:

$$\frac{\partial b(y)}{\partial y} \times \frac{\partial f(e)}{\partial e} + I\frac{\partial \delta(y)}{\partial y} \times \frac{\partial f(e)}{\partial e} = \frac{\partial c(e,i)}{\partial e}$$

Let us now consider the difference of effort between work environments. We assume that marginal effort costs are higher in the environments with outside options or alternative activities. Therefore, our work environment manipulation increases the marginal effort costs in INET and FREE compared to FIX.

If we keep the incentive scheme as well as intrinsic and extrinsic marginal incentives constant effort changes only via a change in the marginal costs. It is easy to see that both work environments FREE and INET increase the marginal costs of effort. Therefore, the optimal effort level e^* and its associated output decrease.

If we now compare the two piecerate incentive schemes, within a work environment, we only change the marginal benefits of effort. The marginal benefit equals the piecerate pr, which is larger in PIECERATE-HIGH than in PIECERATE-LOW. Therefore, the optimal effort level, i.e., output, increases in the piecerate. However, it could be that subjects provide effort close to E and therefore output differences, i.e., differences in effort levels, are negligible. Still, effort (i.e., output) in PIECERATE-HIGH should always be higher than in PIECERATE-LOW.

Bonus Incentives:
Bonus incentive schemes provide marginal extrinsic incentives only immediately at the target. However, they are not differentiable at that point. Therefore, we have to consider corner solutions and check the participation constraint. In the following, let $\hat{e}$ be the effort level that is needed to meet the target, i.e., $g = f(\hat{e})$. We start by looking at the case without intrinsic motivation.

Case 1: Bonus incentives without intrinsic motivation
Without intrinsic motivation the maximization problem simplifies to

$$\max_{e \geq 0} u(e) = \bar{w} + b(y) - c(e,i).$$

Without intrinsic motivation it can never be optimal to exert effort $e \in (0, \hat{e})$, since the agent could always decrease effort, and therefore his costs, without losing any benefit. Similarly, it is easy to see that no effort above $\hat{e}$ can be optimal. The agent considers either exerting exactly the effort level $\hat{e}$, which is needed to reach the target ($g = f(\hat{e})$), or npt exerting any effort at all. He exerts effort if the participation constraint is fulfilled, i.e.,

$$\bar{w} + b(g) - c(\hat{e}, i) \geq \bar{w}.$$

Therefore, the agent exerts effort if reaching the target is beneficial for him, i.e., when the bonus payment is larger than the cost ($b(g) \geq c(\hat{e}, i)$).

Case 2: Bonus incentives with intrinsic motivation

With intrinsic motivation, additional solutions can arise. These solutions include points on the two intervals $[0, \hat{e})$ and $(\hat{e}, E]$. On these two intervals, the marginal benefits equal zero and therefore possible solutions have to fulfill the following condition.

$$I\frac{\partial \delta(y)}{\partial y} \times \frac{\partial f(e)}{\partial e} = \frac{\partial c(e,i)}{\partial e} \tag{1.D.1}$$

This is equal to the first-order condition without marginal benefits. Let $\tilde{e}$ be the solution to this equation. This effort level $\tilde{e}$ can generally be above or below the effort level $\hat{e}$, which is needed to reach the target.

Case 2 a: Consider $\tilde{e} \geq \hat{e}$

If this is the case, $\tilde{e}$ is also an optimum, since monetary benefits are equal in both situations and providing effort above the target is optimal even in the absence of monetary incentives. Subjects exert effort until the marginal intrinsic motivation equals the marginal costs, which results in an even higher effort level as a subject needs in order to reach the target g.

This implies that for Bonus-Easy we would expect only few outputs above 50, since any additional output above the target would only be driven by workers who have a high intrinsic motivation. In Bonus-Hard subjects will not reach the threshold of 100 and therefore this case does not apply.

Case 2 b: Consider $\tilde{e} < \hat{e}$

If this is the case, the agent has to check this local solution against the decision to exert an effort level $\hat{e}$, i.e., work until he reaches the target. Therefore he has to compare $\bar{w} + b(g) + I\delta(g) - c(\hat{e}, i)$ with $\bar{w} + I\delta(f(\tilde{e})) - c(\tilde{e}, i)$, where the exact solution depends on the exact form of the functions. The agent decides to exert effort level $\hat{e}$ if the additional costs of exerting the effort are lower than the additional benefits, i.e.,

$$b(g) + I\delta(f(\hat{e})) - I\delta(f(\tilde{e})) \geq c(\hat{e}, i) - c(\tilde{e}, i)$$

.

Otherwise the agent will provide effort level $\tilde{e}$, which produces an output below the target. This case shows that there can be workers who exert a positive effort level, which leads to an output below the target. Especially in treatment Bonus-Hard subjects will not reach the threshold of 100. Therefore, observed output is due to workers for which the intrinsic motivation induces the optimal effort $\tilde{e}$, such that $0 < \tilde{e} < \hat{e}$.

Let us now consider the difference of effort between work environments. Let us first consider *Case 1*. If the effort costs increase due to a change in the work environment, it becomes more difficult to fulfill the participation constraint. Therefore some subjects will now exert less effort. For *Case 2 a*, we can see that optimal effort decreases if marginal effort costs increase. Since the intrinsic motivation does

not change, higher marginal costs will induce lower effort. Also for *Case 2 b*, effort can only decrease if effort costs increase. Consider first those subjects who exert an effort level $\hat{e}$. Some of these subjects might still exert effort until the target is reached. However, for some subjects it might be optimal to exert less effort if effort costs increase. Those subjects, who already exerted an effort level below $\hat{e}$ will also decrease their effort.

Appendix 1.E Structural Estimation

1.E.1 Parametrization of Conceptual Framework

We have to parametrize the functions in order to estimate the model structurally. We will focus on the piecerate treatments, as those have inner solutions and can be easily estimated. In general, our parametrization closely follows common specifications in the real-effort literature (e.g., DellaVigna and Pope, 2018). The production technology $y = f(e) = e$ translates effort to output. The fixed wage, i.e., a lump-sum payment, is represented by $\bar{w}$. The intrinsic motivation is represented by $I\delta(y)$. We parametrize this in a linear way, i.e., all agents are intrinsically motivated by $\delta(y) = s * e$.

The effort costs are represented by $C(e, i)$, which includes the explicit as well as implicit effort costs. We assume that the parameter i increases the marginal effort costs depending on the outside options available to the agent. Setting up the effort costs in this setup is crucial, since our treatment variations are changing the implicit effort costs. There are two possible ways how effort costs could change. First, physical effort costs could potentially be multiplied, i.e., every effort unit is more expensive as an agent faces opportunity costs. This would multiply the physical effort costs, for example, with a factor o_i. Second, agents face an additional cost of effort for every unit, i.e., the foregone utility of spending this effort differently. Therefore, we add another term to the cost function, for simplicity a linear term $a_i * e$. These two possibilities lead to the following effort costs $C(e, i)$, where our treatments might potentially change only o or a (see discussion below):

$$C(e, i) = exp(o_i)c(e) + a_i e$$

.

For $c(e)$, we can use two versions of effort costs commonly used in the literature: a power cost function and an exponential cost function.

1. Power Cost Function:

$$c(e) = \exp(k) \ \frac{e^{1+\gamma}}{1+\gamma}$$

The power cost function has a constant elasticity with respect to the value of effort (i.e., $s + p$) of $1/\gamma$ and can be scaled with some (positive) parameter $exp(k)$.[43]

2. Exponential Cost Function:

A natural alternative is a function with a decreasing elasticity, one function with such a structure being the exponential cost function:

$$c(e) = \exp(k) \ \frac{\exp(\gamma e)}{\gamma}.$$

43. $exp(k)$ is used to ensure that the parameter is positive.

Given this parametrization, the agent solves the following maximization problem:

$$\max_{e>0} u(e) = \bar{w} + (p+s) * e - exp(o_i)c(e) - a_i * e$$

This will lead to a first-order condition which holds with equality due to the properties of $c(e)$. It is setting the marginal costs of effort equal to the marginal benefit. Given the two parametrizations of $c(e)$, this leads to the following solutions for the optimal effort. For **power costs**:

$$log(e^*) = \frac{1}{\gamma}[log(p+s-a_i) - k - o_i] \tag{1.E.1}$$

and for the **exponential cost function**:

$$e^* = \frac{1}{\gamma}[log(p+s-a_i) - k - o_i] \tag{1.E.2}$$

1.E.2 Structural Estimation

To estimate the above model structurally with non-linear least squares, we need to add some noise term. If we add a noise term to the cost of effort function, the cost function $C(e,i)$ of worker j is

$$C_j(e,i) = exp(o_i)c(e) * exp(-\gamma * \epsilon_j) + a_i e$$

.

This will lead to the following two equations:

$$log(e_j) = \frac{1}{\gamma}[log(p+s-a_i) - k - o_i] + \epsilon_j \tag{1.E.3}$$

$$e_j = \frac{1}{\gamma}[log(p+s-a_i) - k - o_i] + \epsilon_j. \tag{1.E.4}$$

For the case without opportunity costs, $a_i = 0$ and $o_i = 1$, and both first-order conditions have three unknown parameters (γ, s, k). In each work environment with opportunity costs, both first order conditions have two additional unknown parameters (a_{Inet}, o_{Inet} and a_{Free}, o_{Free}); the work environments with opportunity costs add four additional parameters in total.

In order for our model to be identified, we use the following restrictions: First, we only allow changes of opportunity costs to matter due to changes in a_i, i.e., each effort unit produces some additional costs. This idea is also in line with the standard idea of how opportunity costs should enter. This shifts the marginal effort costs upwards and, by this, potentially decreases effort. We therefore restrict $o_i = 1$ in all

environments.[44, 45] Furthermore, we can add an additional parameter to account for potential heterogeneous effort costs due to differentiability. For this, we can multiply the effort costs with an additional parameter $exp(ability) * performance\_trial$. This allows for heterogeneous effort cost functions due to ability.

1.E.3 Results

Table 1.E.1. Structural Parameters of Effort Costs

	Power Costs			Exponential Cost		
	(1)	(2)	(3)	(4)	(5)	(6)
γ	6.1402	3.7928	3.5236	0.1659	0.1925	0.0835
	(8.2545)	(5.6732)	(2.5826)	(0.1854)	(0.2733)	(0.0571)
s	14.4257	36.7609	8.5172*	16.8910	16.8010	11.3334**
	(20.9882)	(150.3160)	(5.0409)	(17.6572)	(21.9920)	(4.6492)
a_{Inet}	15.6364	36.6680	8.4170	17.4483	17.8837	9.1744**
	(22.9803)	(148.1987)	(5.4080)	(20.2475)	(24.2706)	(3.9922)
a_{Free}	16.3858	38.4273	10.1634**	18.3353	18.4262	10.9095***
	(21.2072)	(149.5680)	(5.0685)	(19.1204)	(23.2552)	(3.8420)
k	-21.7657	-11.5388	-8.7176	-6.5898	-8.1466	0.5525
	(32.3766)	(24.5867)	(8.3318)	(10.1751)	(15.2176)	(1.7276)
o		-1.0136			0.0976	
		(3.7866)			(0.5420)	
ability			-0.5118			-0.4680
			(0.3772)			(0.3190)
N	283	283	281	287	287	285
R^2	.18	.18	.26	.15	.15	.33

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. Table presents structural estimates of 1.E.3 and 1.E.4 using non-linear least squares. Standard errors in parentheses.

Table 1.E.1 presents the results of the estimation. We first focus on the main parameters of interest, a_{Inet} and a_{Free}. We observe that both parameters are positive across all specifications. In both environments, subjects have to pay an additional cost for each unit produced. However, in our main specifications (1) and (4), both

44. As a robustness check, we allow one additional parameter that is the same for both opportunity cost environments.

45. We estimate a version where we allow only for changes in a_i. We observe a very low elasticity with respect to changes in p in Fix. A model which tries to fit these moments will therefore estimate a very high γ. Multiplying this kind of effort costs curve with a parameter smaller than 1 will reduce effort (as marginal costs are higher), yet will be unable to match the elastic response to effort in the other two environments.

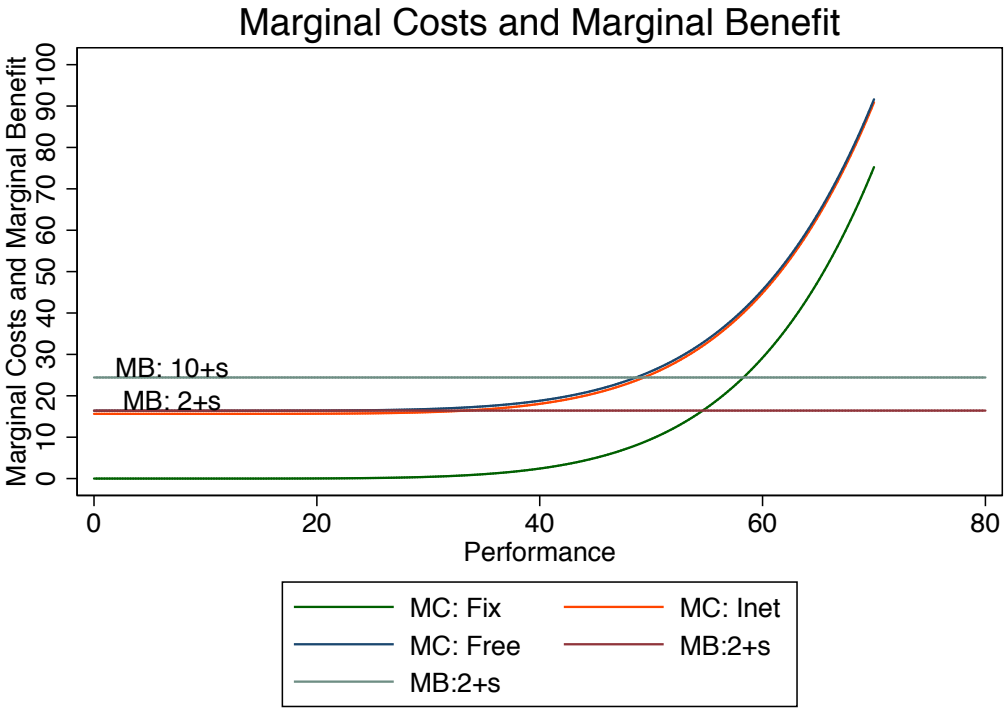


Figure 1.E.1. Marginal Costs and Marginal Benefit with Power Costs

parameters are not significant. Only if we allow ability to enter in (3) and (6) parameters become significant. This is due to the high variance in our data. Overall, the coefficients are in line with the hypothesis that opportunity costs shift the effort costs upwards and that the shift is slightly higher in FREE than in INET. Our estimates also show that implicit costs are of a similar size as the intrinsic motivation parameter. The effect of introducing opportunity costs therefore has a similar effect in terms of magnitude as setting off intrinsic motivation.

To check the fit of our model, we tabulate the predicted output of the model and the actual output in Table 1.E.2, using the model in column (1). In Figure 1.E.1, we plot the results of column (1) to illustrate the results and the mechanics. In INET and FREE, the marginal cost curve is shifted upwards due to a_{Inet} and a_{Free}. As the marginal benefits are constant across the environments, this reduces both the observed and estimated outputs.

Table 1.E.2. Output and Predicted Output, Piecerate Based Incentives

	Fix		Inet		Free	
	Piecerate-Low	Piecerate-High	Piecerate-Low	Piecerate-High	Piecerate-Low	Piecerate-High
	Mean	Mean	Mean	Mean	Mean	Mean
Predicted	54.63255	58.27958	33.32347	49.34271	20.49268	48.63174
Output	57.5625	59.40426	41.87234	53.6875	36.20833	52.34694
Observations	48	47	47	48	48	49

Bonus-based incentives

We can use our estimated parameters and calculate a prediction for the bonus-based treatments using some simplifications and assumptions. We use the estimates of column(1) and show the results of this exercise in Table 1.E.3. For Fix, we can simply drop the marginal piece rate incentives. In the case Bonus-easy, Case 2a from above is fulfilled. The intrinsic motivation equilibrium predicts effort slightly above 50. For Bonus-hard, Case 2b from above is fulfilled. However, reaching the target is too costly. Therefore we get the same prediction as before.

For both environments Inet andFree intrinsic motivation is not high enough to induce an equilibrium with pure intrinsic motivation, as $s < a$. Therefore we only have to check whether the subjects target the goal or not. We can simply compare, costs and benefits of working until the goal is reached. We therefore predict an output of 50 for Bonus-easy and 0 for Bonus-hard.

For the fix environment, our model makes predictions that are generally in line with our results. For the other two environments, the point predictions are also close to the observed outcome in Bonus-Easy. The model, however, fails to predict the observed outcome in the Bonus-Hard environments. Generally, the model makes very sharp predictions, although it does not take into account the heterogeneity in intrinsic motivation.

Table 1.E.3. Output and Predicted Output, Bonus Based Incentives

	Fix		Inet		Free	
	Bonus-Easy	Bons-Hard	Bonus-Easy	Bons-Hard	Bonus-Easy	Bons-Hard
	Mean	Mean	Mean	Mean	Mean	Mean
Predicted	53.48947	53.48947	50	0	50	0
Output	60.08889	58.20833	53.1875	54.42553	45.9375	38.6875
Observations	45	48	48	47	48	48

Appendix 1.F Time Used to Work on the Task

Our treatment variation gave subjects the possibility to adjust their effort at the extensive margin in INET and FREE by using an outside option. In the following, we show that subjects actually used the outside option. In addition, we show that our treatment differences mainly result from the time worked on the task (extensive margin) rather than the speed (intensive margin).

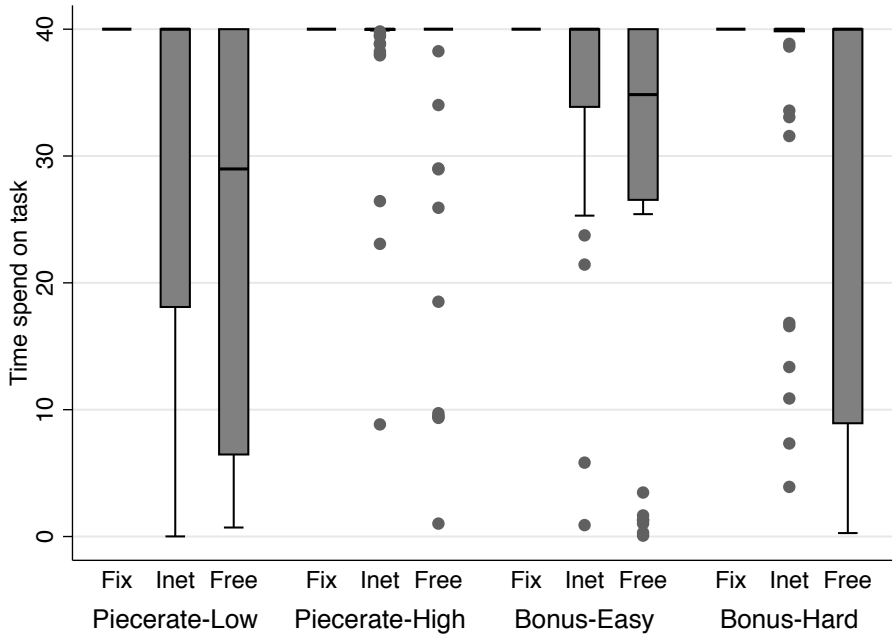


Figure 1.F.1. Boxplot of Time Spent Working on the Task

Notes: Bold lines give the median outputs, boxes the 25th and 75th quartiles, and whiskers the 1.5xIQR. Circles present outliers, i.e., single observations outside of the whiskers.

Across all treatments, the time spent working on the task significantly correlated with the output ($\rho = 0.6048, p < 0.01$). Figure 1.F.1 shows the distribution of time spend working on the task. Obviously, subjects worked on average significantly less than 40 minutes on the task in the INET- and FREE-treatments (both $p < 0.01$, Wilcoxon signed-rank test). Table 1.F.1 presents OLS regressions which show how our work environments affect the time subjects work on the task, analogously to the treatment effect tables in the paper.

In a next step we adjust the output by the total time an individual worked on the task. If the time spent working on the task is the main driver of the treatment differences, we should not observe significant differences between treatments anymore. Figure 1.F.2 gives the result of this exercise. The previously reported significant differences between FIX, INET, and FREE turn insignificant when controlling for the time spend working on the task: In PIECERATE-LOW, the p-value changes

Table 1.F.1. Treatment Effects for Time Working on the Task

	Over All Incentives		Piecerate-Low		Piecerate-High		Bonus-Easy		Bonus-Hard	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Inet	-4.88***	-5.23***	-9.36***	-8.22***	-1.48*	-1.73*	-4.59***	-5.00***	-4.16***	-4.20***
	(0.74)	(0.75)	(2.03)	(1.89)	(0.77)	(0.96)	(1.22)	(1.28)	(1.42)	(1.47)
Free	-10.82***	-10.64***	-15.79***	-14.06***	-4.62***	-4.91***	-10.07***	-9.66***	-12.94***	-12.73***
	(1.05)	(1.05)	(2.35)	(2.42)	(1.48)	(1.49)	(1.88)	(1.85)	(2.30)	(2.38)
Constant	40.00***	35.10***	40.00***	18.08**	40.00	39.36***	40.00***	46.62***	40.00***	39.26***
	(0.00)	(3.12)	(0.00)	(7.73)	(.)	(3.85)	(0.00)	(7.65)	(0.00)	(7.16)
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N	571	568	143	142	144	142	141	141	143	143
R^2	.16	.18	.22	.29	.076	.084	.17	.22	.2	.21
(Inet and Free = 0)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
(Inet = Free)	0.00	0.00	0.04	0.06	0.06	0.07	0.02	0.03	0.00	0.00

Notes: Table presents results from an OLS estimation with output as dependent variable. * p < 0.1, ** p < 0.05, *** p < 0.01. Robust Standard errors in parentheses.

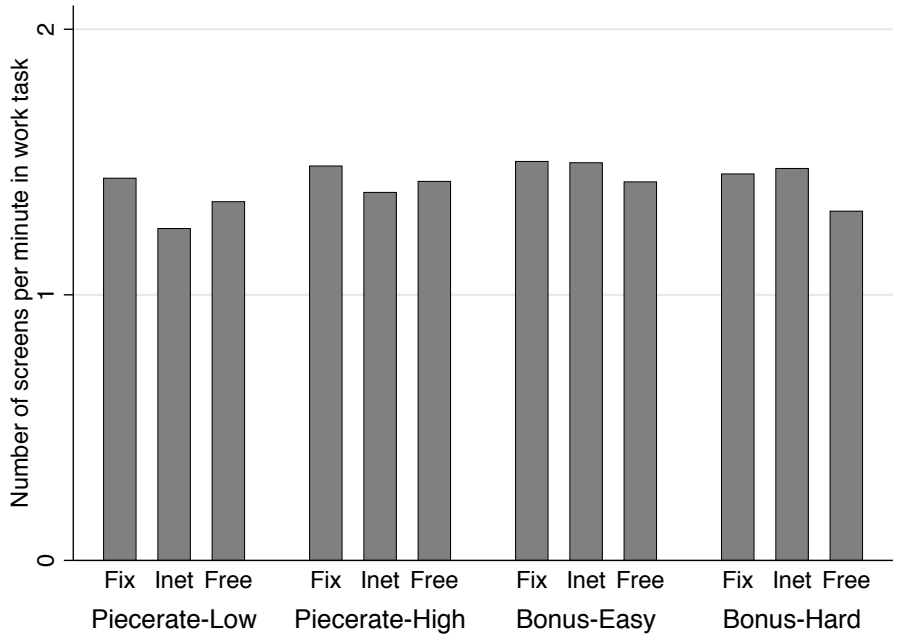


Figure 1.F.2. Output, Adjusted by Working Time

from $p < 0.01$ to $p = 0.1277$; in Bonus-Easy from $p < 0.01$ to $p = 0.8592$, in Bonus-Hard from $p < 0.01$ to $p = 0.2579$; and in Piecerate-High from $p = 0.1636$ to $p = 0.3316$ (all Kruskal-Wallis test).[46]

46. The Kruskal-Wallis test is a multi-sample generalization of the two-sample Mann-Whitney u-test.

We additionally analyze the time subjects need to complete one screen to measure effort adjustments at the intensive margin. Table 1.F.2 gives the average and median time needed per screen. Testing the median time per screen reveals no significant differences in the medians for all incentive schemes, except for Piecerate-High (all $p > 0.446$, median test). For Piecerate-High, the median does differ across work environments ($p = 0.098$, median test).[47] Comparing the means gives no significant differences for any incentive scheme (all $p > 0.2250$, Kruskal-Wallis test).

Table 1.F.2. Time per Screen

		Over All Incentives	By Incentive Scheme			
			Piecerate-Low	Piecerate-High	Bonus-Easy	Bonus-Hard
Fix	Mean	45.14	47.60	42.88	42.17	47.67
	Median	39.34	40.00	39.34	40.00	38.10
	SD	22.36	30.95	11.49	11.79	27.30
	N	188	48	47	45	48
Inet	Mean	46.48	52.89	47.65	42.04	43.70
	Median	40.96	41.90	41.39	38.83	41.74
	SD	18.51	25.20	20.79	10.05	13.08
	N	188	45	48	48	47
Free	Mean	45.69	49.20	44.24	40.62	48.49
	Median	40.96	41.90	41.39	38.83	41.74
	SD	18.21	25.03	12.98	11.67	19.01
	N	185	47	48	44	46

Notes: SD: standard deviation, N: number of independent observations

47. One possibility which influences the time subjects need to complete a screen is the existence of opportunity costs and the salience of those in a given situation. See Kurzban et al. (2013) for a discussion in Pyschology.

Appendix 1.G Explaining the Usage of the Outside Option

We repeat the analysis from the paper and check whether our questionnaire measures can explain the usage of the outside option. Table 1.G.1 presents results for the intensive margin, i.e., the time subjects work on the task, and Table 1.G.2 presents the effects on the extensive margin, i.e., on the probability of using the outside option. For the latter we present marginal effects using a logit model with an indicator if someone used the outside option. Overall, results are in line with the analysis in the paper for both the time subjects spent working on the task and the probability of using the outside option.

Table 1.G.1. Time Spent Working

	(1)	(2)	(3)
INET x CRT score	0.0753		0.5809
	(0.6505)		(0.6573)
FREE x CRT score	-1.8688**		-1.7050*
	(0.8899)		(1.0040)
INET x Conscientiousness		1.0099**	0.9303**
		(0.3992)	(0.4246)
FREE x Conscientiousness		0.6986	0.2125
		(0.5936)	(0.6767)
INET x Ability			0.5703
			(0.4986)
FREE x Ability			0.3815
			(0.6483)
Treatment FE	Yes	Yes	Yes
Gender and Age	No	No	Yes
Big 5 (w/o Cons), Risk	No	No	Yes
N	383	383	380
R^2	.14	.14	.2

Notes: Table presents least squares regression using the time worked on the task as dependent variable. * p < 0.1, ** p < 0.05, *** p < 0.01. Robust Standard errors in parentheses. Big 5 (w/o Cons.) controls for the other Big 5 traits and risk for general risk attitudes.

Table 1.G.2. Probability of Using Outside Option

	(1)	(2)	(3)
used outside option			
INET x CRT score	0.1343		0.1027
	(0.1309)		(0.1427)
FREE x CRT score	0.3374**		0.2100
	(0.1421)		(0.1651)
INET x Conscientiousness		-0.1764**	-0.1271
		(0.0872)	(0.0899)
FREE x Conscientiousness		-0.2006**	-0.0936
		(0.0970)	(0.1183)
INET x Ability			-0.1110
			(0.1078)
FREE x Ability			0.0600
			(0.1007)
Constant	-0.7592**	0.7431	2.9188
	(0.3825)	(0.5236)	(1.8338)
Treatment FE	Yes	Yes	Yes
Gender and Age	No	No	Yes
Big 5 (w/o Cons), Risk	No	No	Yes
N	383	383	380

Notes: Table presents odd ratios from a logit model with an indicator if someone used the outside option as dependent variable. * p < 0.1, ** p < 0.05, *** p < 0.01. Robust Standard errors in parentheses. Big 5 (w/o Cons.) controls for the other Big 5 traits and risk for general risk attitudes.

Appendix 1.H Experimental Instructions

1.H.1 General Information

General Instructions

Thank you for participating in today's study. Please read the following instructions carefully. If you have questions, you can can ask them at the end of the introduction. To carry out the study, it is very important that you do not communicate with other participants. Therefore, you are not allowed to talk to others. If you communicate with another participant regardless of this, you will have to leave the experiment and will receive no payment.

In this study, you will have the possibility to earn money. The payment at the end of the study is done individually and no other participant will know how much you earned in this study.

Instructions for the task

The task in this study is to set as many sliders as possible to the middle position (position 50) in a given time. Each slider is located at the left end (position 0) and can be moved in steps of 1 to the right end of the scale (position 100). The current position of the slider is displayed to the right of the scale. Please use your mouse to move the slider on the scale as desired. Only when all sliders on the screen are located at the center (position 50) will a red button appear. By pressing this button, you confirm that all sliders are in the middle and you will earn a point.

Please note: You will only earn a point if all sliders are in the middle and you pressed the red button. The task will start simultaneously for all participants. You can see your personal score at the top right corner of the screen. We will now start with a trial round. In this trial round, you can familiarize yourself with the task. Following the trial round, you will receive further information.

1.H.2 Treatment-Specific Instructions

Performance-oriented remuneration

We ask you to do your job carefully. Please try to finish as many screens with sliders as possible within the next 40 minutes. It is not possible to terminate the task before that time is up, since the payment will only be done at the end of the experiment.[48] Generally, the more points you earn, the higher the payment, which you will receive from us immediately afterwards in cash. The following will be applied:

[**PR Treatments:**]

48. For FREE treatments, this sentence was replaced by the following: You can stop working on the task at any time. If you decide to stop working on the task, you can collect your payment and the study is finished for you.

- You will receive a basic wage of 10 Euros. That means that you will earn at least 10 euros for the 40 minutes.
- In addition to your basic wage, you will receive a bonus payment. The size of this bonus payment is based on the number of points you collect:

  For each point you will get an additional 2 [10] cents.

  [**Bonus Treatments:**]

- You will receive a basic wage of 10 Euros. That means that you will earn at least 10 euros for the 40 minutes.
- In addition to your basic wage, you may receive a bonus payment of 5 [10] Euro. It depends on the number of points you accumulate whether you receive this bonus or not. We set a personal goal for you that is 50 points. If you do not reach this goal within 40 minutes, you will not receive the bonus. If you reach the goal, you will receive the bonus payment of 5 [10] Euros.
  - *Example:* You will receive a bonus of 5 [10] Euros as soon as you have collected 50 points or more (also, if you have collected, for example, 105 points). If you have accumulated less than 50 points, you will not receive the bonus.

You will only earn a point if all sliders are in the middle and you have pressed the red button.

[**Inet Treatments:**] During the next 40 minutes, you can also surf the Internet. You can access the internet by clicking the "Internet" button. If you click on this button, Internet Explorer will open. As long as you are on the Internet, your work will be interrupted. To continue working on the task, close Internet Explorer and click "Proceed". You can also interrupt your work several times and return to the task at any time. After 40 minutes, Internet Explorer closes automatically and you can no longer return to the task either.

[**Free treatments:**] You can stop working on the task at any time. If you decide to stop working on the task, your payment is based on all the points you have earned up to this point.

Appendix 1.I Screenshot and Implementation

Original screen resolution was 1920 x 1200 and is adjusted for the screenshots to 1024 x 768.

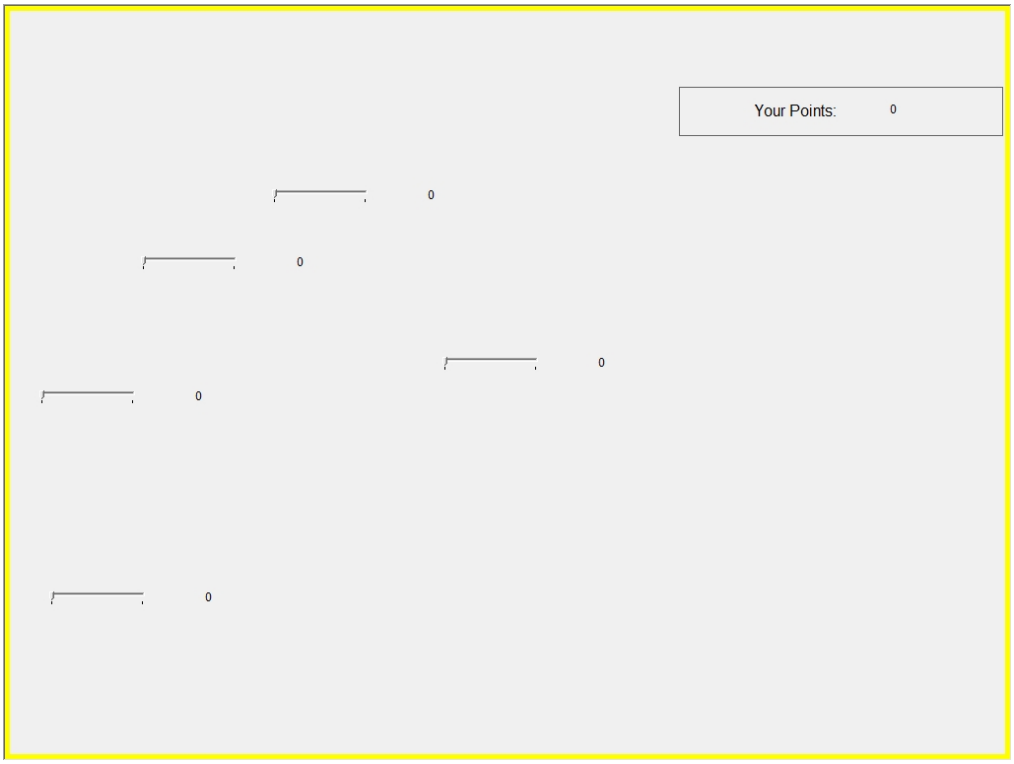


Figure 1.I.1. Screenshot of Real-Effort Screen in Fix

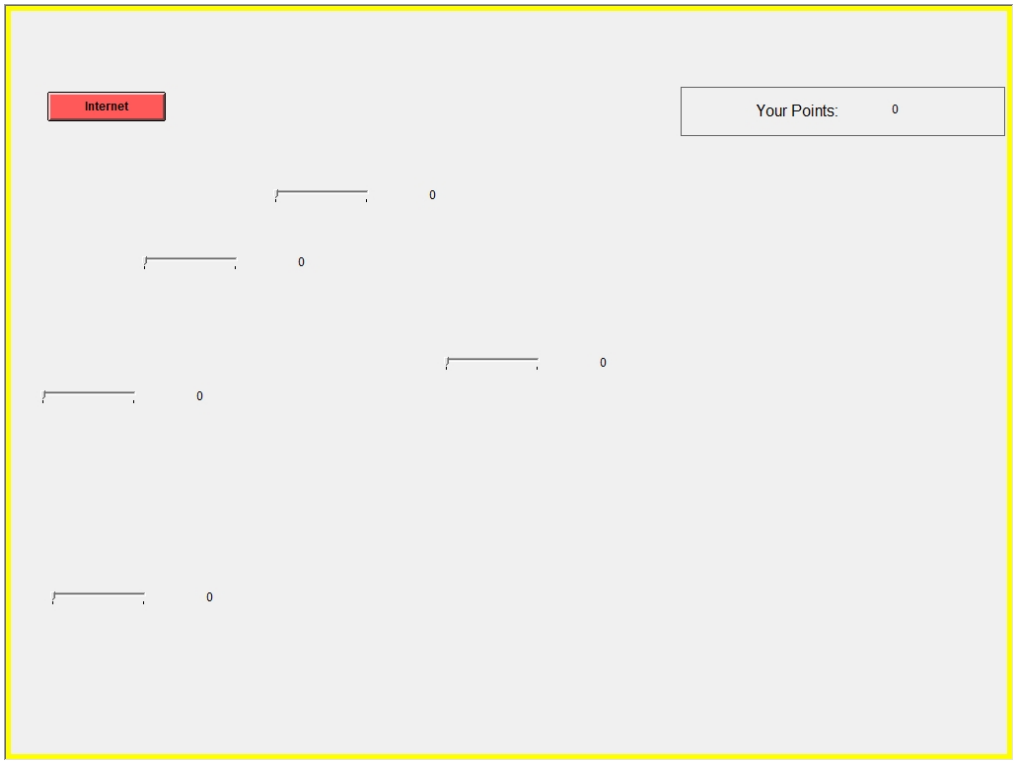


Figure 1.I.2. Screenshot of Real-Effort Screen in INET

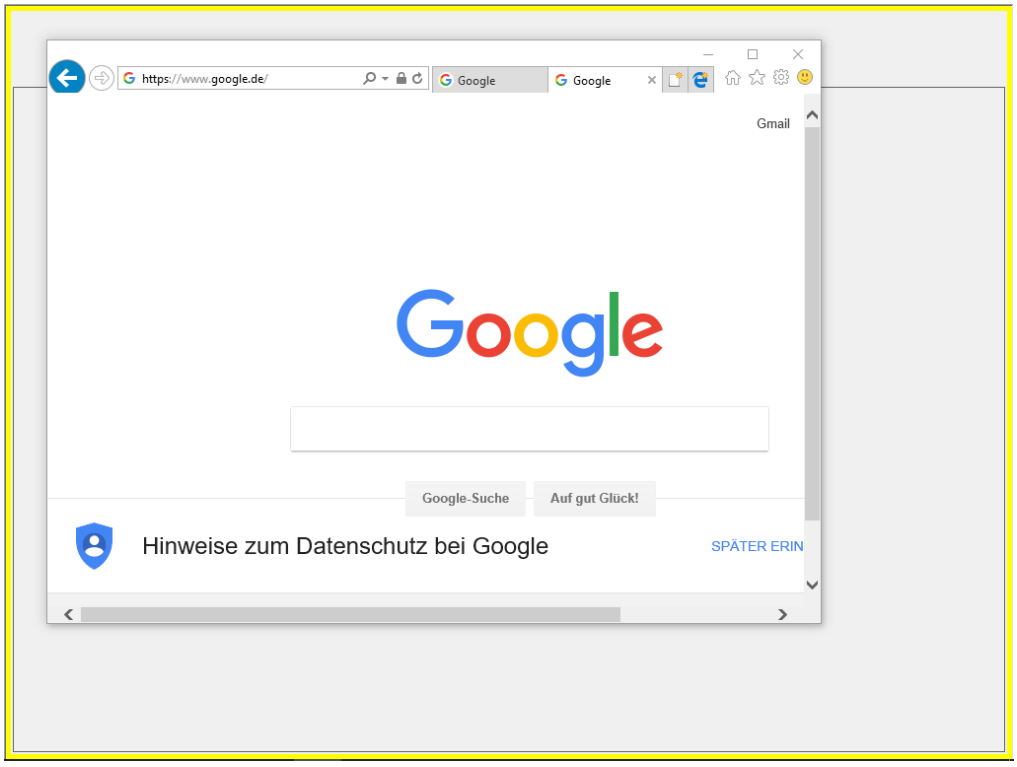


Figure 1.I.3. Screenshot of Internet Access Screen in INET

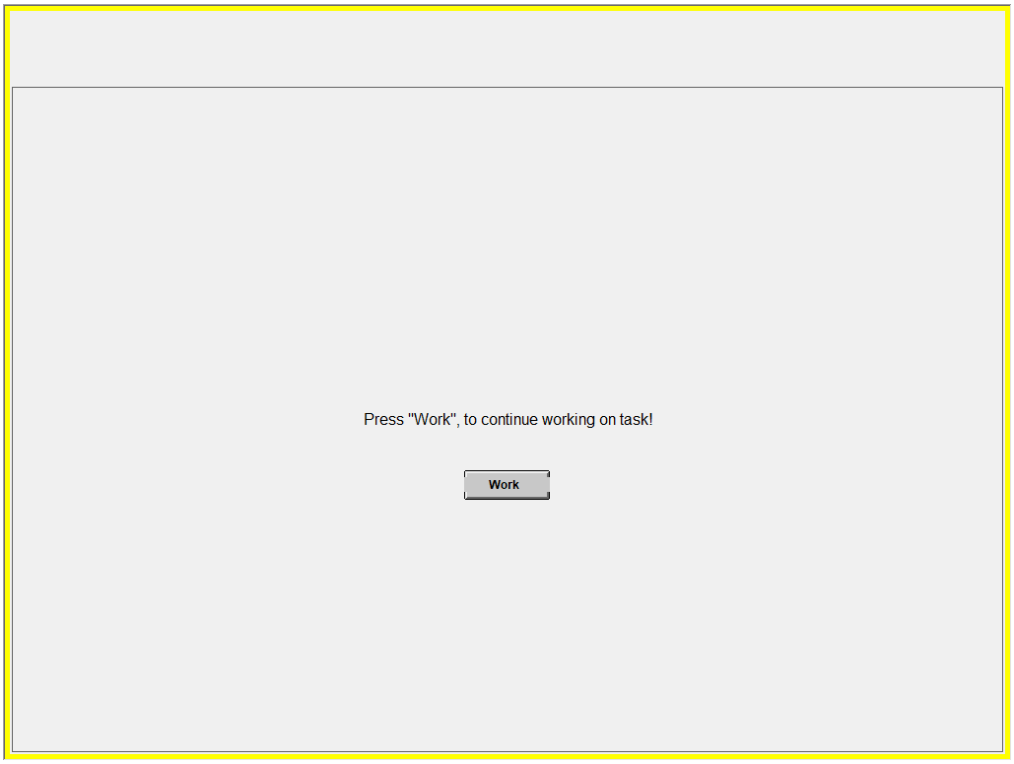


Figure 1.I.4. Screenshot of Blocked Real-Effort Screen in INET

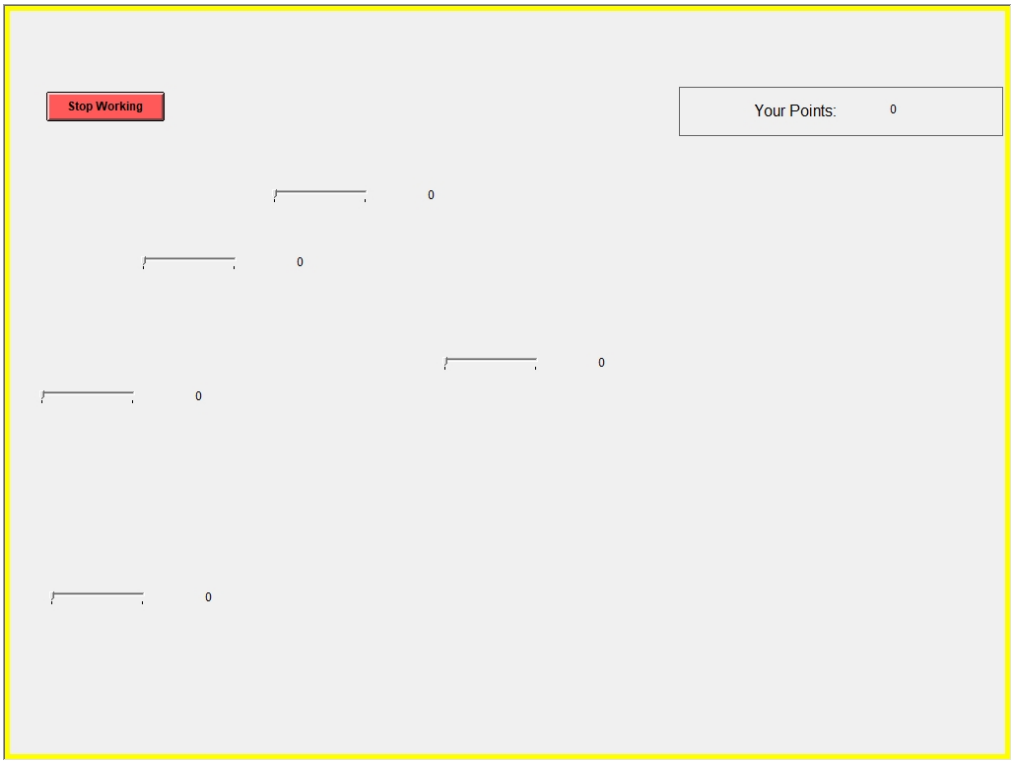


Figure 1.I.5. Screenshot of Real-Effort Screen in FREE

1.I.1 Implementation of INET and FREE

We implemented the INET environment by adding a button to the screen which allowed subjects to open Internet Explorer on their computer (see Figure 1.I.2 and Figure 1.I.6). This button calls an external program, i.e., in our case Internet Explorer. Furthermore, the button indicates that the subjects is online and switches an indicator variable *Internet* to 1. If the indicator variable is switched to 1 a box is displayed, which covers the real-effort task (see Figure 1.I.4). Thus, upon pressing the button, an Internet Explorer window opens, the slider task is blocked, and subjects can surf the web. If subjects want to return to the task, they can close the Internet Explorer or click on the zleaf window. Subjects would automatically return to zleaf, as it is still running in full screen mode. The blocking screen entailed a button that switched the indicator variable back to 0 and the real-effort task would be displayed again. The key "Alt" on the subjects' keyboard was disabled; therefore they could not switch between windows or close any other window except for the Internet Explorer window. Figure 1.I.6 gives details on the implementation. We implemented FREE in a similar way. The only difference was that, upon clicking the button 'Stop Working', subjects would leave the work stage instead of accessing the internet.

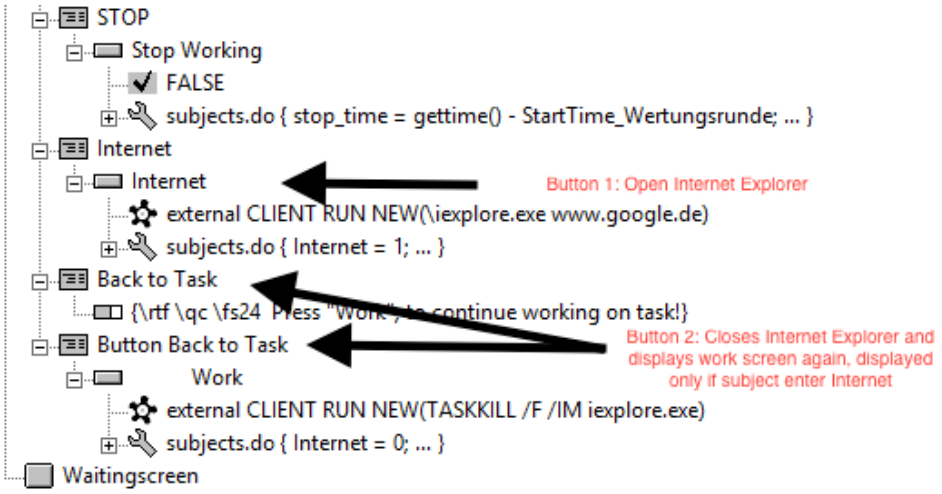


Figure 1.I.6. Implementation of INET in Z-Tree

1.I.2 Payment and Procedures in FREE

Subjects in FREE could arrive during a given time window on a given day. Subjects entered the laboratory through the entrance and were quietly directed to the second room by an experimenter (see Figure 1.I.7). Registration took place in the second room and subjects would proceed with the experiment in their computer cabin in the first room. Cabins are separated by walls and subjects work behind a closed curtain (see Figure 1.I.8). All cabins are accessible without anyone being disturbed. Upon completion subjects are told to go back to the second room quietly and payment was done there in private.

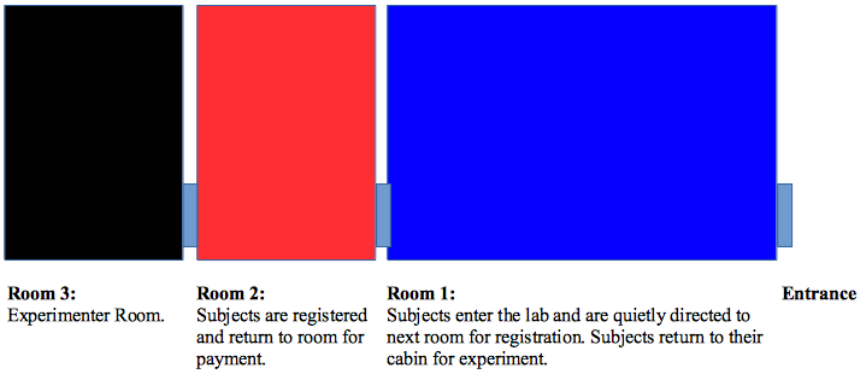


Figure 1.I.7. Sketches of BonnEconLab

(a) Computer cabins 1

(b) Computer cabins 2

(c) Computer cabins 3

Figure 1.I.8. Photos of Laboratory Room

References

Abeler, Johannes, Armin Falk, Lorenz Goette, and David Huffman. 2011. "Reference Points and Effort Provision." *American Economic Review* 101 (2): 470–92. DOI: doi:10.1257/aer.101.2.470. [5, 32]

Alder, G. Stoney, Terry W. Noel, and Maureen L. Ambrose. 2006. "Clarifying the effects of Internet monitoring on job attitudes: The mediating role of employee trust." *Information & Management* 43 (7): 894–903. DOI: https://doi.org/10.1016/j.im.2006.08.008. [24]

Araujo, Felipe A., Erin Carbone, Lynn Conell-Price, Marli W. Dunietz, Ania Jaroszewicz, Rachel Landsman, Diego Lamé, Lise Vesterlund, Stephanie W. Wang, and Alistair J. Wilson. 2016. "The Slider Task: An Example of Restricted Inference on Incentive Effects." *Journal of the Economic Science Association* 2 (1): 1–12. DOI: 10.1007/s40881-016-0025-7. [6, 22, 23]

Ariely, Dan, Uri Gneezy, George Loewenstein, and Nina Mazar. 2009. "Large Stakes and Big Mistakes." *Review of Economic Studies* 76 (2): 451–69. DOI: 10.1111/j.1467-937X.2009.00534.x. eprint: /oup/backfile/content_public/journal/restud/76/2/10.1111_j.1467-937x.2009.00534.x/3/76-2-451.pdf. [5]

Arni, Patrick, and Amelie Schiprowski. 2017. "Job Search Requirements, Effort Provision and Labor Market Outcomes." [23]

Asch, BJ. 1990. "Do Incentives Matter? The Case of Navy Recruiters." *Industrial & Labor Relations Review* 43 (3): 89S–106S. URL: http://ilr.sagepub.com/content/43/3/89S.short. [5]

Barrick, M. R., and M.K. Mount. 1991. "The Big 5 Personality Dimensions and Job Performance: a Meta-Analysis." *Personnel Psychology* 44 (1): 1–26. [20]

Berger, J., C. Harbring, and D. Sliwka. 2013. "Performance Appraisals and the Impact of Forced Distribution – An Experimental Investigation." *Management Science* 59 (1): 54–68. DOI: 10.1287/mnsc.1120.1624. [33]

Blumkin, Tomer, Bradley J. Ruffle, and Yosef Ganun. 2012. "Are Income and Consumption Taxes Ever Really Equivalent? Evidence from a Real-Effort Experiment with Real Goods." *European Economic Review* 56 (6): 1200–19. DOI: http://dx.doi.org/10.1016/j.euroecorev.2012.06.001. [33]

Bock, Olaf, Ingmar Baetge, and Andreas Nicklisch. 2014. "hroot: Hamburg Registration and Organization Online Tool." *European Economic Review* 71: 117–20. DOI: http://dx.doi.org/10.1016/j.euroecorev.2014.07.003. [9]

Bradler, Christiane, Robert Dur, Susanne Neckermann, and Arjan Non. 2016. "Employee Recognition and Performance: A Field Experiment." *Management Science* 62 (11): 3085–99. [3]

Camerer, Colin F, and Roberto A Weber. 2013. "Experimental Organizational Economics." *Handbook of Organizational Economics*, 213–62. [5]

Carpenter, Jeffrey. 2016. "The Labor Supply of Fixed-Wage Workers: Estimates from a Real Effort Experiment." *European Economic Review* 89: 85–95. [3]

Carpenter, Jeffrey, Peter Hans Matthews, and John Schirm. 2010. "Tournaments and Office Politics: Evidence from a Real Effort Experiment." *American Economic Review* 100 (1): 504–417. [5]

Charness, Gary, and Peter Kuhn. 2011. "Lab Labor: What Can Labor Economists Learn from the Lab?" *Handbook of Labor Economics* 4: 229–330. [4, 5]

Charness, Gary, David Masclet, and Marie Claire Villeval. 2014. "The Dark Side of Competition for Status." *Management Science* 60 (1): 38–55. DOI: 10.1287/mnsc.2013.1747. [5, 33]

Cohn, Alain, Ernst Fehr, and Lorenz Goette. 2015. "Fair Wages and Effort Provision: Combining Evidence from a Choice Experiment and a Field Experiment." *Management Science* 61 (8): 1777–94. [3]

Corgnet, Brice, Joaquín Gómez-Miñambres, and Roberto Hernán-González. 2015. "Goal Setting and Monetary Incentives: When Large Stakes Are Not Enough." *Management Science* 61 (12): 2926–44. DOI: 10.1287/mnsc.2014.2068. [3, 5, 34]

Corgnet, Brice, Roberto Hernan-Gonzalez, et al. 2015. "Revisiting the Tradeoff between Risk and Incentives: The Shocking Effect of Random Shocks." [34]

Corgnet, Brice, Roberto Hernán-González, and Ricardo Mateo. 2015. "Cognitive Reflection and the Diligent Worker: An Experimental Study of Millennials." *PLoS ONE* 10 (11): DOI: doi:10.1371/journal.pone.0141243. [20]

Corgnet, Brice, Roberto Hernán-González, and Matthew W. McCarter. 2015. "The Role of the Decision-Making Regime on Cooperation in a Workgroup Social Dilemma: An Examination of Cyberloafing." *Games* 6 (4): 588–603. URL: http://www.mdpi.com/2073-4336/6/4/588. [24]

Corgnet, Brice, Roberto Hernan-Gonzalez, and Stephen Rassenti. 2015. "Peer Pressure and Moral Hazard in Teams: Experimental Evidence." [34]

Corgnet, Brice, Roberto Hernán-González, and Stephen Rassenti. 2015. "Firing Threats: Incentive Effects and Impression Management." *Games and Economic Behavior* 91: 97–113. DOI: http://dx.doi.org/10.1016/j.geb.2015.02.015. [34]

Corgnet, Brice, Roberto Hernán-González, and Eric Schniter. 2015. "Why Real Leisure Really Matters: Incentive Effects on Real Effort in the Laboratory." *Experimental Economics* 18 (2): 284–301. DOI: 10.1007/s10683-014-9401-4. [6, 10, 22, 31, 37]

Corgnet, Brice, Ludivine Martin, Peguy Ndodjang, and Angela Sutan. 2015. "On the Merit of Equal Pay: When Influence Activities Interact with Incentive Setting." [33]

Danilov, Anastasia, and Timo Vogelsang. 2016. "Time for Helping." *Journal of the Economic Science Association* 2 (1): 36–47. [5]

Deci, Edward L. 1971. "The Effects of Externally Mediated Rewards on Intrinsic Motivation." *Journal of Personality and Social Psychology* 18: 105–15. [3]

Deci, Edward L, James P Connell, and Richard M Ryan. 1989. "Self-determination in a work organization." *Journal of applied psychology* 74 (4): 580. [24]

Deci, Edward L, and R Ryan. 1995. "Intrinsic Motivation and Self-Determination in Human Behavior." *New York*, [24]

DellaVigna, S, and D Pope. 2018. "What Motivates Effort? Evidence and Expert Forecasts." *Review of Economic Studies,* 85 (2): 1029–69. [21, 41]

Dickinson, David L. 1999. "An Experimental Examination of Labor Supply and Work Intensities." *Journal of Labor Economics* 17 (4): 638–70. URL: http://www.jstor.org/stable/10.1086/209934. [6, 31]

Dohmen, Thomas, and Armin Falk. 2011. "Performance Pay and Multidimensional Sorting: Productivity, Preferences, and Gender." *American Economic Review* 101 (2): 556–90. DOI: 10.1257/aer.101.2.556. [5]

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G Wagner. 2011. "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences." *Journal of the European Economic Association* 9 (3): 522–50. [7, 20]

Eckartz, Katharina. 2014. "Task Enjoyment and Opportunity Costs in the Lab: The Effect of Financial Incentives on Performance in Real Effort Tasks." *Jena Economic Research Papers*, [6, 31]

Eriksson, Tor, Anders Poulsen, and Marie Claire Villeval. 2009. "Feedback and Incentives: Experimental Evidence." *Labour Economics* 16 (6): 679–88. DOI: 10.1016/j.labeco.2009.08.006. [35]

Erkal, Nisvan, Lata Gangadharan, and Boon Han Koh. 2018. "Monetary and non-monetary incentives in real-effort tournaments." *European Economic Review* 101: 528–45. DOI: https://doi.org/10.1016/j.euroecorev.2017.10.021. [6, 31]

Falk, Armin, and David Huffman. 2007. "Studying Labor Market Institutions in the Lab: Minimum Wages, Employment Protection, and Workfare." *Journal of Institutional and Theoretical Economics* 163 (1): 30–45. DOI: doi:10.1628/093245607780182044. [35]

Falk, Armin, and Michael Kosfeld. 2006. "The Hidden Costs of Control." *American Economic Review* 96 (5): 1611–30. [3]

Fischbacher, Urs. 2007. "z-Tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics* 10 (2): 171–78. [9]

Frederick, S. 2005. "Cognitive reflection and decision making." *Journal of Economic Perspectives* 19 (4): 25–42. URL: http://www.ingentaconnect.com/content/aea/jep/2005/00000019/00000004/art00002. [7, 20]

Gächter, Simon, Lingbo Huang, and Martin Sefton. 2016. "Combining "Real Effort" with Induced Effort Costs: the Ball-Catching Task." *Experimental Economics* 19 (4): 687–712. DOI: 10.1007/s10683-015-9465-9. [5]

Gill, David, and Victoria Prowse. 2012. "A Structural Analysis of Disappointment Aversion in a Real Effort Competition." *American economic review* 102 (1): 469–503. [4, 7]

Gill, David, and Victoria L Prowse. 2017. "Using Response Times to Measure Strategic Complexity and the Value of Thinking in Games." [20]

Gill, David, Victoria Prowse, and Michael Vlassopoulos. 2013. "Cheating in the Workplace: An Experimental Study of the Impact of Bonuses and Productivity." *Journal of Economic Behavior & Organization* 96: 120–34. [8]

Gneezy, Uri, Stephan Meier, and Pedro Rey-Biel. 2011. "When and Why Incentives (Don't) Work to Modify Behavior." *Journal of Economic Perspectives* 25 (4): 191–210. [5]

Gneezy, Uri, and Aldo Rustichini. 2000. "Pay Enough or Don't Pay at All." *Quarterly Journal of Economics* 115 (3): 791–810. [5]

Goerg, Sebastian J, Sebastian Kube, and Ro'i Zultan. 2010. "Treating Equals Unequally: Incentives in Teams, Workers' Motivation, and Production Technology." *Journal of Labor Economics* 28 (4): 747–72. [3]

Goerg, Sebastian, and Sebastian Kube. 2012. "Goals (th)at Work, Goals , Monetary Incentives , and Workers ' Performance." [3]

Hammermann, Andrea, and Alwine Mohnen. 2014. "The pric(z)e of hard work: Different incentive effects of non-monetary and monetary prizes." *Journal of Economic Psychology* 43: 1–15. DOI: 10.1016/j.joep.2014.04.003. [35]

Hayashi, Andrew T., Brent K. Nakamura, and David Gamage. 2013. "Experimental Evidence of Tax Salience and the Labor–Leisure Decision." *Public Finance Review* 41 (2): 203–26. DOI: 10.1177/1091142112460726. eprint: https://doi.org/10.1177/1091142112460726. [35]

Heckman, James J., and Tim Kautz. 2012. "Hard Evidence on Soft Skills." *Labour Economics* 19: 451–64. [20]

Herbst, Daniel, and Alexandre Mas. 2015. "Peer Effects on Worker Output in the Laboratory Generalize to the Field." *Science* 350 (6260): 545–49. [5]

Herweg, Fabian, Daniel Müller, and Philipp Weinschenk. 2010. "Binary Payment Schemes: Moral Hazard and Loss Aversion." *American Economic Review* 100 (5): 2451–77. DOI: 10.1257/aer.100.5.2451. [3]

Holmstrom, Bengt, and Paul Milgrom. 1987. "Aggregation and Linearity in the Provision of Intertemporal Incentives." *Econometrica: Journal of the Econometric Society* 55 (2): 303–28. [3]

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, & Organization* 7: 24–52. [4]

James, Harvey S. 2005. "Why Did You Do that? An Economic Examination of the Effect of Extrinsic Compensation on Intrinsic Motivation and Performance." *Journal of Economic Psychology* 26 (4): 549–66. [10]

Kajackaite, Agne. 2015. "If I Close my Eyes, Nobody Will Get Hurt: The Effect of Ignorance on Performance in a Real-Effort Experiment." *Journal of Economic Behavior & Organization* 116: 518–24. DOI: 10.1016/j.jebo.2015.05.020. [35]

Kessler, Judd B., and Michael I. Norton. 2016. "Tax Aversion in Labor Supply." *Journal of Economic Behavior and Organization* 124: 15–28. DOI: 10.1016/j.jebo.2015.09.022. [6, 36]

Kiessling, Lukas, Jonas Radbruch, and Sebastian Schaube. 2018. "The Impact of Self-Selection on Performance." *IZA Discussion Papers*, (11365): [24]

Koch, Alexander K., and Julia Nafziger. 2016. "Gift Exchange, Control, and Cyberloafing: A Real-Effort Experiment." *Journal of Economic Behavior & Organization* 131 (11): 409–26. DOI: http://dx.doi.org/10.1016/j.jebo.2016.09.008. [6, 10, 22, 24, 32, 37]

Koch, Alexander K, and Julia Nafziger. 2011. "Self-Regulation through Goal Setting." *Scandinavian Journal of Economics* 113 (1): 212–27. [3]

Kosfeld, Michael, and Susanne Neckermann. 2011. "Getting More Work for Nothing? Symbolic Awards and Worker Performance." *American Economic Journal: Microeconomics* 3 (3): 86–99. [3]

Kurzban, Robert, Angela Duckworth, Joseph W Kable, and Justus Myers. 2013. "An Opportunity Cost Model of Subjective Effort and Task Performance." *Behavioral and Brain Sciences* 36 (6): 661–79. [5, 48]

Larkin, Ian. 2014. "The Cost of High-Powered Incentives: Employee Gaming in Enterprise Software Sales." *Journal of Labor Economics* 32 (2): 199–227. URL: http://www.jstor.org/stable/10.1086/673371. [5]

Lazear, Edward P., and Paul Oyer. 2012. "Personnel Economics." In. *The Handbook of Organizational Economics*. Princeton University Press, 479–519. URL: http://www.jstor.org/stable/j.ctt1r2ggg.16. [4]

Lazear, P Edward. 2000. "Performance Pay and Productivity." *American Economic Review* 90 (5): 1346–61. DOI: 10.1257/aer.90.5.1346. [3]

Mohnen, Alwine, Kathrin Pokorny, and Dirk Sliwka. 2008. "Transparency, Inequity Aversion, and the Dynamics of Peer Pressure in Teams: Theory and Evidence." *Journal of Labor Economics* 26 (4): 693–720. [5, 36]

Murdock, Kevin. 2002. "Intrinsic Motivation and Optimal Incentive Contracts." *RAND Journal of Economics* 33 (4): 650–71. [10]

Nalbantian, Haig R, and Andrew Schotter. 1997. "Productivity Under Group Incentives: An Experimental Study." *American Economic Review* 87 (3): 314–41. [5]

Niederle, Muriel, and Lise Vesterlund. 2007. "Do Women Shy Away From Competition? Do Men Compete Too Much?" *Quarterly Journal of Economics* 122 (3): 1067–101. DOI: 10.1162/qjec.122.3.1067. [5]

Noussair, Charles, and Jan Stoop. 2015. "Time as a Medium of Reward in Three Social Preference Experiments." *Experimental Economics* 18 (3): 442–56. [5]

Ordóñez, Lisa D., Maurice E. Schweitzer, Adam D. Galinsky, and Max H. Bazerman. 2009. "Goals Gone Wild: The Systematic Side Effects of Overprescribing Goal Setting." *Academy of Management Perspectives* 23 (1): 6–16. DOI: 10.5465/AMP.2009.37007999. [5]

Pokorny, Kathrin. 2008. "Pay–But Do Not Pay too Much: An Experimental Study on the Impact of Incentives." *Journal of Economic Behavior & Organization* 66 (2): 251–64. [5]

Rammstedt, Beatrice, and Oliver P. John. 2007. "Measuring Personality in one Minute or Less: A 10-Item Short Version of the Big Five Inventory in English and German." [7, 20]

Roberts, John. 2007. *The Modern Firm: Organizational Design for Performance and Growth*. Oxford university press. [23]

Rosaz, Julie, Robert Slonim, and Marie Claire Villeval. 2016. "Quitting and Peer Effects at Work." *Labour Economics* 39: 55–67. DOI: http://dx.doi.org/10.1016/j.labeco.2016.02.002. [5, 6, 36]

Waber, Ben, Jennifer Magnolfi, and Greg Lindsay. 2014. "Workspaces that Move People." *Harvard Business Review* 92 (10): 68–77. [24]

Wagner, Julie, and Dan Watch. 2017. *Innovation Spaces: The New Design of Work*. Brookings Institute. [24]

Winter, Eyal. 2004. "Incentives and Discrimination." *American Economic Review* 94 (3): 764–73. DOI: 10.1126/science.151.3712.867-a. [3]

Chapter 2

Passive Choices and Cognitive Spillovers*

Joint with Steffen Altmann and Andreas Grunewald

2.1 Introduction

Passive behavior is a widespread phenomenon. In many situations, we can choose between various alternatives—yet we remain passive and do not take any decision. As a consequence, we stick to our current health insurance plans (Handel, 2013; Heiss, McFadden, Winter, Wuppermann, and Zhou, 2016), fail to cancel contracts with auto-renewal policies (DellaVigna and Malmendier, 2006), buy (partially) pre-configured products (Levav, Heitmann, Herrmann, and Iyengar, 2010), and do not make use of saving subsidies, tax benefits, or other social support programs (Chetty, Friedman, Leth-Petersen, Nielsen, and Olsen, 2014; Bhargava and Manoli, 2015; Finkelstein and Notowidigdo, 2018). Often, our passivity is associated with leaving money on the table compared to other available alternatives (see, e.g., Bhargava, Loewenstein, and Sydnor, 2017). In light of such observations, academics and policy makers have proposed a variety of interventions to foster active decision-making. To overcome passivity, we remind people of decisions that are to be taken (Altmann and Traxler, 2014; Calzolari and Nardotto, 2016; Karlan, McConnell, Mullainathan,

* Financial support from Volkswagen Foundation is gratefully acknowledged. We thank participants of the Grüneburg Seminar in Frankfurt for helpful comments and suggestions. We also benefited from comments by seminar and conference participants of the Workshop on Passive Choices at the University of Copenhagen, the 2018 Sloan Nomis Workshop on the Cognitive Foundations of Economic Behavior, the 20th Colloquium on Personnel Economics in Zurich, the Bonn Mannheim PhD Workshop 2017, the Spring Meeting of Young Economists in Halle, the 3rd Maastricht Behavioral Economic Policy Symposium, the Annual Conference of the German Economic Association 2017, the Max Planck Institute for Research on Collective Goods, the CPB Netherlands Bureau for Economic Policy Analysis, and the University of Mainz.

and Zinman, 2016; Damgaard and Gravert, 2018), provide them with information (Kling, Mullainathan, Shafir, Vermeulen, and Wrobel, 2012; Fellner, Sausgruber, and Traxler, 2013; Kaufmann, Müller, Hefti, and Boes, 2018), impose deadlines (Heffetz, O'Donoghue, and Schneider, 2016; Altmann, Traxler, and Weinschenk, 2017), or force them to make active decisions (Carroll, Choi, Laibson, Madrian, and Metrick, 2009; Stutzer, Goette, and Zehnder, 2011).

In this paper, we study the role of scarce cognitive resources as a source of passive behavior. Cognitive resources are fundamental for any economic decision. Making choices requires us to pay attention, process information, and evaluate trade-offs between the available alternatives. A growing body of literature documents that our resources for carrying out these tasks are inherently limited (see, e.g., Caplin, Dean, and Martin (2011), Mullainathan and Shafir (2013), and Gabaix (2017) for a comprehensive overview). Moreover, we commonly face multiple tasks or judgments that require our attention simultaneously. This in turn may further curtail the resources available for each of the judgments and, as a result, diminish our propensity to make active decisions. While the link between cognitive resources and passive behavior is intuitively plausible, a number of questions are not well understood. We focus on three of them that appear particularly important. First, does cognitive resource scarcity lead to an increase in passive decision-making? Second, how do interventions that foster active decision-making affect the choices of individuals when cognitive resources are scarce versus abundant? Third, does fostering active choice in one domain reduce the amount of cognitive resources devoted to others, i.e., do choice-promoting policies lead to negative "cognitive spillovers" on other decisions?

We study these questions in a controlled laboratory setting. Three features of our experiment make it ideally suited towards this end. First, to identify the causal impact of cognitive resource scarcity on passive behavior, we can exogenously vary the scarcity of individuals' cognitive resources across different treatments of our experiment. Second, we can gather information on the cognitive resources underlying individuals' decisions, shedding light on the mechanisms through which cognitive resource scarcity affects passivity. Third, we can assess the consequences of fostering active choice in the targeted domain, but also in terms of potential cognitive spillovers in other decision domains.

Participants in our experiment work on two tasks simultaneously—a "background task" and a "decision task". For the background task, subjects memorize and recall numbers, requiring them to bring up cognitive resources. In the first treatment dimension, we manipulate how demanding the background task is. In doing so, we exogenously vary whether subjects' cognitive resources are scarce or ample (denoted as Scarce and Ample condition, respectively). We then examine how the induced difference in cognitive resource scarcity affects participants' inclination to stay passive in the decision task, in which they have to find the correct solution to simple math problems with three possible solutions. If individuals do not actively choose an option in the decision task, a randomly selected default governs their

choices. Participants' propensity to stick to the default option gives us a direct measure of passivity. Moreover, the default option provides a natural opportunity for participants in our experiment to abstain from devoting any cognitive resources to the decision task, and rather focus on the background task alone. A key feature of our experiment is that we can readily measure this allocation of cognitive resources. Specifically, in our Baseline environment, participants enter the decision task by pressing a button on the keyboard. If they do not hold the corresponding button, they face a blank screen. This feature allows us to track whether subjects attend to the decision task at all, and how much time they dedicate to the task—the amount of visual attention they allocate to the decision task.

The data from our experiment demonstrate that cognitive resource scarcity causes a strong increase in passive decision-making. Participants in the Baseline-Ample condition predominantly decide actively and stick to the default option only as often as it is expected to be the correct choice (32%). In contrast, subjects in Baseline-Scarce remain passive significantly more often, following the default in 60% of the cases. Hence, scarcity of cognitive resources leads to an increase in passivity. We further show that the strong treatment difference in behavior is attributable to a re-allocation of cognitive resources. In particular, under cognitive resource scarcity, subjects shift their attention away from the decision task. This shift happens both at the extensive and intensive margin. If cognitive resources are scarce, subjects completely disregard the decision task in about 32.0% of cases, while they do so in only 2.5% of cases in Baseline-Ample. In the same spirit, subjects in Baseline-Scarce also devote less cognitive resources to the decision task, conditional on paying any attention to the task.

In the second part of our experiment, we study how choice-promoting interventions—i.e., policies that encourage active decision-making—affect individuals' choices and the allocation of cognitive resources. We study behavior in two additional decision environments. These capture essential features of commonly observed policies to help people overcome passivity, by directing their attention to a particular task or decision (Kling et al., 2012; Calzolari and Nardotto, 2016; Karlan et al., 2016), or by asking or effectively forcing them to make an active decision (Carroll et al., 2009; Stutzer, Goette, and Zehnder, 2011). Specifically, in the Directed Attention environment, we steer participants' attention to the decision task by permanently displaying the task on their screen. We do, however, still allow for passive behavior, by leaving one option preselected as the default. In contrast, the decision task in the Active Choice environment features no default option. Subjects in this environment are, thus, required to take an active decision.

The Directed Attention and Active Choice environments, therefore, allow us to examine the behavioral consequences of choice-promoting interventions, and to study how the effects of the policies depend on the relative scarcity of individuals' cognitive resources. In particular, we would expect the influence of the interventions on the allocation of cognitive resources—and, hence, choices—to be particularly

pronounced under cognitive resource scarcity. This is indeed what we observe: when cognitive resources are scarce, passive behavior decreases from 60% in Baseline-Scarce to 41% in Directed-Scarce, while the rate of passive choices is almost identical in Baseline-Ample (32%) and Directed-Ample (30%). At the same time, we find that passive behavior under cognitive resource scarcity is not eliminated entirely by the Directed Attention intervention. Relative to the Active Choice environment, a choice alternative that is preselected as default in the Directed Attention environment is about 10 percentage points more likely to be chosen than the identical (non-default) alternative in the Active Choice environment.

Last but not least, our experiment allows us to examine how choice-promoting interventions affect the quality of individuals' decisions. In both the Directed Attention and Active Choice environment, we observe that the higher frequency of active decisions comes along with an improvement in the quality of individuals' choices in the decision task, relative to the Baseline condition. If cognitive resources are scarce, however, inducing active choice in some decision domain might also lead to a reduction in cognitive resources devoted to other domains, with potentially negative consequences for decisions in the latter. Our experiment has the unique feature that we can readily measure whether such cognitive spillovers occur. Our data show that choice-promoting interventions can indeed have detrimental effects on other decisions: relative to the Baseline environment, both the Directed Attention and Active Choice policy impair the quality of individuals' decisions in the background task. Indeed, in our experiment, these negative spillovers on the background task completely offset the observed gains in the decision task. As a result, subjects' overall payoffs do not differ across decision environments.

Our findings contribute to the literature that studies the cognitive and perceptual foundations of passive behavior. By establishing a direct causal link between a person's available cognitive resources and passive behavior, we advance the literature that has studied the relationship between cognitive resources and passivity, using proxies of cognitive capacity such as financial literacy (Brown, Farrell, and Weisbenner, 2011; Rooij and Teppa, 2014), self-rated knowledge of the decision situation (Levav et al., 2010), or exhaustion of decision makers (Danziger, Levav, and Avnaim-Pesso, 2011). Our paper also complements findings by Caplin and Martin (2016) and Caplin and Martin (2017), who show that better defaults lead to more passive behavior, indicating that individuals devote fewer cognitive resources to decisions that feature high-quality defaults.

Cognitive resource scarcity as a driver of passive behavior might be aggravated by scarcity of other resources. There is, for instance, an ongoing discussion whether concerns about financial resources, hunger, and other aspects of poverty induce a "tax" on individuals' cognitive resources or bandwidth (Mani, Mullainathan, Shafir, and Zhao, 2013; Mullainathan and Shafir, 2013; Carvalho, Meier, and Wang, 2016; Shah, Mullainathan, and Shafir, 2018; Sharafi, 2018). To the extent that this is the case, our results may also shed further light on the behavioral consequences of

poverty. For example, the finding that defaults are more sticky among subjects with scarce cognitive resources suggests that the correlation between household income and default adherence that has been observed in cross-sectional data (Brown, Farrell, and Weisbenner, 2011; Bhargava, Loewenstein, and Sydnor, 2017) might, at least partially, work through a reduction in bandwidth among subjects with lower financial resources. When seen through this lens, our results also suggest that well-chosen defaults can yield a double dividend for (financially or cognitively) deprived parts of the population. They do not only mechanically improve outcomes for passive individuals, but they may also "free up" cognitive resources that are sorely needed for other tasks. Conversely, however, cognitive resource scarcity is also likely to make individuals more susceptible to being exploited by "bad" defaults imposed by parties with misaligned interests, e.g., firms attempting to sell particular preconfigured products.

On a more general level, our results suggest that choices in decision domains that are typically evaluated in isolation should be considered jointly when they compete for a person's cognitive resources. This is of particular interest for the evaluation of "nudges" and behavioral policy interventions. Many of these policies remind people of upcoming tasks and available choice options (Altmann and Traxler, 2014; Calzolari and Nardotto, 2016; Karlan et al., 2016) or provide additional information (Kling et al., 2012; Tiefenbeck, Goette, Degen, Tasic, Fleisch, et al., 2016; Kaufmann et al., 2018), thereby drawing subjects' attention to one particular decision. Our findings inform the design and evaluation of such policies in two ways. First, they help to understand for whom the interventions are likely to have the strongest effects: individuals whose available cognitive resources are limited (e.g., because they have a low stock of cognitive resources or face multiple demanding tasks simultaneously) are more likely to remain blissfully ignorant about some decisions and, hence, they are also more likely to be affected by interventions that redirect their attention. Second, our results indicate that evaluating choice-promoting interventions solely based on individuals' decisions in the targeted choice domain may not suffice to demonstrate the interventions' usefulness: Directing individuals' cognitive resources to one choice domain may come at the cost of negative cognitive spillovers on other domains, which ultimately could lead to worse outcomes overall.

By showing that a "nudge" in one decision domain can affect the quality of choices in other domains, we also add to the literature that warns about possible unintended consequences of behavioral policy interventions. First indications for such unintended consequences have been established by research in psychology showing that directing people's attention to one task may induce them to overlook other, unexpected events (e.g., Simons and Chabris (1999)). It has also been documented that libertarian paternalistic interventions can backfire as they may impair individual (Caplin and Martin, 2016; Haan and Linde, 2017) or social learning about the decision environment (Carlin, Gervais, and Manso, 2013), or as firms' strategic responses limit the effectiveness of the intervention (Duarte and Hastings, 2012).

To understand whether differences in passivity are driven by differences in the allocation of cognitive resources, we enrich our behavioral data with measures of the attentional processes underlying subjects' choices. From a methodological perspective, our paper thus stands in the tradition of studies that use related process-tracing methods, such as Mouselab (e.g., Johnson, Payne, Schkade, and Bettman, 1989; Gabaix, Laibson, Moloche, and Weinberg, 2006), eye-tracking (e.g., Wang, Spezio, and Camerer, 2010), or data on search processes (Caplin, Dean, and Martin, 2011). By providing insights on the factors that shape individuals' allocation of attention, our results can also inform a growing literature that theoretically explores the behavioral implications of limited attention and the determinants of attention allocation (Bordalo, Gennaioli, and Shleifer, 2012; Kőszegi and Szeidl, 2012; Bordalo, Gennaioli, and Shleifer, 2013; Gabaix, 2014; Mackowiak, Matějka, and Wiederholt, 2018), and tests the corresponding results in the lab (Caplin and Dean, 2013; Dertwinkel-Kalt, Gerhardt, Riener, Schwerter, and Strang, 2016; Dean and Neligh, 2017; Martin, 2017; Nielsen, Sebald, and Sørensen, 2018) or field (Bartoš, Bauer, Chytilová, and Matějka, 2016).

The remainder of the paper is organized as follows. In the next section, we present the design of our experiment. Section 2.3 discusses behavioral hypotheses for the different treatments. Section 2.4 presents our empirical results and Section 2.5 concludes.

2.2 Design of the Experiment

The goal of our experiment is to study the impact of cognitive resource scarcity on passive behavior. For this purpose, we set up a stylized decision environment that captures two key features of situations in which people commonly stay passive. First, individuals regularly have to juggle various tasks or decisions simultaneously. For instance, they prepare an important meeting with clients at work, take care of their kids, visit their doctor for a check-up appointment, choose a restaurant for a family dinner, and additionally decide on their next mobile phone contract and health care plan. Each of the tasks requires some cognitive resources. If these are scarce, people might make some decisions based on a cursory first glance, or disregard them entirely and remain passive. Second, in many decision environments it is specified what happens if people stay passive—i.e., there are explicit or implicit defaults that prevail unless a decision maker actively decides otherwise.

To capture these features in a laboratory setting, we implemented a simple decision environment in which participants are simultaneously confronted with two tasks—a "background task" and a "decision task". The background task functions as an abstract representation of the bundle of tasks and choices that a decision maker has to handle, with the exception of the decision task. For the implementation of the background task, we build on a well-established paradigm from cognitive psy-

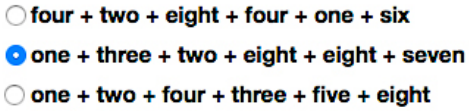


Figure 2.1. Example of a Decision Task

chology for which it is straightforward to manipulate the level of difficulty and, hence, the amount of cognitive resources required to solve the task correctly (see, e.g., Sprenger, Dougherty, Atkins, Franco-Watkins, Thomas, et al., 2011; Carpenter, Graham, and Wolf, 2013; Huh, Vosgerau, and Morewedge, 2014; Deck and Jahedi, 2015). Specifically, as *background task*, subjects in the experiment had to memorize and recall numbers. At the beginning of each round of the experiment, a new number was displayed for 10 seconds on subjects' screens. Subsequently, the number disappeared and subjects had to keep it in mind. After another 30 seconds, subjects had to type in the memorized number in a field on their screen, earning €0.40 if their answer was correct and €0 otherwise. We made sure that subjects had no opportunity to write down the numbers of the background task: they had no access to scratch paper and had to hand over their mobile phones for the duration of the experiment.

During the 30 seconds in which they had to keep the number from the background task in mind, subjects additionally faced the *decision task*. Specifically, subjects were presented with three summations, each of which consisted of six addends. Their task was to decide which of the three options yielded the highest sum (see Figure 2.1 for an example). Subjects earned €0.10 for a correct answer and €0 otherwise. The decision task featured a default option that was implemented if subjects did not make an active decision. In particular, in each round, one randomly selected option of the decision task was displayed as the default choice (cp. the middle option in Figure 2.1). Subjects were informed about the existence of a default and that it was randomly determined which option was the default in a given round of the experiment.

We designed the decision task to resemble multi-attribute choices that feature a payoff-maximizing option, the identification of which requires cognitive resources (e.g., finding the cheapest health-care plan for a known expected demand profile; cp. Caplin, Dean, and Martin (2011) and Kaufmann et al. (2018)). More broadly, our task can also be thought of as representing menu choices between products or services with uncertain value. Such uncertainty might arise if individuals need to inquire the different attributes of the product and/or their own preferences for the attributes. A consumer who decides about a retirement savings plan, for example, might need to bring up cognitive resources to learn about the risk/return profiles of

Table 2.1. Treatment Overview

	Baseline	Directed Attention	Active Choice
Ample	Baseline-Ample	Directed-Ample	Active-Ample
Scarce	Baseline-Scarce	Directed-Scarce	Active-Scarce

the assets included in the plan, and his own valuation of the respective fluctuations in wealth.

Two points are worth noting regarding our choices of parameters in the experiment. First, to hold the difficulty of the decision task roughly constant across different rounds of the experiment, each option resulted in a sum between 20 and 34. Second, the rewards for correctly solving the background task are relatively high compared to the ones for the decision task. We opted for this parameter constellation to ensure "treatment take-up", i.e., to make sure that subjects tried to correctly solve the background task, even if this was demanding.

2.2.1 Treatments

We implemented a 3x2 between-subjects design. In the first treatment dimension, we exogenously varied the amount of cognitive resources needed to solve the background task. Subjects in treatments with Scarce cognitive resources had to memorize seven-digit numbers in the background task. In contrast, the numbers to memorize in treatments with Ample resources had only two digits. Solving the background task in the Ample condition thus essentially requires zero cognitive resources, whereas the more difficult task in the Scarce condition will induce cognitive resource scarcity. Comparing behavior between the two conditions thus allows us to identify the causal impact of cognitive resource scarcity on subjects' propensity to stay passive in the decision task.

In the second treatment dimension, we varied the characteristics of the decision environment faced by participants. In particular, we study three decision environments, denoted as Baseline, Directed Attention, and Active Choice environment (cp. Table 2.1). These environments differed only in how the decision task was displayed to subjects. In the Baseline environment, subjects faced a blank screen after the number to memorize for the background task had disappeared.[1] To access the decision task, they had to press and hold a key on their keyboard. The blank baseline screen only contained information on which key subjects had to hold in order to see the decision task. If they released the key, the decision task disappeared

1. A translated version of all screens can be found in Figure 2.B.2a–2.B.2d in 2.B.

and subjects returned to the blank screen again. Subjects were informed about this procedure in advance.

This feature of the Baseline environment allows for the possibility that subjects may not devote any cognitive resources to the decision task, in line with the idea that passivity might be triggered by individuals not even entering "decision mode" (see Sunstein, 2014; Heiss et al., 2016). Furthermore, we can directly track whether subjects in the Baseline environment entered the decision task in a given round of the experiment, i.e., whether they pressed the key at least once. We can therefore distinguish whether subjects remained passive because they completely ignored the decision task, or whether they followed the default despite paying attention to the task. Moreover, for a subsample of participants in the Baseline environment, we additionally gathered detailed information on the precise length of the time spans in which subjects attended to the decision task.[2] These attention spans provide us with an intensive-margin measure of the amount of cognitive resources that subjects in Baseline-Scarce and Baseline-Ample allocate to the decision task.[3]

The two remaining decision environments, Directed Attention and Active Choice, were designed to investigate how encouraging active decision-making affects individuals' behavior. These environments mirror some core features of commonly observed policies that aim to reduce passivity by directing people's attention to a particular decision or task (e.g., Altmann and Traxler, 2014; Calzolari and Nardotto, 2016; Karlan et al., 2016) or forcing them to make an active decision (e.g., Carroll et al., 2009; Stutzer, Goette, and Zehnder, 2011). In the Directed Attention environment, we steered participants' attention to the decision task by permanently displaying the task on their screen. Hence, we impaired subjects' ability to completely disregard the decision task. The Directed Attention environment, however, still allowed for passive behavior—the decision task involved a (randomly) preselected default option. The latter feature was removed in the Active Choice environment, in which the decision task was also displayed permanently, but none of the options was preselected. Hence, subjects had to actively choose one of the op-

2. We elicited this enriched attention data in 50% of sessions of the Baseline environment (balanced across the Baseline-Scarce and Baseline-Ample condition). Subjects were not aware that their attention spans were recorded.

3. Dean, Schilbach, and Schofield (2017) define attention as the "ability to focus on particular pieces of information by engaging in a selection process that allows for further processing of incoming stimuli.[..]". Attention is, thus, one part of cognitive functioning. Solving the decision task in our experiment, however, also involves memory and higher-order cognitive functions. Attention spans therefore only provide a proxy of the total amount of resources allocated to the problem. Yet, as attention is a necessary prerequisite to solve the decision task, it moderates other cognitive resources in the decision process. In particular, devoting zero visual attention to the decision task is analogous to not devoting cognitive resources to the task.

tions in the decision task.[4] To study how the interplay between the choice environment and cognitive resource scarcity shapes behavior, we implemented treatments with AMPLE and SCARCE cognitive resources for all decision environments. This leaves us with a total of 6 different treatment cells (see Table 2.1 for an overview).

2.2.2 Procedures

Each session of the experiment consisted of four parts. In the first and second part, we familiarized subjects with the background task and decision task, respectively. In addition, these parts of the experiment also serve as a validation check to make sure that there are no systematic treatment differences in subjects' ability of solving the tasks. The first part consisted of ten rounds in which subjects had to memorize numbers of varying difficulty (from 5 to 9 digits). As in the main experiment, numbers were displayed for 10 seconds and subjects had to keep them in mind for 30 seconds. Subjects earned €0.40 if they correctly recalled the number and €0 otherwise. In the second part of the experiment, subjects worked on the decision task (but no background task) for ten rounds. Each round lasted 30 seconds and subjects earned €0.10 per correct answer. The third and main part of the experiment consisted of 20 rounds in which subjects simultaneously faced the background task and decision task, as described above. Only after the end of the third part, subjects received feedback on their performance in the different parts of the experiment. The experiment ended with a short post-experimental questionnaire. Table 2.C.1 in the appendix summarizes descriptive statistics and balancing checks for the baseline ability measures from Phase 1 and 2 and a number of sociodemographic characteristics.

At the beginning of each part of the experiment, subjects received written on-screen instructions explaining the rules and details of the corresponding part.[5] In all rounds of the experiment, subjects could never leave a screen by themselves, but were automatically forwarded to the next screen when the time for a given screen had elapsed. The tasks, numbers, defaults, and their order were identical across all subjects and treatments. Furthermore, to eliminate potential session-level effects in the corresponding treatment comparisons (see Fréchette, 2012), we randomized between individuals within a given session whether cognitive resources were SCARCE or AMPLE.

The experiments were conducted in the BonnEconLab at the University of Bonn, implemented with Otree (Chen, Schonger, and Wickens, 2016), and the online recruitment system by Bock, Baetge, and Nicklisch (2014). A total of 564 subjects participated in our experiment. We conducted 8 sessions each for the BASELINE,

4. In 5.35% of cases, subjects in the ACTIVE CHOICE environment nevertheless remained passive and did not choose any of the three options in the decision task. As they had not picked the correct solution, subjects' earnings for the decision task were €0 in these cases.

5. A translation of the instructions of the experiment can be found in 2.B.

Directed Attention, and Active Choice environment, corresponding to approximately 96 subjects in each treatment cell (cp. Table 2.C.1 in the appendix). On average, sessions lasted 75 minutes. Subjects' mean earnings in the experiment were €16.53, including a show-up fee of €4.

2.3 Behavioral Predictions

In what follows, we discuss how we expect cognitive resource scarcity to influence passive behavior. Our behavioral predictions are informed by an illustrative theoretical framework, which builds on the premise that individuals have a limited stock of cognitive resources and optimally allocate these resources across tasks. While both of these assumptions naturally provide a simplistic perspective on individuals' decision processes in the experiment, they help to structure thoughts about the behavioral consequences of cognitive resource scarcity. We discuss the theoretical framework in more detail in 2.A, where we also derive a set of conditions under which the following behavioral hypotheses hold. Here, we focus our attention on the intuitions leading to the hypotheses.

A key building block of our experiment is that participants face two tasks, both of which require cognitive resources to be solved. Formally, individual j is endowed with the fixed stock of cognitive resources X^j and solves the following decision problem:

$$\begin{aligned} &\max_{x_B, x_D} u(x_B, x_D) \;=\; \pi_B(x_B)u_B + \pi_D(x_D)u_D \\ &s.t.\;\; x_B + x_D \;\leq\; X^j \end{aligned} \tag{2.1}$$

where x_B, x_D are the cognitive resources devoted to the background task (B) and decision task (D), respectively, $\pi_B(\cdot)$ [$\pi_D(\cdot)$] denotes the probability of solving task B [D] correctly, and u_B, u_D denote the individual's payoffs from solving task B and D, respectively.

As discussed in Section 3.2, our experiment design rests on the idea that memorizing two-digit numbers essentially requires zero cognitive resources. We conceptualize this idea by setting $\pi_B(x_B) = 1$ for all $x_B \geq 0$ in the Ample conditions. Individuals in Baseline-Ample can thus solve the background task even with minimal cognitive resources and, consequently, they should dedicate all available resources to the decision task. The data from our experiment allow for a straightforward test of whether subjects indeed do. Specifically, we can compare the frequency at which subjects in Baseline-Ample correctly solve the decision task in the main part of the experiment (83.6%) to the corresponding number in the second phase of the experiment (86%), in which they work on the decision task, but face no background task. The difference between the two frequencies is small and not statistically signif-

icant (Wilcoxon signed-rank test, p=0.308), lending support to the notion that the background task in Baseline-Ample requires no cognitive resources.[6]

In contrast, individuals in Baseline-Scarce face a trade-off when allocating cognitive resources between the background and decision task: dedicating more resources to the background task increases the probability of solving this task, but comes at the costs of allocating fewer resources to the decision task, with resulting negative consequences for the probability of solving the latter. The optimal solution to this trade-off will naturally depend on the overall stock of cognitive resources available to an individual. Intuitively, when cognitive resources are scarce, subjects with a relatively small stock of cognitive resources will find it profitable to completely ignore the decision task and instead allocate all their available resources to the (more highly incentivized) background task. As a result, we expect an extensive-margin reduction in cognitive resources devoted to the decision task, relative to the Baseline-Ample environment: while all subjects devote cognitive resources to solving the decision task in Baseline-Ample, some subjects will find it optimal not to pay any attention to the decision task when cognitive resources are scarce. For subjects with a relatively large stock of cognitive resources, in turn, it will still be optimal to devote some resources to the decision task. These subjects thus divide their resources between both tasks in Baseline-Scarce, whereas they devote all resources to the decision task in Baseline-Ample. For subjects with a relatively large stock of cognitive resources, we therefore expect an intensive-margin reduction in cognitive resources allocated to the decision task, relative to Baseline-Ample. Both effects imply that fewer resources are allocated to the decision task when cognitive resources are scarce. Hence, the cumulative distribution of cognitive resources devoted to the decision task in Baseline-Scarce should first-order stochastically dominate the one in Baseline-Ample.

Hypothesis 1. *The fraction of subjects who devote no cognitive resources to the decision task is larger in Baseline-Scarce than in Baseline-Ample. The cumulative distribution of cognitive resources that subjects allocate to the decision task in Baseline-Scarce first-order stochastically dominates the corresponding distribution in Baseline-Ample.*

Next, we consider how these differences in cognitive resource allocation affect the degree of passive behavior across treatments. We expect that both the extensive- and intensive-margin reduction in cognitive resources dedicated to the decision task will result in more passive behavior. It is directly apparent that individuals who decide not to pay any attention to the decision task—i.e., those who show an extensive-margin reaction—will automatically follow the default. Furthermore, individuals who devote positive, but lower amounts of cognitive resources to the decision task will stay passive more often, as long as (i) a reduction in cognitive resources leads

6. Furthermore, in roughly 98% of cases, subjects in Baseline-Ample also solve the background task correctly (cp. Table 2.2 in Section 2.4).

to a higher likelihood of making mistakes (a lower $\pi_D(\cdot)$ in our framework) and (ii) individuals who dedicate fewer resources to a task are more likely to follow an ill-specified default than to actively opt for another wrong option (see 2.A for further details). As a result of both the extensive- and intensive-margin reduction in cognitive resources devoted to the decision task, we thus expect higher rates of passive behavior in Baseline-Scarce as compared to Baseline-Ample.

The above arguments also imply that the difference in passive behavior between Baseline-Scarce and Baseline-Ample should be particularly pronounced in situations in which the default is incorrect. In this case, the reduction in cognitive resources devoted to the decision task increases the rate of passive behavior in Baseline-Scarce for two reasons. First, some individuals completely disregard the decision task and stay passive. Second, individuals who partially reduce the cognitive resources allocated to the decision task are more likely to make a wrong decision and (mistakenly) follow the incorrectly specified default. In contrast, in situations in which the default is correct, individuals who completely disregard the task still stay passive, whereas a partial reduction in cognitive resources might induce people to opt out too frequently of a correctly specified default. Which of the two effects dominates is, ex ante, unclear. Hence, differences in default adherence between Baseline-Scarce and Baseline-Ample should be especially strong for situations involving "bad" defaults.

Hypothesis 2. *Passive decision making is more pronounced in Baseline-Scarce than in Baseline-Ample. Likewise, passivity rates are higher in Baseline-Scarce than in Baseline-Ample if the default option is incorrect. If the default option is correct, the treatment comparison is ambiguous.*

Our framework also provides a natural setting to examine how the choice-promoting interventions in the Directed Attention and Active Choice environments affect the allocation of cognitive resources, and the resulting decisions. We assume that, as a result of the interventions, some amount of cognitive resources x_T—which depends on treatment $T \in \{Directed, Active\}$—is exogenously directed towards the decision task.

As subjects in the Ample environment devote all available resources to the decision task anyways, the interventions will not distort their allocation of cognitive resources. We thus predict that there should be no systematic differences between Active-Ample, Directed-Ample, and Baseline-Ample in terms of cognitive resource allocation, subjects' behavior, and the resulting quality of their decisions. In contrast, in Directed-Scarce and Active-Scarce, the additional constraint ($x_D \geq x_T$) will be binding for subjects who would otherwise allocate no or only few cognitive resources to the decision task. Relative to the Baseline-Scarce treatment, these subjects will increase the amount of cognitive resources devoted to the decision task. Following the same arguments as above, we should thus observe lower rates of passivity in Directed-Scarce compared to Baseline-Scarce. Nevertheless,

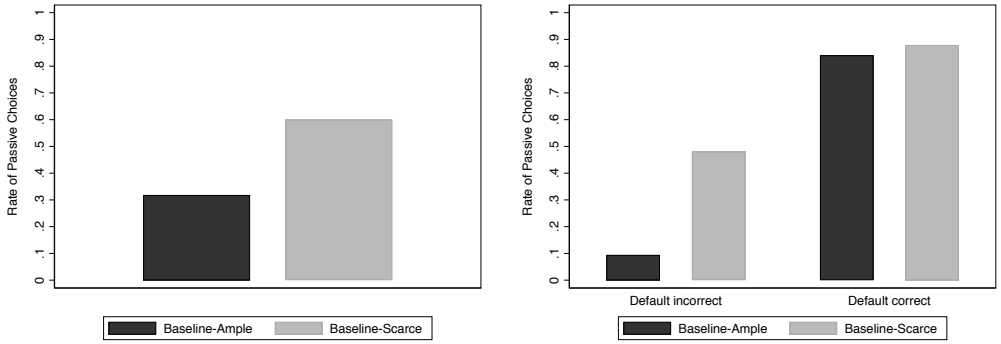
the defaults in Directed-Scarce will attract choices at a higher-than-random frequency (see 2.A). As a result, we expect subjects in Directed-Scarce to choose the default option more often than they choose the corresponding choice alternative in Active-Scarce.

Hypothesis 3. *There are no differences in behavior between Baseline-Ample, Directed-Ample, and Active-Ample. Under cognitive resource scarcity, subjects stay passive more often in Baseline-Scarce than in Directed-Scarce. The options which are displayed as defaults are chosen more frequently in Directed-Scarce than in Active-Scarce.*

Last but not least, we can extend our analysis to examine the consequences of our treatment interventions for the quality of subjects' decisions. In line with the arguments above, the Directed Attention and Active Choice interventions should only affect subjects in treatments with Scarce resources. It is also directly apparent that the decision problem of subjects in the Active and Directed environments is a constrained version of the one in Baseline. As a consequence, average payoffs should be weakly higher in Baseline-Scarce compared to Directed-Scarce and Active-Scarce. This overall effect is composed of two countervailing sub-effects. On the one hand, subjects who devote no or only few resources to the decision task in Baseline-Scarce experience an increase in resources allocated to that task. Hence, we expect the quality of choices in the decision task to be higher in Directed-Scarce and Active-Scarce than in Baseline-Scarce. On the other hand, as the cognitive-resource constraint is binding, the exogenous reallocation of resources to the decision task forces subjects to withdraw scarce cognitive resources from the background task. This reallocation should have negative consequences for the quality of subjects' decisions in the background task. As a consequence of this cognitive spillover, we expect decision quality in the background task to be higher in Baseline-Scarce than in Directed-Scarce and Active-Scarce.

Hypothesis 4. *There are no differences in decision quality between Baseline-Ample, Directed-Ample, and Active-Ample. Relative to Baseline-Scarce, overall payoffs should be weakly lower in Directed-Scarce and Active-Scarce. While performances in the decision task are better in Directed-Scarce and Active-Scarce than in Baseline-Scarce, the opposite holds true for performances in the background task.*

To summarize, our framework provides four main predictions for subjects' behavior in the experiment: (1) A more demanding background task causes subjects to withdraw cognitive resources from the decision task at the extensive and intensive margin; (2) This reduction in cognitive resources leads to a higher rate of passive decision-making when cognitive resources are scarce; (3) Directing subjects' attention to the decision task or implementing an active-choice environment reduces passivity in the decision task; (4) Both interventions improve subjects' performance in the decision task, but they also cause a withdrawal of cognitive resources from the background task, leading to negative cognitive spillovers and worse performance in



(a) Overall default adherence **(b)** Default adherence conditional on default quality

Figure 2.2. Passive Behavior in BASELINE

Notes: Panel (a) depicts average default adherence rates in BASELINE-AMPLE and BASELINE-SCARCE. Panel (b) shows default adherence rates in BASELINE-AMPLE and BASELINE-SCARCE conditional on quality of the default option.

this domain. The discussion of our empirical results in the following sections will be structured according to these main predictions.

2.4 Results

2.4.1 Cognitive Resource Scarcity and Passive Choices

To analyze how scarcity of cognitive resources affects passivity, we first examine how frequently subjects stick to the default option in the decision task in the BASELINE environment. Panel (a) of Figure 4.1 compares default adherence rates between the BASELINE-SCARCE and BASELINE-AMPLE condition. In line with the first part of Hypothesis 2, the figure shows a striking increase in default adherence under cognitive resource scarcity. While subjects stick to the default in only 31.8% of cases in BASELINE-AMPLE, the default adherence rate increases to 60.0% in BASELINE-SCARCE.[7] The difference in default adherence is highly significant (Mann-Whitney-U test, $p < 0.01$).[8]

Panel (b) of Figure 4.1 further shows that the treatment difference between BASELINE-AMPLE and BASELINE-SCARCE is predominantly driven by an increase in default adherence in situations in which sticking to the default is a "bad" choice.

7. Note that the randomly determined default option ended up being correct in 6 out of 20 rounds, i.e., in 30% of cases.

8. Unless otherwise noted, all non-parametric tests are based on subject-level averages across the 20 rounds of the experiment. The reported parametric tests, which are based on observations at the individual subject-round level, account for potential clustering at the subject level. Reported p-values are always two-sided.

Specifically, the figure depicts default adherence rates separately for situations in which the (randomly determined) default option did versus did not coincide with the correct solution to the decision task. We find only small differences in default adherence rates if the default option corresponds to the correct solution; subjects stick to the default in 84.0% (Baseline-Ample) versus 87.9% (Baseline-Scarce) of cases (Mann-Whitney-U test, $p = 0.069$). In contrast, we observe a strong divergence in the rate of passive choices if the stipulated default option is incorrect. In this case, subjects follow the default in 48.1% of cases when cognitive resources are scarce. This compares to only 9.4% of cases in Baseline-Ample. Supporting the second part of Hypothesis 2, the difference across treatments is statistically significant (Mann-Whitney-U test, $p < 0.01$). The finding that subjects are substantially more prone to stick to "bad" defaults in Baseline-Scarce is a first indication that the high rate of passive choices in this treatment is indeed driven by a reduction of cognitive resources devoted to the decision task: when the background task is more demanding, subjects remain passive in situations where an active choice would improve their decisions, but would also require them to spend scarce cognitive resources. We will return to this point in Section 2.4.2 below.

Result 1. *Scarcity of cognitive resources causes an increase in passive decision-making. Subjects who face a cognitively more demanding background task are significantly less likely to make active choices in the decision task.*

The strong overall treatment difference raises the question whether certain groups of participants exhibit stronger increases in passivity than others. A natural source of heterogeneity to consider in this respect is individuals' "stock" of cognitive resources. Intuitively, subjects with abundant cognitive resources might be less affected by the exogenously induced scarcity in Baseline-Scarce than subjects with lower cognitive capacity. This is indeed what we observe. As a proxy for participants' stock of cognitive resources, we use their performance in a short test for fluid intelligence that was administered as part of the post-experimental questionnaire for a random subset of participants.[9] Among subjects whose test score lies below the median in our sample, the likelihood to follow defaults is 31.4 percentage points higher in Baseline-Ample than in Baseline-Scarce (see Figure 2.3). In contrast, the size of the treatment effect is only 11.7 percentage points for subjects with an above-median test score. The treatment difference for subjects with above-median test scores is significantly smaller than the difference for those with below median test scores (t-test, $p = 0.022$).[10] In fact, treatment differences in passivity rates even

9. Specifically, our measure is based on a 10-item version of Raven's Progressive Matrices, in which 50% of participants in both the Baseline-Ample and Baseline-Scarce condition participated.

10. The reported p-value for the "diff-in-diff" effect is obtained from a regression framework in which a treatment dummy is interacted with a dummy that is one if the subject's score in the Raven test lies above the sample median and zero otherwise. Standard errors account for potential clustering at the subject level.

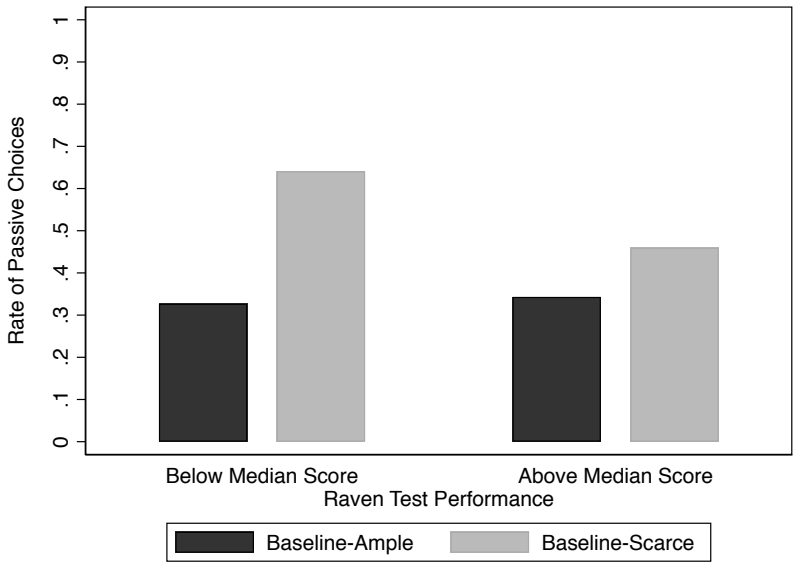


Figure 2.3. Passive Behavior by Raven Scores

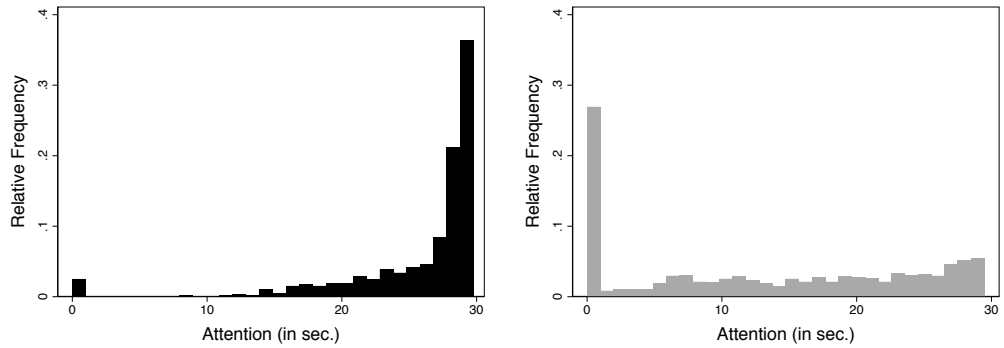
Notes: The figure depicts default adherence rates in Baseline-Scarce and Baseline-Ample, separately for subjects whose performance in a Raven matrices test lies above / below the median test score.

vanish entirely for subjects in the top quartile of the test-score distribution (37.3% default adherence in Baseline-Ample vs. 33.6% Baseline-Scarce, Mann-Whitney-U test, $p = 0.8661$).

2.4.2 Re-allocation of Cognitive Resources

The differences in passive behavior between Baseline-Ample and Baseline-Scarce are consistent with the hypothesized consequences of cognitive resource scarcity. In a next step, we analyze whether the underlying mechanisms are also in line with those discussed in Section 2.3. In particular, we study whether the observed behavioral effects can be linked to treatment differences in how subjects' allocate their cognitive resources across tasks.

As our first measure of cognitive resource allocation, we examine how much visual attention subjects devote to the decision task. In particular, we consider the total number of seconds that a subject dedicates to the decision task in a given round of the experiment. Figure 2.4 depicts histograms of the attention spans in Baseline-Ample and Baseline-Scarce. The figure demonstrates strong, systematic, and statistically significant treatment differences in how subjects allocate their attention (Kolmogorov-Smirnov test on subject-round level, $p < 0.01$). The difference between treatments is particularly striking when considering the modes of the distributions: while the modal behavior in Baseline-Ample is paying maximal attention, subjects most frequently devote zero attention to the decision task when



(a) Baseline-Ample **(b)** Baseline-Scarce

Figure 2.4. Attention Spans in Baseline

Notes: The figure plots distributions of attention spans devoted to the decision task, as measured by the total number of seconds devoted to the task in a given period. Histograms for Baseline-Ample (left panel) and Baseline-Scarce (right panel).

cognitive resources are scarce. On average, the amount of time dedicated to the decision task decreases from 25.76 seconds in Baseline-Ample to 13.15 seconds in Baseline-Scarce (t-test, $p < 0.01$).

The documented effect on average attention spans can be decomposed into an extensive-margin and an intensive-margin effect. Subjects completely ignore the decision task in 32.1% of cases in Baseline-Scarce, whereas they do so in only 2.5% of cases in Baseline-Ample ($p < 0.01$, Fisher-exact test).[11] Two points are worth noting about this result. First, paying zero attention to the decision task directly implies that subjects devote no cognitive resources to solving the task. Hence, in line with the first part of Hypothesis 1, subjects are more likely not to devote any resources to the decision task when cognitive resources are scarce. Second, complete inattention to the decision task automatically translates into passive acceptance of the stipulated default option. The extensive-margin reduction in cognitive resources thus accounts for a considerable share of the overall treatment difference in the frequency of passive choices. In particular, both the treatment difference in passive choices as well as the treatment difference in completely inattentive choices are approximately 30 percentage points. At first glance, this observation might suggest that the extensive-margin reduction in attention accounts for the entire increase in passivity across treatments. Note, however, that even if subjects attend to the task, they follow the default in roughly one third of the cases. Hence, about one third of

11. The reported numbers are based on the full sample of the Baseline environment and therefore differ slightly from the values in Figure 2.4 (recall that the exact length of attention spans depicted in Figure 2.4 was recorded only for a 50% subsample in both treatments). The corresponding figures in the sample underlying Figure 2.4 are 26.3% (Baseline-Scarce) and 2.5% (Baseline-Ample), respectively ($p < 0.01$, Fisher-exact test).

the decisions that are taken under complete inattention in Baseline-Scarce would also have resulted in a "default choice" in Baseline-Ample. The extensive-margin effect therefore accounts for approximately two thirds of the observed increase in passivity. An Oaxaca-Blinder decomposition (Blinder, 1973; Oaxaca, 1973; Fortin, Lemieux, and Firpo, 2011), presented in more detail in 2.C.3, confirms this observation: the extensive-margin reduction is estimated to account for approximately 70% of the overall treatment difference in passive behavior.

The second part of Hypothesis 1 predicts that the amount of cognitive resources devoted to the decision task should also decrease at the intensive margin. To examine whether this is the case, we first analyze differences in the amount of attention that subjects devote to the decision task conditional on paying any attention to the task (i.e., conditional on entering the decision task at least once in a given round of the experiment). In line with our hypothesis, we observe that the conditional average attention span in Baseline-Scarce is significantly shorter than in Baseline-Ample (17.8 vs. 26.4 seconds; t-test, $p < 0.01$). This shift in attention spans is driven by the entire population of subjects. Figure 2.5 depicts the cumulative distribution of mean attention spans that subjects devote to the decision task. The cumulative distribution in Baseline-Ample first-order stochastically dominates its counterpart in Baseline-Scarce, corroborating the second part of Hypothesis 1.[12] Hence, scarcity of cognitive resources causes subjects to shift their attention away from the decision task at both the intensive and the extensive margin.

In sum, our findings on how individuals allocate their attention in Baseline-Ample and Baseline-Scarce are consistent with our behavioral hypotheses. While the attention data provide valuable insights into subjects' allocation of cognitive resources, they have to be treated with some caution. Specifically, it is a priori not clear that attention spans provide a precise measure for the allocation of cognitive resources at the intensive margin. For example, it could be the case that subjects in Baseline-Scarce attend to the decision task, but are not able to effectively deploy cognitive resources to solve the task, as the background task simply takes up too much of their bandwidth. Conversely, subjects could still devote some cognitive resources to the task at times where they do not visually attend to it. We therefore consider a second measure for treatment differences in cognitive resource allocation at the intensive margin: the quality of individuals' choices in the decision task. Since the decision task requires cognitive resources to be solved correctly, and since more resources devoted to the task will translate into better decisions, the decision quality constitutes a natural "outcome-based" measure of the resources allocated to the decision task. To examine intensive-margin treatment differences based on this measure, we compare the quality of subjects' choices for all cases in which subjects entered the

12. First-order stochastic dominance is also observed when considering the CDFs of decision-level data instead of subject-level averages (see Figure 2.C.1 in the appendix).

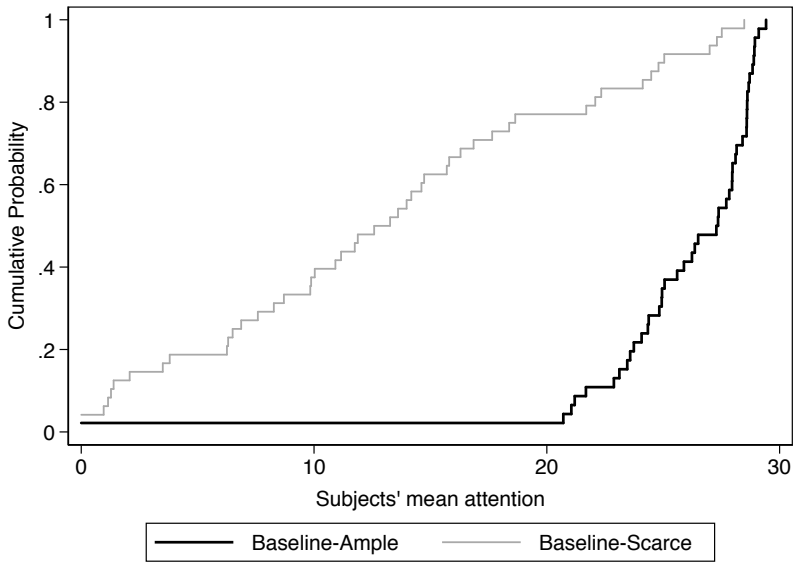


Figure 2.5. Individuals' Attention Levels in BASELINE

Notes: The figure depicts the cumulative distribution of subjects' mean attention levels in BASELINE-AMPLE and BASELINE-SCARCE

decision task at all. On average, subjects in BASELINE-AMPLE correctly solve the decision task in 85.0% of these cases. The corresponding number for BASELINE-SCARCE is 12.1 percentage points lower (t-test, $p < 0.01$). The data on decision quality thus ascertain that subjects do not only react at the extensive margin, but also reduce the cognitive resources allocated to the decision task at the intensive margin.

Result 2. *Under cognitive resource scarcity, subjects devote fewer resources to the decision task, as measured by the amount of attention devoted to the task as well as the quality of individuals' choices.*

2.4.3 How Choice-promoting Interventions Affect Passivity

The observed differences between BASELINE-AMPLE and BASELINE-SCARCE demonstrate that scarce cognitive resources can be an important source of passive behavior. This insight raises two interesting questions regarding the effects of policies that aim at encouraging active decision-making. First, do the consequences of such choice-promoting policies depend on whether decision makers act under cognitive resource scarcity or with abundant cognitive resources? Second, how does fostering active choice in one domain affect individuals' decisions in other tasks or choice domains?

To shed light on these questions, we analyze the behavior of subjects in the two additional decision environments (cp. Table 2.1). Figure 2.6 compares default adherence rates between the BASELINE and DIRECTED ATTENTION environment. In line

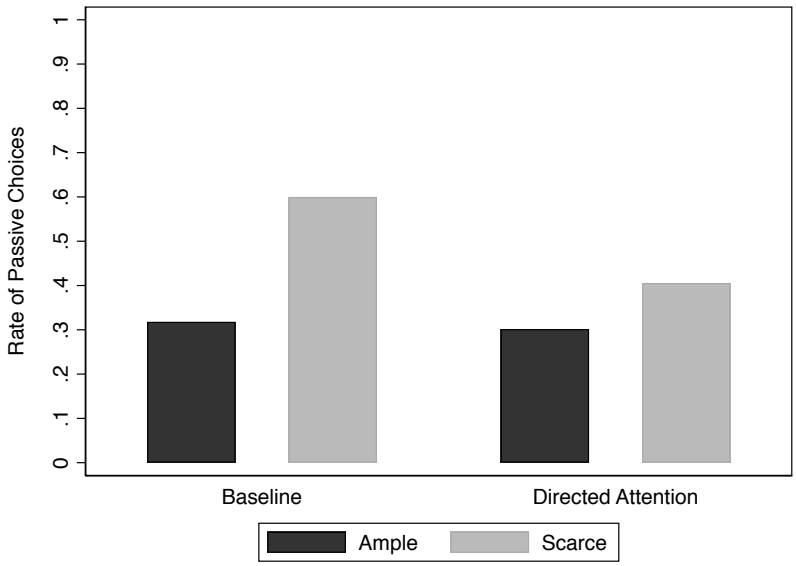


Figure 2.6. Passive Behavior Across Choice Environments

Notes: The figure shows the average default adherence rates for Baseline and Directed Attention.

with Hypothesis 3, the frequency of passive behavior is very similar in Baseline-Ample and Directed-Ample in which the background task puts little strain on individuals' cognitive resources (31.8% vs. 30.2% in Baseline-Ample and Directed-Ample, respectively; Mann-Whitney-U test, $p = 0.243$). This finding corroborates our previous observation that subjects in Baseline-Ample essentially devote all of their cognitive resources to the decision task (cp. Figure 2.4 and Figure 2.5). An intervention that aims at steering individuals' attention to the decision task thus has only negligible effects on the allocation of cognitive resources, and the resulting choices.

Under cognitive resources scarcity, this picture changes substantially. Comparing the frequency of passive choices between Baseline-Scarce and Directed-Scarce reveals strong treatment differences. Default adherence rates drop from 60.0% to 40.6% when individuals' attention is directed to the decision task, by simply displaying the task permanently on their screens. The difference in the frequency of passive choices is statistically significant (Mann-Whitney-U test, $p < 0.01$). At the same time, however, the rate of passive choices in Directed-Scarce still lies significantly above the one in Directed-Ample (Mann-Whitney-U test, $p < 0.01$). Directing attention to the decision task thus fosters active decision making, but it does not fully eliminate the passivity caused by cognitive resource scarcity.

It is also informative to compare default adherence in Baseline-Scarce and Directed-Scarce to the frequency with which subjects choose the corresponding decision alternative in Active-Scarce (that featured no defaults), i.e., to the

rate at which subjects in Active-Scarce choose the option that happened to be the default in the *exact same version* of the decision task in Baseline-Scarce and Directed-Scarce. In Active-Scarce, subjects choose this option in 33.8% of cases. This number lies significantly below the rate of passive choices in both Baseline-Scarce and Directed-Scarce (Mann-Whitney-U test, $p < 0.01$ for Active-Scarce vs. Baseline-Scarce; $p = 0.018$ for Active-Scarce vs. Directed-Scarce), corroborating the second part of Hypothesis 3.[13]

Result 3. *Encouraging active decision-making through an active-choice intervention or by directing individuals' attention to a task reduces passivity if cognitive resources are scarce. Directing subjects' attention to a specific decision, however, does not fully eliminate passive behavior relative to an active-choice environment. Both interventions do not affect passivity if subjects have ample cognitive resources.*

2.4.4 Consequences for Choice Quality

The increase in active decision making in response to choice-promoting interventions also leads subjects in Directed-Scarce and Active-Scarce to make *better* decisions. As the first row of Table 2.2 shows, subjects in Baseline-Scarce on average solve 59.1% of decision tasks correctly. This number increases to 68.8% and 72.0% in Active-Scarce and Directed-Scarce, respectively. The observed increase in decision quality is statistically significant for both Active-Scarce and Directed-Scarce (Mann-Whitney-U test, $p < 0.01$ in both cases), whereas the latter two treatments do not differ significantly from each other ($p = 0.338$). Hence, when cognitive resources are scarce, both choice-promoting interventions succeed in their primary goal: they help individuals to make better decisions by encouraging them to choose actively.

In the three conditions with Ample cognitive resources, we observe only minor differences in the quality of choices in the decision task (see row (4) in the bottom panel of Table 2.2).[14] This is not surprising, as we would expect subjects in Baseline-Ample to devote essentially all of their cognitive resources to the decision task. Hence, policies that direct attention to this task or force subjects to make an active decision should not yield improvements in choices.

13. A question of detail concerns the evaluation of cases in which subjects failed to make a decision in the Active Choice environment (cp. Footnote 4). For sake of comparability, the numbers reported above treat these cases as "passive choices"—since the same behavior of ignoring the decision task entirely would result in a default choice in the Baseline and Directed Attention environments. If we instead drop the corresponding cases from our calculations, the passive-choice frequencies slightly change to 27.5% (Active-Scarce) and 28.5% (Active-Ample), respectively.

14. The difference in choice quality between Baseline-Ample and Directed-Ample turns out to be statistically significant (Mann-Whitney-U test, $p = 0.035$), but is relatively small in magnitude. The corresponding differences for Baseline-Ample vs. Active-Ample and Directed-Ample vs. Active-Ample are both statistically insignificant ($p = 0.265$ and $p = 0.300$, respectively).

Table 2.2. Decision Quality and Payoffs

			Baseline	Directed Attention	Active Choice
			Scarce		
(1)	Decision Task	% correct	59.11 (21.47)	71.98 (18.45)	68.76 (21.92)
(2)	Background Task	% correct	77.89 (20.30)	73.33 (20.82)	74.44 (18.47)
(3)	Total Payoff	Earnings per round	37.07 (8.15)	36.53 (8.65)	36.65 (7.58)
			Ample		
(4)	Decision Task	% correct	83.56 (14.00)	87.45 (11.98)	85.47 (14.05)
(5)	Background Task	% correct	97.98 (3.46)	98.78 (2.40)	96.93 (10.87)
(6)	Total Payoff	Earnings per round	47.55 (2.16)	48.26 (1.68)	47.32 (5.34)

Notes: The table presents proportions of correctly solved tasks in the decision task and background task, as well as the average total payoff of subjects in one round of the experiment. The reported standard deviations (in parentheses) are calculated based on subject-level averages in decision qualities and payoffs, respectively (i.e., they refer to the between-subject SDs).

While reductions in passivity and the potential gains resulting from more deliberate, active decisions are typically the core criteria to evaluate choice-promoting interventions, our experiment is designed to also shed light on potential cognitive spillovers to other domains or tasks that decision makers have to handle simultaneously. Specifically, we can investigate how encouraging active decision making in the decision task affects individuals' performance in the background task. The second row of Table 2.2 indicates that the interventions have a negative effect on the quality of subjects' decisions in the background task. The likelihood to correctly recall the number decreases from 77.9% in Baseline-Scarce to 73.3% in Directed-Scarce and 74.4% in Active-Scarce. In both cases, the differences relative to Baseline-Scarce are (weakly) significant (Mann-Whitney-U tests, $p = 0.071$ for Baseline-Scarce vs. Directed-Scarce, $p = 0.086$ for Baseline-Scarce vs. Active-Scarce).

The results show that the studied choice-promoting interventions cause two countervailing effects. On the one hand, they reduce passivity and thereby improve decisions in the targeted choice domain. On the other hand, they encourage subjects to withdraw scarce cognitive resources from other choice domains, which in turn deteriorates the quality of their decisions in these domains. To evaluate the overall consequences of the Directed Attention and Active Choice intervention, it is therefore crucial to assess whether the increase in payoffs in the targeted domain is "worth" the accompanying negative cognitive spillovers on other decisions. We address this question by comparing treatment differences in subjects' average total payoff from solving the decision task and the background task.[15] Notably, individuals' average total payoff (reported in the third row of Table 2.2) is essentially identical for all three decision environments. If anything, payoffs are slightly higher in Baseline-Scarce compared to Directed-Scarce and Active-Scarce. All pairwise treatment comparisons turn out to be statistically insignificant (Mann-Whitney-U tests, $p = 0.659$ for Baseline-Scarce vs. Directed-Scarce, $p = 0.415$ for Baseline-Scarce vs. Active-Scarce, and $p = 0.791$ for Directed-Scarce vs. Active-Scarce). Hence, the positive impact on the quality of choices in the decision task and the negative spillovers on the background task cancel each other out, such that the resulting net effect on subjects' overall payoffs is essentially zero.

Result 4. *Compared to Baseline-Scarce, the quality of choices in the decision task increases in Directed-Scarce and Active-Scarce. At the same time, choice quality in the background task decreases from Baseline-Scarce to Directed-Scarce and Active-Scarce. The resulting overall payoffs do not differ significantly across decision environments.*

15. Recall that the payoff for solving the background task correctly is relatively high (see Section 3.2). Hence, while the percentage change in the decision quality of this task may appear to be small, the observed differences can be relatively high for subjects' overall payoffs.

2.5 Conclusion

We conclude by discussing practical implications of our findings for institutions that design or evaluate default rules and choice-promoting policies. Typically, the success of such policies is examined solely with respect to the outcomes in the decision domain that is the subject of the intervention. This approach presumes that the policies do not trigger negative spillovers to other domains. Our results indicate that this assumption might frequently be violated. Whenever different tasks or decisions compete for people's scarce cognitive resources, interventions that foster active choice in one domain can induce negative cognitive spillovers to others, which may dilute or even fully offset the policies' positive effects. While for some interventions the net effects may still be positive—e.g., in the case of high-quality personalized recommendations (Kling et al. (2012), Kaufmann et al. (2018))—examining the existence and magnitude of cognitive spillovers is crucial in order not to systematically overestimate the benefits of choice-promoting policies.

Our findings also shed new light on the question of which "types" of decision makers are especially prone to stick to defaults, and under which conditions default specifications might be beneficial for consumers. The fact that passive decision making is more pronounced under cognitive resources scarcity implies that default specifications have particularly strong consequences for subgroups of the population that face scarce cognitive resources. To the extent that poverty is a driver of such resource scarcity (see for example Mani et al., 2013), the group that remains passive will be relatively more likely to be poor. When stipulating defaults, policy makers should account for these differences in the incidence of default effects, and their potential distributional consequences. In the context of defaults that are set by firms to sell preconfigured goods, the same reasoning implies that decision makers with scarce cognitive resources will be particularly susceptible to exploitation. On a more positive note, our findings also suggest that high-quality defaults can generate positive cognitive spillovers to choices in other domains of decision makers' lives. In particular, defaults in our experiment were chosen at random and, therefore, passive behavior resulted in relatively poor choices in the decision task. Well-chosen defaults, in contrast, might yield a double dividend when cognitive resources are scarce: they do not only improve outcomes for passive decision makers, but also "free up" scarce cognitive resources, allowing people to focus on other particularly pressing tasks or decisions.

Appendix 2.A Formal Derivation of Hypotheses

This section presents the formal arguments underlying the behavioral predictions discussed in Section 2.3. To fix ideas, we present a simple theoretical framework reflecting the key ideas of our experiment design. Within this framework, we describe a set of sufficient conditions for the hypotheses discussed in Section 2.3. We base our analysis upon the premise that agents have a limited stock of cognitive resources and optimally allocate these resources across tasks. While both of these assumptions represent significant simplifications, this approach is valuable for deriving qualitative predictions on how we should expect behavior to differ across treatments. We build upon a simplified version of the framework proposed by Alonso, Brocas, and Carillo (2014) to analyze resource allocation in the brain.[16] The main idea of Alonso, Brocas, and Carillo's model is that different cognitive tasks are executed by different systems of neurons. These systems simultaneously demand resources, which are allocated by a central executive system. For more details on the underlying research in neuroscience, we refer the reader to the literature review in Alonso, Brocas, and Carillo (2014) and Brocas (2012).

To understand how the allocation of cognitive resources affects outcomes across treatments, we incorporate two key features of our experiment. First, we assume that subjects face two tasks that simultaneously require resources—the decision task and the background task. Second, the decision task features three options, one of which is randomly preselected as the default. Formally, suppose that every individual j is equipped with a stock of cognitive resources X^j and faces a background task B and a decision task D to which she can allocate resources $x_i,\ i \in \{B, D\}$, such that $x_B + x_D \leq X^j$. Allocating resources x_B to a background task of difficulty $\theta \in \{\theta_L, \theta_H\}$ results in a likelihood of $\pi_B(x_B, \theta)$ to correctly solve the background task and obtain utility u_B. $\pi_B(x_B, \theta_H)$ is increasing, strictly concave, and continuously differentiable in x_B. Moreover, we assume that $\pi_B(x_B, \theta_L) = 1 \ \ \forall x_B \in \mathbb{R}^+$, in line with the idea that keeping in mind a two-digit number essentially requires no cognitive resources. Allocating resources x_D to the decision task results in a likelihood $\pi_D(x_D, d)$ to solve the task and obtain utility u_D, where $d \in \{c, inc, no\}$ specifies whether the stipulated default option is correct, incorrect, or nonexistent (as in the Active Choice environment). If the default is specified at random, a subject's expected probability to solve the task correctly is thus $\frac{1}{3}\pi_D(x_D, c) + \frac{2}{3}\pi_D(x_D, inc)$, which we assume to be strictly increasing, differentiable, and concave in x_D. If $x_D = 0$, the individual stays passive and automatically follows the default option (if there is one). Making use of the above notation we can now state the subjects' decision problem. Since it is ex ante unknown to subject j whether the default is correct, she allocates her cognitive

16. We deviate from Alonso, Brocas, and Carillo (2014) in two respects. First, we abstract from asymmetric information across different regions of the brain and revert to the case of perfect knowledge. Second, we impose slightly more structure on payoffs, in line with our experimental setup.

resources in the BASELINE environment according to:

$$\max_{x_B, x_D} u(x_B, x_D) \;=\; \pi_B(x_B, \theta) u_B + \left(\frac{1}{3}\pi_D(x_D, c) + \frac{2}{3}\pi_D(x_D, inc)\right) u_D \qquad (2.A.1)$$
$$s.t.\;\; x_B + x_D \;\leq\; X^j$$

The optimal allocation of cognitive resources thus depends on the shape of the functions $\pi_B(\cdot)$ and $\pi_D(\cdot)$ as well as the payoffs obtained from each of the tasks. In a next step, we discuss a set of plausible sufficient conditions on the shape of $\pi_D(x_D, d)$, under which the hypotheses from Section 2.3 hold.

Condition 1. For all x_D, the probabilities $\pi_D(x_D, d)$ satisfy the following conditions:
(i) $\pi_D(x_D, c) > \pi_D(x_D, no) > \pi_D(x_D, inc)$
(ii) $\frac{\partial \pi_D(x_D, c)}{\partial x_D} < \frac{\partial \pi_D(x_D, inc)}{\partial x_D}$
(iii) $\pi_D(x_D, no) = \frac{1}{3}\pi_D(x_D, c) + \frac{2}{3}\pi_D(x_D, inc)$

Condition 1 imposes three restrictions. First, correct default options catalyze correct choices and incorrect default options catalyze incorrect choices. Hence, holding x_D constant, the probability to make a correct choice is highest if the default option is correct and lowest if it is incorrect. This feature captures the intuition and widespread observation that people tend to stick to defaults disproportionally often, even if these are specified at random (see, e.g., Haan and Linde (2017), Altmann, Falk, Heidhues, Jayaraman, and Teirlinck (2019)).[17] Second, there is more to be gained from allocating cognitive resources to the decision task if the default option is incorrect, i.e., if the baseline probability to make a mistake is relatively high. Third, for a given amount of cognitive resources, subjects are equally likely to solve the decision task correctly irrespectively of whether there is a random default option or no default option. This condition implies that there is no arbitrage opportunity in terms of decision qualities if a random default instead of no default is established.

Derivation of Hypothesis 1

We start by deriving the optimal allocation of cognitive resources. If a subject faces a background task with difficulty θ_L, all cognitive resources will be allocated to the decision task since $\pi_B(x_B, \theta_L) = 1 \quad \forall x_B \in \mathbb{R}^+$. Consider now a subject who faces a background task with difficulty θ_H. Since $\pi_B(x_B, \theta_H)$ is strictly increasing, all cognitive resources will be used in any optimum. The maximization problem can then be rewritten as

$$\max_{x_D} \pi_B(X_j - x_D, \theta_H) u_B + \pi_D(x_D) u_D,$$

17. We abstract from explicitly modeling the potential sources of this attraction, such as a status quo bias, omission/commission biases, etc. (see Sunstein (2013) for a comprehensive overview).

where $\pi_D(x_D) \equiv \frac{1}{3}\pi_D(x_D, c) + \frac{2}{3}\pi_D(x_D, inc)$. This objective function is strictly concave in x_D. The derivative with respect to x_D yields

$$-\pi'_B(X_j - x_D, \theta_H)u_B + \pi'_D(x_D)u_D.$$

As a consequence, the optimal solution (x_D^*, x_B^*) will satisfy $x_D^* = 0$ if and only if:

$$\pi'_D(0)\frac{u_D}{u_B} \leq \pi'_B(X^j, \theta_H).$$

To solve the optimization problem we have to consider two cases.

Case I: $\pi'_D(0)\frac{u_D}{u_B} < \pi'_B(0, \theta_H)$

Due to the concavity of π_B and π_D, there exists a threshold $\bar{X} \in \mathbb{R}^+$ such that subjects with $X^j \leq \bar{X}$ will abstain from devoting cognitive resources to the decision task in Baseline-Scarce. In Case I, $\bar{X}$ will be strictly positive, because subjects with a minimal stock of cognitive resources have a higher marginal incentive to allocate cognitive resources to the background task compared to the decision task. We should therefore observe an effect at the extensive margin of attention: More subjects in Baseline-Scarce completely ignore the decision task than in Baseline-Ample, where subjects should spend all their resources on the decision task.

For all subjects with $X^j > \bar{X}$, $x_D^* > 0$ holds. Furthermore, x_B^* will also be strictly positive for these subjects by construction of $\bar{X}$. Hence, we expect subjects with $X^j > \bar{X}$ to react at the intensive margin: They dedicate less resources to the decision task in Baseline-Scarce than in Baseline-Ample but they attend to the decision task in both conditions. If the population of participants is heterogeneous enough in terms of X^j, we should therefore observe extensive-margin as well as intensive-margin effects. As a consequence, the cumulative distribution over cognitive resources dedicated to the decision task in Baseline-Scarce first order stochastically dominates the corresponding cumulative distribution in Baseline-Ample, yielding Hypothesis 1.

Note that we chose parameters in the experiment to ensure treatment take-up, i.e., we choose incentives to ensure that subject try to solve the background task in all treatments (see Section 3.2 for details). We therefore derive all our hypotheses for the experiment from Case I only. For sake of completeness, we also solve the optimization problem if the condition does not hold.

Case II: $\pi'_D(0)\frac{u_D}{u_B} \geq \pi'_B(0, \theta_H)$

In this case, subjects always dedicate positive amounts of cognitive resources to the decision task. Then, there exists a threshold $\tilde{X}$ such that subjects dedicate all resources to the decision task if $X^j \leq \tilde{X}$ and distribute their resources across both tasks otherwise. Hence, subjects with $X^j \leq \tilde{X}$ do not react to the treatment and subjects with $X^j > \tilde{X}$ react at the intensive margin. Again, if the population of participants is heterogeneous enough in terms of X^j, we should observe the cumulative distribution over cognitive resources dedicated to the decision task in Baseline-Scarce to first-order stochastically dominate the corresponding cumulative distribution in Baseline-Ample.

Derivation of Hypothesis 2

Recall that subjects who completely ignore the decision task in Baseline-Scarce automatically stay passive. Passivity rates for subjects with a small stock of cognitive resources $X^j \leq \bar{X}$ are therefore higher in Baseline-Scarce compared to Baseline-Ample. Second, consider subjects with larger stocks of cognitive resources, who react at the intensive margin. Let ρ denote the probability of default adherence for $x_D > 0$, conditional on the default being incorrect and the subject choosing either the default or the other incorrect option. As indicated in Section 2.3, our hypotheses hold as long as individuals who dedicate fewer resources to a task are more likely to follow an incorrect default option than to actively opt for another wrong option, i.e., as long as $\rho \geq \frac{1}{2}$. In this case, a subject with $x_D > 0$ will stick to the default with probability

$$\frac{1}{3}\pi_D(x_D, c) + \frac{2}{3}\rho(1 - \pi_D(x_D, inc)).$$

To determine the direction of the effect of a reduction in cognitive resources at the intensive margin on default adherence, it suffices to consider the derivative of the above expression with respect to x_D, which is given by:

$$\frac{1}{3}\pi'_D(x_D, c) - \frac{2}{3}\rho\pi'_D(x_D, inc) \;\leq\; \frac{1}{3}[\pi'_D(x_D, c) - \pi'_D(x_D, inc)] \;\leq\; 0,$$

where the last inequality holds because there is more to be gained from allocating cognitive resources to the decision task if the default option is incorrect than if it is correct (part (ii) of Condition 1). Both the extensive and intensive margin effects therefore imply that default adherence is lower in Baseline-Ample than in Baseline-Scarce, yielding the first part of Hypothesis 2.

Focusing only on cases in which the default option is incorrect yields a likelihood to stick to the default of $\rho(1 - \pi_D(x_D, inc))$ for subjects who choose a strictly positive x_D^*. This probability decreases in x_D, because $\pi_D(x_D)$ increases with x_D and the second part of Condition 1. For this case, extensive- and intensive-margin effects are thus aligned and default adherence unambiguously increases from Baseline-Ample to Baseline-Scarce. Focusing instead on cases in which the default option is correct yields a default adherence probability of $\pi_D(x_D, c)$ for subjects with strictly positive x_D. Hence, subjects with $X^j \geq \bar{X}$ may follow the default more often in Baseline-Ample than in Baseline-Scarce, because they make fewer mistakes. However, there is a countervailing extensive-margin effect: more subjects choose $x_D^* = 0$ in Baseline-Scarce, which in turn leads to higher default adherence in Baseline-Scarce relative to Baseline-Ample. If the default option coincides with the correct solution, the net effect is thus ambiguous. These insights establish the second part of Hypothesis 2.

Derivation of Hypothesis 3

Our theoretical framework also provides a natural setting to examine how the choice-promoting interventions in the Directed Attention and Active Choice environments affect the allocation of cognitive resources, and the resulting decisions. We assume that, as a result of the interventions, some strictly positive amount of cognitive resources $0 < x_T < X^j \ \ \forall j$—which depends on treatment $T \in \{Directed, Active\}$—is exogenously directed towards the decision task. Given an amount of cognitive resources x_D, the expected rate of correct choices in the decision task is identical in Active Choice and Directed Attention because of the no-arbitrage condition (part (iii) of Condition 1). A subject thus faces the following decision problem in both treatments:

$$\begin{aligned} \max_{x_B, x_D} u(x_B, x_D) \ &= \ \pi_B(x_B, \theta) u_B + \left(\frac{1}{3} \pi_D(x_D, c) + \frac{2}{3} \pi_D(x_D, inc) \right) u_D \qquad \text{(2.A.2)} \\ s.t. \ \ x_D \ &\geq \ x_T \quad \text{and} \quad x_B + x_D \ \leq \ X^j \end{aligned}$$

In all environments with Ample resources, the constraint will not be binding since $x_T < X^j = x_D^* \ \ \forall j$. Hence, there will be no behavioral reaction to the treatment interventions and thus no differences in the quality of subjects' decisions, yielding the first part of Hypothesis 3. Directing individuals' attention to the decision task under cognitive resource scarcity (Directed-Scarce), however, increases the amount of resources allocated to the decision task for subjects with $X^j \leq \bar{X}$, relative to Baseline-Scarce. Following the same arguments as above, this increase implies lower default adherence. Nevertheless, individuals in Directed-Scarce will in expectation stick to the default in more than one third of cases. To see this, note that, for $\rho \geq \frac{1}{2}$,

$$\frac{1}{3} \pi_D(x_D, c) + \frac{2}{3} \rho (1 - \pi_D(x_D, inc)) \ \geq \ \frac{1}{3} [\pi_D(x_D, c) + (1 - \pi_D(x_D, inc))] \ \geq \ \frac{1}{3},$$

where the last inequality holds as a consequence of part (i) of Condition 1. Hence, options that are (randomly) preselected as default will, in expectation, be chosen more often in Directed-Scarce compared to the same (non-default) options in Active-Scarce.

Derivation of Hypothesis 4

In all environments with Ample resources, the constraint of devoting x_T to the decision task will not be binding since $x_T < X_j \ \ \forall j$. Hence, there will be no behavioral reaction to the treatment interventions, which yields the first part of Hypothesis 4. Since the optimization problem in (2.A.2) is a constrained version of the optimization in (2.A.1), it is clear that overall profits should be weakly higher in Baseline-Scarce compared to Directed-Scarce and Active-Scarce, which yields the second part of Hypothesis 4.

As stated in the derivation of Hypothesis 3, directing individuals' attention to the decision task under cognitive resource scarcity (Directed-Scarce), increases the amount of resources allocated to the decision task for subjects with $X^j \leq \bar{X}$, relative to Baseline-Scarce. As $\pi_D(x_D)$ is increasing in x_D, the increase in cognitive resources also implies an increase in the likelihood of making correct choices in the decision task. As the cognitive-resource constraint is binding in Directed-Scarce and Active-Scarce, the increase in resources devoted to the decision task implies a reduction of resources devoted to the background task and, hence, a lower likelihood of correctly solving this task (last part of Hypothesis 4).

Appendix 2.B Supplementary Information about the Experiment

2.B.1 Instructions

2.B.1.1 First Part

The first part of the experiment consists of a total of 10 rounds. The sequence of events in each round is as follows:

- First, a number is displayed on your screen.
- Your task is to memorize this number.
- After 10 seconds, the number will not be displayed anymore and you face a blank screen for 30 seconds.
- After 30 seconds have elapsed, a screen with an input field appears. In this field you can enter the number you have memorized at the beginning of the round. You have 20 seconds to enter the number.
- If you enter the correct number, you will receive a payoff of **40 cents**.
- If you enter nothing or a wrong number, you will receive **0 cents**.
- After you entered the number, the round is over and the next round begins.

2.B.1.2 Second Part

The second part of the experiment consists of a total of 10 rounds. The sequence of events in each round is as follows:

- In each round you will see three options, as illustrated in the figure below. Each option corresponds to a calculation task.
- Your task is to select the option that yields the **highest sum**.
- To select an option, please click the box in front of the respective option. You have 30 seconds to do so.
- If the selected option is the correct choice, you will receive a payoff of **10 cents** for this round.
- If the option you selected is not correct or you did not select any option, you will receive **0 cents** for this round.

2.B.1.3 Third Part

The third part of the experiment consists of a total of 20 rounds. In each round you can work on the task from Part 1 and the task from Part 2. The sequence of events in each round is as follows:

- First, a number is displayed on the screen for 10 seconds. As in Part 1, your task is to memorize this number.
- Following this screen, the number will not be displayed for 30 seconds. During this time, you can work on a calculation task as in Part 2. [Only in Baseline] *To work on the task from Part 2, you have to press and hold a key on the keyboard. The key you need to press will be displayed on the screen.*

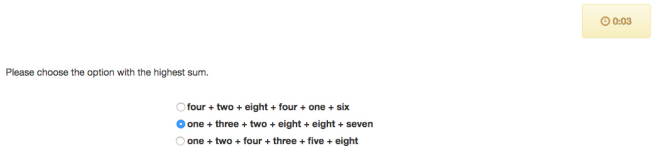
- [Only in BASELINE and DIRECTED] *In the task from Part 2, one of the three options is preselected in each round. If you do not work on the task, this option will be considered as your choice. The preselected option is determined randomly. This means that in about one third of the cases, the preselected option also corresponds to the correct option.*
- After 30 seconds have elapsed, a screen with an input field appears. As in Part 1, you can enter the number you have memorized at the beginning of the round in this field. You have 20 seconds to enter the number.
- If you enter the correct number in the task from Part 1, you will receive a payoff of **40 cents** for this task. Otherwise, you will receive **0 cents** for this task.
- If the chosen option in the task from Part 2 is correct, you will receive a payoff of **10 cents** for this task. If the chosen option is not correct, you will receive **0 cents** for this task.
- The payoff in one round is the sum of the payoffs from both tasks.

2.B.2 Screenshots

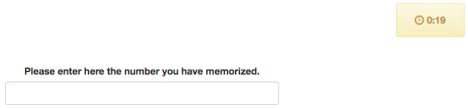
(a) Screen 1: Background Task

(b) Screen 2: Decision Task (only in BASELINE)

Figure 2.B.1. Screenshots



(c) Screen 3: Decision Task



(d) Screen 4: Background Task

Figure 2.B.1. Screenshots

Table 2.C.1. Descriptives

	Baseline		Directed Attention		Active Choice		Kruskal-Wallis
	Ample	Scarce	Ample	Scarce	Ample	Scarce	p-value
Decision Task ability	0.86	0.85	0.86	0.84	0.83	0.84	0.38
	(0.13)	(0.14)	(0.12)	(0.13)	(0.14)	(0.14)	
Background Task ability	0.83	0.81	0.79	0.81	0.80	0.81	0.64
	(0.12)	(0.14)	(0.14)	(0.15)	(0.16)	(0.14)	
Age	23.13	23.36	23.33	23.90	24.74	23.48	0.20
	(4.30)	(4.10)	(3.54)	(4.11)	(6.09)	(4.08)	
Female	0.59	0.62	0.62	0.57	0.55	0.53	0.82
	(0.50)	(0.49)	(0.49)	(0.50)	(0.50)	(0.50)	
% Economists	0.18	0.16	0.13	0.15	0.16	0.18	0.92
	(0.39)	(0.37)	(0.34)	(0.35)	(0.36)	(0.39)	
Raven Score[18]	6.09	6.35	6.20	6.35	6.06	5.83	0.55
	(1.87)	(1.67)	(1.63)	(1.47)	(1.97)	(1.46)	
N	94	95	94	96	96	89	

Notes: The table shows the average for basic characteristics of our sample in each treatment. Standard deviations are reported in parentheses.

Appendix 2.C Supplementary Analysis

2.C.1 Sample Descriptives

In the following table, we provide an overview of baseline characteristics for our sample. The first two rows in Table 2.C.1 depict treatment-level averages for the ability measures from Part 1 and Part 2 of the experiment (see Section 3.2 for details). The remaining rows depict average values of sociodemographic characteristics across treatments. The last column of Table 2.C.1 reports p-values of Kruskal-Wallis tests for the identity of means across treatments. We do not find any significant differences across treatments.

18. Number of correct answers in a 10-item version of Raven's Progressive Matrices. Raven scores were only elicited for 50% of participants in each treatment.

2.C.2 Attention and Re-allocation of Cognitive Resources

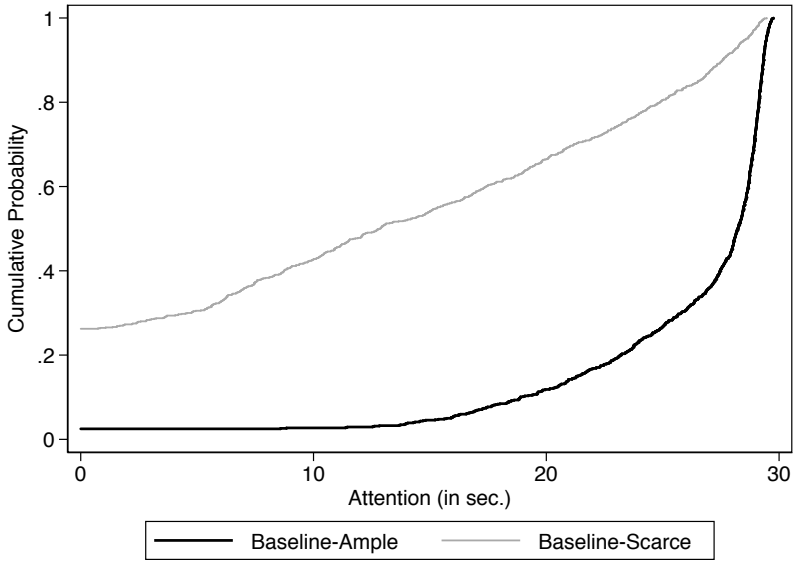


Figure 2.C.1. Attention Levels in BASELINE-AMPLE and BASELINE-SCARCE

Notes: The figure depicts CDFs of subjects' attention devoted to the decision task in BASELINE-AMPLE and BASELINE-SCARCE. Calculations are based on individual attention spans at the subject-round level.

2.C.3 Blinder-Oaxaca Decomposition

Using an Oaxaca-Blinder decomposition (Blinder, 1973; Oaxaca, 1973; Fortin, Lemieux, and Firpo, 2011), we can decompose the average treatment effect between Baseline-Ample and Baseline-Scarce. This approach provides a descriptive estimate of the extent to which the differences in passive choices between Baseline-Ample and Baseline-Scarce are accounted for by the differences in decisions that are taken without even entering the decision task.

Given the standard assumptions of the decomposition, the difference in the dependent variable Y, i.e., the treatment difference in the rate of passive choices, can be decomposed in three components. Denote by $I_{Attention=0}$ the dummy variable that takes the value of 1 if a subject does not enter the decision task in a given round of the experiment. Then, rewriting the overall difference in outcomes yields

$$\begin{aligned} E[Y_S] - E[Y_A] =& \beta_0^{\mathrm{S}} - \beta_0^{\mathrm{A}} + E[I_{Attention=0}|T = \mathrm{S}]\ (\beta_1^{\mathrm{S}} - \beta_1^{\mathrm{A}}) \\ &+ (E[I_{Attention=0}|T = \mathrm{S}] - E[I_{Attention=0}|T = \mathrm{A}])\beta_1^{\mathrm{A}}, \end{aligned}$$

where β_0^T and β_1^T are the coefficients from treatment specific linear regression models of Y on $I_{Attention=0}$, with $T \in \{S(carce); A(mple)\}$. The first two summands represent the effect due to changes in β_0 and β_1. The last part is the composition effect, i.e., the part of the treatment effect which is due to the change of the fraction of decisions that are taken by completely inattentive subjects. All components of the decomposition can be estimated using OLS regressions, replacing the expected values by the sample averages (see Table 2.C.2 for the corresponding numbers). The overall treatment difference in passivity is 28.19 percentage points. The decomposition reveals that 73.32% of this effect can be accounted for by the change in the fraction of decisions that are taken without entering the decision task. The effect at the extensive margin, therefore, accounts for a difference in passivity rates of 20.67 percentage points.

Table 2.C.2. Attention and Passive Choices

	Means		Coef.	
	(1) Ample	(2) Scarce	(3) Ample	(4) Scarce
$E[I_{Attention=0}\|T]$	0.0250	0.3205		
	(0.0036)	(0.0107)		
Attention = 0, β_1			0.6994	0.5887
			(0.0077)	(0.0221)
Constant, β_0			0.3006	0.4113
			(0.0077)	(0.0221)
Observations	1880	1900	1880	1900

Notes: The first two columns present the fraction of subjects in the Baseline-Ample and Baseline-Scarce treatment who devote no attention to the decision task. Columns (3) and (4) present OLS estimates using default adherence as the outcome variable. Robust standard errors (in parentheses) account for potential clustering on subject level.

References

Alonso, Ricardo, Isabelle Brocas, and Juan Carillo. 2014. "Resource Allocation in the Brain." *Review of Economic Studies* 81 (2): 501–34. [90]

Altmann, Steffen, Armin Falk, Paul Heidhues, Rajshri Jayaraman, and Marrit Teirlinck. 2019. "Defaults and Donations: Evidence from a Field Experiment." *Review of Economics and Statistics* forthcoming: [91]

Altmann, Steffen, and Christian Traxler. 2014. "Nudges at the Dentist." *European Economic Review* 72: 19–38. [65, 69, 73]

Altmann, Steffen, Christian Traxler, and Philipp Weinschenk. 2017. "Deadlines and Cognitive Limitations." *IZA Discussion Paper No. 11129*, [66]

Bartoš, Vojtěch, Michal Bauer, Julie Chytilová, and Filip Matějka. 2016. "Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition." *American Economic Review* 106 (6): 1437–75. [70]

Bhargava, Saurabh, George Loewenstein, and Justin Sydnor. 2017. "Choose to Lose: Health Plan Choices from a Menu with Dominated Option." *Quarterly Journal of Economics* forthcoming: [65, 69]

Bhargava, Saurabh, and Dayanand Manoli. 2015. "Psychological Frictions and the Incomplete Take-up of Social Benefits: Evidence from an IRS Field Experiment." *American Economic Review* 105 (11): 3489–529. [65]

Blinder, Alan. 1973. "Wage Discrimination: Reduced Form and Structural Estimates." *Journal of Human Resources* 8 (4): 436–55. [83, 102]

Bock, Olaf, Ingmar Baetge, and Andreas Nicklisch. 2014. "hroot: Hamburg Registration and Organization Online Tool." *European Economic Review* 71: 117–20. DOI: http://dx.doi.org/10.1016/j.euroecorev.2014.07.003. [74]

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2012. "Salience Theory of Choice Under Risk." *Quarterly Journal of Economics* 127 (3): 1243–85. [70]

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2013. "Salience and Consumer Choice." *Journal of Political Economy* 121 (5): 803–43. [70]

Brocas, Isabelle. 2012. "Information Processing and Decision-making: Evidence from the Brain Sciences and Implications for Economics." *Journal of Economic Behavior & Organization* 83 (3): 292–310. [90]

Brown, Jeffrey R., Anne M. Farrell, and Scott J. Weisbenner. 2011. "The Downside of Defaults." *Working Paper*, [68, 69]

Calzolari, Giacomo, and Mattia Nardotto. 2016. "Effective Reminders." *Management Science* 63 (9): 2915–32. [65, 67, 69, 73]

Caplin, Andrew, and Mark Dean. 2013. "Behavioral Implications of Rational Inattention with Shannon entropy." *NBER Working Paper No. 19318*, [70]

Caplin, Andrew, Mark Dean, and Daniel Martin. 2011. "Search and Satisficing." *American Economic Review* 101 (7): 2899–922. [66, 70, 71]

Caplin, Andrew, and Daniel Martin. 2016. "The Dual-Process Drift Diffusion Model: Evidence from Response Times." *Economic Inquiry* 54 (2): 1274–82. [68, 69]

Caplin, Andrew, and Daniel Martin. 2017. "Defaults and Attention: The Drop Out Effect." *Revue Économique* 68 (5): 747–55. [68]

Carlin, Bruce Ian, Simon Gervais, and Gustavo Manso. 2013. "Libertarian Paternalism, Information Production, and Financial Decision Making." *Review of Financial Studies* 26 (9): 2204–28. [69]

Carpenter, Jeffrey, Michael Graham, and Jesse Wolf. 2013. "Cognitive Ability and Strategic Sophistication." *Games and Economic Behavior* 80: 115–30. [71]

Carroll, Gabriel D., James J. Choi, David Laibson, Brigitte C. Madrian, and Andrew Metrick. 2009. "Optimal Defaults and Active Decisions." *Quarterly Journal of Economics* 124 (4): 1639–74. [66, 67, 73]

Carvalho, Leandro S, Stephan Meier, and Stephanie W Wang. 2016. "Poverty and Economic Decision-Making: Evidence from Changes in Financial Resources at Payday." *American Economic Review* 106 (2): 260–84. [68]

Chen, Daniel, Martin Schonger, and Chris Wickens. 2016. "oTree – An Open-source Platform for Laboratory, Online, and Field Experiments." *Journal of Behavioral and Experimental Finance* 9: 88–97. [74]

Chetty, Raj, John N. Friedman, Søren Leth-Petersen, Torben Heien Nielsen, and Tore Olsen. 2014. "Active vs. Passive Decisions and Crowd-Out in Retirement Savings Accounts: Evidence from Denmark." *Quarterly Journal of Economics* 129 (3): 1141–219. [65]

Damgaard, Mette Trier, and Christina Gravert. 2018. "The Hidden Costs of Nudging: Experimental Evidence from Reminders in Fundraising." *Journal of Public Economics* 157: 15–26. [66]

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. 2011. "Extraneous Factors in Judicial Decisions." *Proceedings of the National Academy of Sciences* 108 (17): 6889–92. [68]

Dean, Emma Boswell, Frank Schilbach, and Heather Schofield. 2017. "Poverty and Cognitive Function." In *The Economics of Asset Accumulation and Poverty Traps*. University of Chicago Press. [73]

Dean, Mark, and Nathaniel Neligh. 2017. "Experimental Tests of Rational Inattention." *Working Paper*, [70]

Deck, Cary, and Salar Jahedi. 2015. "The Effect of Cognitive Load on Economic Decision Making: A Survey and New Experiments." *European Economic Review* 78: 97–119. DOI: https://doi.org/10.1016/j.euroecorev.2015.05.004. [71]

DellaVigna, Stefano, and Ulrike Malmendier. 2006. "Paying not to Go to the Gym." *American Economic Review* 96 (3): 694–719. [65]

Dertwinkel-Kalt, Markus, Holger Gerhardt, Gerhard Riener, Frederik Schwerter, and Louis Strang. 2016. "Concentration Bias in Intertemporal Choice." *Working Paper*, [70]

Duarte, Fabian, and Justine S. Hastings. 2012. "Fettered Consumers and Sophisticated Firms: Evidence from Mexico's Privatized Social Security Market." Working paper. [69]

Fellner, Gerlinde, Rupert Sausgruber, and Christian Traxler. 2013. "Testing Enforcement Strategies in the Field: Threat, Moral Appeal and Social Information." *Journal of the European Economic Association* 11: 634–60. [66]

Finkelstein, Amy, and Matthew J. Notowidigdo. 2018. "Take-up and Targeting: Experimental Evidence from SNAP." *Working Paper*, [65]

Fortin, Nicole, Thomas Lemieux, and Sergio Firpo. 2011. "Decomposition Methods in Economics." In *Handbook of Labor Economics*. Vol. 4, Elsevier, 1–102. [83, 102]

Fréchette, Guillaume. 2012. "Session-effects in the Laboratory." *Experimental Economics* 15 (3): 485–98. [74]

Gabaix, Xavier. 2014. "A Sparsity-Based Model of Bounded Rationality." *Quarterly Journal of Economics* 129 (4): 1661–710. [70]

Gabaix, Xavier. 2017. "Behavioral Inattention." *NBER Working Paper No. 24096*, [66]

Gabaix, Xavier, David Laibson, Guillermo Moloche, and Stephen Weinberg. 2006. "Costly Information Acquisition: Experimental Analysis of a Boundedly Rational Model." *American Economic Review* 96 (4): 1043–68. [70]

Haan, Thomas de, and Jona Linde. 2017. "'Good Nudge Lullaby': Choice Architecture and Default Bias Reinforcement." *Economic Journal* forthcoming: [69, 91]

Handel, Benjamin R. 2013. "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts." *American Economic Review* 103 (7): 2643–82. [65]

Heffetz, Ori, Ted O'Donoghue, and Henry S. Schneider. 2016. "Forgetting and Heterogeneity in Task Delay: Evidence from New York City Parking-Ticket Recipients." *NBER Working Paper No. 23012*, [66]

Heiss, Florian, Daniel McFadden, Joachim Winter, Amelie Wuppermann, and Bo Zhou. 2016. "Inattention and Switching Costs as Sources of Inertia in Medicare Part D." *NBER Working Paper No. 22765*, [65, 73]

Huh, Young Eun, Joachim Vosgerau, and Carey K Morewedge. 2014. "Social Defaults: Observed Choices Become Choice Defaults." *Journal of Consumer Research* 41 (3): 746–60. [71]

Johnson, Eric J., John W. Payne, David A. Schkade, and James R. Bettman. 1989. "Monitoring Information Processing and Decisions: The Mouselab System." *Working Paper 89-4, Center for Decision Studies, Fuqua School of Business*, [70]

Karlan, Dean, Margaret McConnell, Sendhil Mullainathan, and Jonathan Zinman. 2016. "Getting to the Top of Mind: How Reminders Increase Saving." *Management Science* 62 (12): 3393–411. [65, 67, 69, 73]

Kaufmann, Cornel, Tobias Müller, Andreas Hefti, and Stefan Boes. 2018. "Does Personalized Information Improve Health Plan Choices when Individuals are Distracted?" *Journal of Economic Behavior and Organization* 149: 197–214. [66, 69, 71, 89]

Kling, Jeffrey R., Sendhil Mullainathan, Eldar Shafir, Lee C. Vermeulen, and Marian V. Wrobel. 2012. "Comparison Friction: Experimental Evidence from Medicare Drug Plans." *Quarterly Journal of Economics* 127 (1): 199–235. DOI: 10.1093/qje/qjr055. eprint: http://qje.oxfordjournals.org/content/127/1/199.full.pdf+html. [66, 67, 69, 89]

Kőszegi, Botond, and Adam Szeidl. 2012. "A Model of Focusing in Economic Choice." *Quarterly Journal of Economics* 128 (1): 53–104. [70]

Levav, Jonathan, Mark Heitmann, Andreas Herrmann, and Sheena S. Iyengar. 2010. "Order in Product Customization Decisions: Evidence from Field Experiments." *Journal of Political Economy* 118 (2): 274–99. [65, 68]

Mackowiak, Bartosz, Filip Matějka, and Mirko Wiederholt. 2018. "Rational Inattention: A Disciplined Behavioral Model." *Working Paper*, [70]

Mani, Anandi, Sendhil Mullainathan, Eldar Shafir, and Jiaying Zhao. 2013. "Poverty Impedes Cognitive Function." *Science* 341 (6149): 976–80. [68, 89]

Martin, Daniel. 2017. "Rational Inattention in Games: Experimental Evidence." *Working Paper*, [70]

Mullainathan, Sendhil, and Eldar Shafir. 2013. *Scarcity: Why Having Too Little Means So Much*. New York, NY: Time Books, Henry Holt & Company LLC. [66, 68]

Nielsen, Carsten S., Alexander C. Sebald, and Peter N. Sørensen. 2018. "Testing for Salience Effects in Choices under Risk." *Working Paper, University of Copenhagen*, [70]

Oaxaca, Ronald. 1973. "Male-female Wage Differentials in Urban Labor Markets." *International Economic Review* 14 (3): 693–709. [83, 102]

Rooij, Maarten van, and Federica Teppa. 2014. "Personal Traits and Individual Choices: Taking Action in Economic and Non-economic Decisions." *Journal of Economic Behavior & Organization* 100: 33–43. DOI: 10.1016/j.jebo.2013.12.019. [68]

Shah, Anuj K, Sendhil Mullainathan, and Eldar Shafir. 2018. "An Exercise in Self-replication: Replicating Shah, Mullainathan, and Shafir (2012)." *Journal of Economic Psychology*, [68]

Sharafi, Zahra. 2018. "The Impact of Financial Hardship on Adolescents' Cognitive Ability." *Working Paper*, [68]

Simons, Daniel J, and Christopher F Chabris. 1999. "Gorillas in our Midst: Sustained Inattentional Blindness for Dynamic Events." *perception* 28 (9): 1059–74. [69]

Sprenger, Amber, Michael Dougherty, Sharona Atkins, Ana Franco-Watkins, Rick Thomas, Nicholas Lange, and Brandon Abbs. 2011. "Implications of Cognitive Load for Hypothesis Generation and Probability Judgment." *Frontiers in Psychology* 2: 129. [71]

Stutzer, Alois, Lorenz Goette, and Michael Zehnder. 2011. "Active Decisions and Prosocial Behaviour: A Field Experiment on Blood Donation." *Economic Journal* 121: F476–F493. [66, 67, 73]

Sunstein, Cass R. 2014. "Choosing not to Choose." *Duke Law Journal* 64: 1. [73]

Sunstein, Cass R. 2013. "Deciding by Default." *University of Pennsylvania Law Review* 162 (1): 1–57. [91]

Tiefenbeck, Verena, Lorenz Goette, Kathrin Degen, Vojkan Tasic, Elgar Fleisch, Rafael Lalive, and Thorsten Staake. 2016. "Overcoming Salience Bias: How Real-Time Feedback Fosters Resource Conservation." *Management Science* forthcoming: DOI: 10.1287/mnsc.2016.2646. eprint: https://doi.org/10.1287/mnsc.2016.2646. [69]

Wang, Joseph Taoyi, Michael Spezio, and Colin F. Camerer. 2010. "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender-Receiver Games." *American Economic Review* 100 (3): 984–1007. DOI: 10.1257/aer.100.3.984. [70]

Chapter 3

Self-selection of Peers and Performance*

Joint with Lukas Kiessling and Sebastian Schaube

"The first thing I would do every morning was look at the box scores to see what Magic did. I didn't care about anything else."
– Larry Bird

3.1 Introduction

Basketball hall of famer Larry Bird motivated himself to train harder not by focusing on any player but rather by looking at his rival Magic Johnson's performance during the previous night's game. Similarly, seeing a specific classmate study long and continuously might also help to concentrate on one's own work. In various dimensions of life – ranging from students in educational settings (Sacerdote, 2001) over cashiers in supermarkets (Mas and Moretti, 2009) and fruit pickers on strawberry fields (Bandiera, Barankay, and Rasul, 2009; Bandiera, Barankay, and Rasul, 2010)

* We thank Viola Ackfeld, Philipp Albert, Thomas Dohmen, Lorenz Goette, Ingo Isphording, Sebastian Kube, Pia Pinger, Ulf Zölitz and audiences at Bonn, MBEPS 2017, VfS 2017, ESA Europe 2017, COPE 2018, ESA World 2018, IZA World Labor Conference, IZA Brown Bag, Rady Spring School in Behavioral Economics 2017, Bonn-Mannheim Ph.D. Workshop, 20th IZA Summer School in Labor Economics, 12th Nordic Conference on Behavioral and Experimental Economics, Max Planck Institute for Research on Collective Goods, EEA 2018, Bergen, and ESWM 2018 for helpful feedback and comments. We also thank the schools and students that participated in the experiments. This research was undertaken while all authors were at the University of Bonn. We did not obtain an IRB approval for this project because at the time of the experiment there did not exist an IRB at the University of Bonn's Department of Economics. However, we would like to stress that the schools' headmasters approved the study, written parental consent was required for students to take part in the study and participation was voluntary. Moreover, the experiment is in line with the requirements of the BonnEconLab. Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 (projects A01 and A02) is gratefully acknowledged.

to fighter pilots during World War II (Ager, Bursztyn, and Voth, 2016) – people affect each other through their presence, performance and choices. Yet, these social influences often stem from specific persons – roommates, frequently interacting coworkers, friends, or former colleagues – that individuals select themselves. This is in stark contrast with settings in which peers are randomly or exogenously assigned. But what actually changes once we allow peers to be self-selected? In general, these settings differ in two aspects: first, self-selection changes with whom one interacts; and, second, having the opportunity to self-select peers fundamentally changes the mode of peer assignment from exogenous (or random) assignment to self-selection. Both of these channels potentially alter an individual's motivation and behavior.

In this paper, we study how different peer assignment rules – self-selection versus random assignment – affect individual performance. In doing so, we examine a key feature of many peer effect studies, namely the absence of self-selection. In a first step, we document differences in performance between treatments which allow for self-selection or random assignment of peers. Subsequently, we analyze the underlying mechanisms. For this purpose, we decompose performance improvements into their two possible sources: an indirect effect stemming from changes in the peer composition and a direct effect from being able to self-select rather than being assigned to a specific peer.

In order to study the effects of self-selection, we conducted a framed field experiment (Harrison and List, 2004) with over 600 students (aged 12 to 16) in physical education classes of German secondary schools. Students took part in two running tasks (suicide runs) – first alone, then with a peer – and filled out a survey in between that elicited preferences for peers, personal characteristics, and the social network within each class. Our treatments exogenously varied the peer assignment in the second run using three different peer assignment rules. We implemented a random matching of pairs (Random) as well as two matching rules that used elicited preferences to implement two notions of self-selection: first, the classroom environment enabled students to state preferences for known peers (*name-based preferences*); and second, using a running task yielded direct measures of performance and thus could be used to select peers based on their relative performance in the first run (*performance-based preferences*). Using these two sets of preferences, we implemented two treatments with self-selection of peers by matching students based on either their name-based preferences (Name) or preferences over relative performance (Performance).

We find that self-selection of peers leads to an average performance improvement of 14–15 percent of a standard deviation relative to randomly assigned peers. While students in Random also improve their performance from the first to the second run, the improvements with self-selected peers almost double. Self-selection changes the peer composition, e.g., students predominantly interact with friends in Name, but tend to choose others with a similar past performance in Performance. Based on this finding, we decompose the overall treatment effect into an indirect

effect that is due to the peer's altered characteristics and a direct effect of being able to self-select a peer. Although we observe substantial peer effects in multiple dimensions (e.g., in relative performance in the first run), a peer's characteristics do not explain treatment differences resulting in an indirect effect close to zero. Instead, our estimates provide evidence that there is a direct effect of peer self-selection on performance. Therefore, the process of self-selection itself increases the performance of students. Borrowing from self-determination theory (Deci and Ryan, 1985; Deci and Ryan, 2000), we interpret this direct effect as a positive effect of having autonomy: being able to self-select peers has a psychological effect that enhances intrinsic motivation and improves subsequent performance. Finally, we simulate other exogenous peer assignment rules that seek to maximize or minimize the productivity differences between students. We document that these alternative rules yield performance improvements close to those observed with randomly assigned peers and therefore lower than those with peer self-selection. These findings thus support our interpretation that self-selection of peers carries a intrinsic value beyond changes in the peer composition.

Our results have three main contributions to the literature on peer effects, social interactions, and autonomy. First, we show that self-selection changes with whom people interact and thereby affects the overall composition of the reference or peer group. Second, we present evidence that self-selection of peers affects behavioral outcomes and has a direct effect on productivity. This highlights a novel channel through which peers and their selection affect behavior and provides the first clean evidence on autonomy in a field setting. Third, we document that peer effects may be present in multiple dimensions and discuss how this limits the effects of exogenous reassignment rules.

We document a strong causal difference in performance between widely-used randomly assigned peer groups and self-selected peers.[1] This focus on random peer assignment is understandable given that researchers aim to identify a clean causal effect of being exposed to peers. However, similar to what has been found in previous studies exploring the selection of students into peer groups (e.g., Tincani, 2017; Cicala, Fryer, and Spenkuch, 2018), our results indicate that the relevant *and* self-selected peer within a group does not equal to a random peer. This systematic selection helps to understand why the impact of certain peer groups differs compared to others: friends and non-friends may have differential effects (Lavy and Sand, forthcoming; Chan and Lam, 2015) and only persons with specific characteristics may

1. The literature on peer effects builds on (conditional) random assignment to identify peer effects and circumvent statistical issues outlined in Manski (1993). See also Sacerdote (2011) and Herbst and Mas (2015) for literature reviews on peer effects in education and a comparison of peer effects from field and lab settings, respectively.

affect performance (Aral and Nicolaides, 2017).[2] In light of our results, such differential peer effects can be due to self-selection of relevant peers. Related to our paper, Chen and Gong (2018) study self-selection of team members and document, consistent with our findings, that teams form endogenously along the social network outperform randomly assigned ones. We move beyond their work in at least three dimensions. First, we focus on a setup with a single peer and individual incentives. Thus, we restrict the possible sources of peer effects to that single peer. Second, we lever a rich dataset of individual characteristics and provide evidence that several attributes of randomly assigned peers matter. Third, by eliciting preferences for peers, we observe a normally unobserved dimension – the fit of a peer. Taken together, these features allow us to document that peer self-selection constitutes a novel behavioral channel through which peers can influence our behavior.

Moreover, our findings help to reconcile mixed evidence on the effectiveness of interventions changing class or work-group compositions to exploit peer effects (e.g., Duflo, Dupas, and Kremer, 2011; Carrell, Sacerdote, and West, 2013; Booij, Leuven, and Oosterbeek, 2017; Garlick, 2018). In our setup, the combination of two effects – the change in the peer composition and the multidimensionality of peer effects – has only a small impact on aggregate performance. More specifically, we move beyond peer effects in a single dimension and allow several characteristics such as productivity, friendship ties, and personality measures to exert peer effects.[3] Our results show that there are sizable peer effects apart from productivity. Consequently, if policy-makers reassign peers based on peer effects in a single dimension only, they neglect the fact that reassigning rules simultaneously change other peer characteristics giving rise to peer effects apart from the targeted dimension. These effects can counterbalance each other and lead to a net effect that is in our case close to zero and in general ambiguous. Hence, studies analyzing peer interactions and reassignment policies need to take into account not only a potential direct effect of self-selection, but also the multidimensionality of peer effects.

Our findings also contribute to the literature studying the effects of autonomy and decision rights on behavioral outcomes. In particular, we provide field evidence that self-selection (of peers) has a direct effect that can increase performance beyond its instrumental value of changing peer characteristics. Therefore, we complement laboratory studies by Bartling, Fehr, and Herz (2014) and Owens, Grossman, and Fackler (2014), who demonstrate that people are willing to pay for autonomy, i.e., the opportunity to actively select relevant aspects of their decision environment (Deci and Ryan, 1985). Similarly, autonomy in the workplace is associated

2. In a companion paper, Kiessling, Radbruch, and Schaube (2019), we study the peer selection process in more depth and relate the selection of peers to individual-level determinants.

3. Thereby we also join a small set of studies explicitly considering the impact of personality traits on educational outcomes or performance (e.g., Chan and Lam, 2015; Golsteyn, Non, and Zölitz, 2017). Yet, these other studies do not consider the implications of multidimensional peer effects.

with higher wages and employee happiness (Bartling, Fehr, and Schmidt, 2013) and leads to increased labor supply (Chevalier, Chen, Rossi, and Oehlsen, forthcoming), while removing autonomy has been found to have negative consequences on employee effort (Falk and Kosfeld, 2006).[4] Our results highlight an additional channel through which autonomy might provide value to employers or policy-makers: the freedom to choose one's own peers or teammates can boost performance similar to other non-monetary incentives such as recognitions and awards (Kosfeld and Neckermann, 2011; Bradler, Dur, Neckermann, and Non, 2016), framing of rewards (Levitt, List, Neckermann, and Sadoff, 2016) or personal goals (Koch and Nafziger, 2011; Corgnet, Gómez-Miñambres, and Hernán-González, 2015).

While the quantitative impact of different assignment mechanisms and the resulting peer composition might be specific to our setting and sample, students are a highly relevant subject group. They have not only been analyzed to study phenomena such as favoritism (Belot and Ven, 2011, and references therein), but peers during high school also have long-lasting effects on an individual's skill formation (Agostinelli, 2018) and hence on subsequent educational attainment. Moreover, the process of self-selecting peers is potentially equally important for settings in which peer effects do not arise due to social comparisons or peer pressure, but from effort or skill complementarities (e.g., Mas and Moretti, 2009; Bandiera, Barankay, and Rasul, 2010), or setting in which learning from peers is important (e.g., Jackson and Bruegmann, 2009; Bursztyn, Ederer, Ferman, and Yuchtman, 2014). The settings across these studies differ enormously, as does the underlying mechanism. Nonetheless, all of these share the notion that the behavior or action of peers imposes an externality on the action or behavior of others. In addition, peers can in principal also be self-selected affecting subsequent peer interactions.

The remainder of the paper is structured as follows. The next section presents our experimental design as well as procedural details. Section 3.3 presents the data and describes our sample of students. We outline our empirical framework in section 3.4. In section 3.5, we analyze how self-selected peers affect performance relative to randomly assigned peers and decompose this effect in a direct effect of self-selection and an indirect effect as a result of changes in the peer composition. We then interpret the direct effect and highlight potential policy implications. Finally, section 3.6 concludes.

4. These studies focus on individual decisions. However, autonomy can also help improve outcomes under collective decision-making. Having the right to vote has been can affect the quality of leadership positively (e.g., Brandts, Cooper, and Weber, 2014) as well as increase the effectiveness of institutions in the presence of social dilemmas (e.g., Bó, Foster, and Putterman, 2010; Sutter, Haigner, and Kocher, 2010).

3.2 Experimental Design

Studying the self-selection of peers and their subsequent impact on performance requires an environment in which subjects can choose peers themselves and where exogenous assignment can be implemented. Subjects must be able to compare their own performance with that of a peer in a task that lends itself to natural up- and downward comparisons. One complication in many settings is that it is difficult to isolate the person who serves as the relevant point of comparison. This is especially true if several potential peers are present at all times, among which only some constitute the set of an individual's relevant peers. As subjects might select those peers for many reasons besides their performance, it is essential not only to observe additional characteristics of all subjects, but also to collect data from an existing social group. In these groups, subjects have a clear impression of other group members and are able to select peers based on additional characteristics such as their social ties.

In this study, we used the controlled environment of a framed field experiment to overcome these challenges. We embedded our experiment in physical education classes of German secondary schools. Students from grades 7 to 10 participated in a running task, first alone and then simultaneously with a peer. Running allowed students to compare their performance with either faster or slower students, while it excluded complementaries in production between the students. Moreover, we focused on pairs as the unit of observation. This reduced the number of peers in the experimental task to a single individual and allows us to cleanly identify his or her impact. Subjects singled out specific peers by either naming them directly (in the treatment NAME) or selecting performance intervals (in PERFORMANCE). The respective treatments used these preferences to form pairs with self-selected peers or pairs were formed at random. Hence, we can compare the effect of self-selected peers with exogenously assigned ones, and can evaluate the effects of each assignment mechanism.

In the following, we present the design of our field experiment in detail and describe the implemented procedures.

3.2.1 Experimental Design

Figure 3.1 illustrates the experimental design. Students participated in a running task commonly known as "suicide runs", a series of short sprints to different lines of a

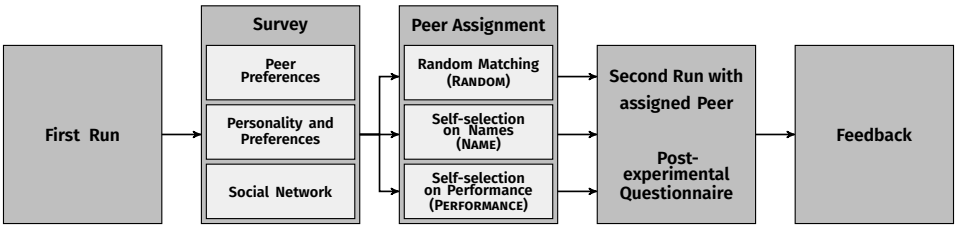


Figure 3.1. Experimental Design

volleyball court.[5,6] The first run – in which students ran alone – served two purposes: first, recorded times can be used as a measure of productivity and to evaluate the time improvement between the two runs; and second, we used (relative) times from the first run in combination with students' preferences to create pairs for the second run in one of the treatments described below. The second run mirrored the first one aside from the fact that students did not run alone, but rather in pairs. This means that two students performed the task simultaneously, while their times were recorded individually. Feedback about performance in both runs was only provided at the end of the experiment.

Between the two runs, students filled out a survey comprising three parts, eliciting preferences for peers, non-cognitive skills and information about the social network within each class. We elicited two kinds of preferences: first, we asked subjects to state the names of those classmates with whom they would like to perform the second run; and second, we asked them to state the relative performance level of their most-preferred peers. Note that we elicited all preferences irrespective of the assigned treatment and used these preferences to match students for the second run in two of the three treatments.

In addition to these preferences, the survey included socio-demographic questions and measures of personality and economic preferences: the Big Five inventory as used in the youth questionnaire of the German socio-economic panel (Weinhardt

5. The exact task is to sprint and turn at every line of the volleyball court. Subjects had to line up at the baseline. From there, they started running to the first attack line of the court (6 meters). After touching this line, they returned to the baseline again, touching the line on arrival. The next sprint took the students to the middle of the court (9 meters), the third to the second attack line (12 meters) and the last to the opposite baseline (18 meters), each time returning back to the baseline. They finished by returning to the starting point. The total distance of this task was 90 meters.

6. The task was chosen for several reasons: (1) the task is not a typical part of the German physical education curriculum, yet it is easily understandable for the students; (2) in contrast to a pure and very familiar sprint exercise as in Gneezy and Rustichini (2004) or Sutter and Glätzle-Rützler (2015), students should only have a vague idea of their classmates' performance and cannot precisely target specific individuals in PERFORMANCE; and (3) due to the different aspects of the task (general speed, quickness in turning as well as some level of endurance or perseverance), the performance across age groups was not expected to (and did not) change dramatically.

and Schupp, 2011), a measure of locus of control (Rotter, 1966), competitiveness[7], general risk attitude (Dohmen, Falk, Huffman, Sunde, Schupp, et al., 2011), and a short version of the INCOM scale for social comparison (Gibbons and Buunk, 1999; Schneider and Schupp, 2011). The survey concluded by eliciting the social network within every class. Subjects were asked to state up to six of their closest friends within the class.

Before and after the second run, we asked students a short set of questions about their peer and their experience during the task. Before the run, we elicited their belief about the relative performance of their peer in the first run, namely who they thought was faster. Following the second run, we asked them whether they would rather run alone or in pairs the next time, how much fun they had as well as how pressured they felt in the second run due to their peer on a five-point Likert scale.

3.2.2 Preference Elicitation

We used the strategy method to elicit two sets of peer preferences, independent of the treatment to which a subject is assigned. The first set elicited preferences for situations in which social information is available (*name-based preferences*). Accordingly, we asked each student to state his or her six most-preferred peers from the same gender within their class, i.e., those people with whom they would like to be paired in the second run. They could select any person of the same gender, irrespective of this person's actual participation in the study or their attendance in class.[8] These classmates had to be ranked, creating a partial ranking of their potential peers.

Second, we elicited preferences solely based on the relative performance in the first run, ignoring the identities of the potential running partners (*performance-based preferences*). For this purpose, we presented subjects with ten categories comprising one-second intervals starting from $(4, 5]$ seconds slower than their own performance in the first run, to $(0, 1]$ seconds slower and $(0, 1]$ seconds faster up to $(4, 5]$ seconds faster. Appendix Figure 3.I.1 presents a screenshot of the elicitation. We chose the range of intervals such that subjects could choose peers from a range of approximately ± 2 SD from their own performance in the first run. Subjects had to indicate from which time interval they would prefer a peer for the second run, irrespective of the potential peer's identity. Similar to the name-based preferences, we elicited a partial ranking for those performance-based preferences. Accordingly, subjects had

7. We implemented a continuous survey measure of competitiveness using a four-item scale. For this, we asked subjects about their agreement to the following four statements on a seven-point Likert scale: (i) "I am a person that likes to compete with others", (ii) "I am a person that gets motivated through competition", (iii) "I am a person who performs better when competing with somebody", and (iv) "I am a person that feels uncomfortable in competitive situations" and extracted a single principal component factor from those four items, of which the fourth item was scaled reversely.

8. All subjects were informed that peers in the second run would always have the same gender as themselves and would also need to participate in the study.

to indicate their most-preferred relative time interval, second most-preferred relative time interval and so on.[9]

3.2.3 Treatments

We exogenously varied how pairs in the second run are formed by implementing one of three matching rules at the class level, where pairs are only formed within genders. The first rule matched students randomly – i.e., we employed a random matching (RANDOM) – and serves as a natural baseline treatment.

The second matching rule used the elicited name-based preferences (NAME) and the third rule formed pairs based on the elicited performance-based preferences (PERFORMANCE). Note that the problem of matching pairs constitutes a typical roommate problem. We thus implemented a "stable roommate" algorithm proposed by Irving (1985) to form stable pairs using the elicited preferences.[10]

Subjects did not know the specific matching algorithm, but were only told that their preferences would be taken into account when forming pairs. Furthermore, we highlighted that the mechanism is incentive-compatible by telling students that it is in their best interest to reveal their true preferences. We informed subjects about the existence of all three matching rules in the survey to elicit both sets of preferences irrespective of the implemented treatment. Just before the second run took place, they were informed about the specific matching rule employed in their class and the resulting pairs.

In addition, we conducted an additional control treatment (NOPEER) in which students ran alone twice and which featured a shortened survey but was otherwise identical to the other treatments.[11] As the focus of this paper is the differential size of peer effects and not their existence per se, this only serves the purpose of excluding learning as a source of time improvements between the two runs. Hence, we exclude it from the main analysis and focus only on the evaluation of different peer assignment rules.

9. Naturally, each time interval could only be chosen once in the preference elicitation, although each interval could potentially include several peers if several subjects had similar times and thus belonged to the same interval. Similarly, some intervals may not contain any peers if no subject in the class had a corresponding time.

10. Given the mechanism proposed by Irving (1985), it is a (weakly) dominant strategy for all participants to reveal their true preferences. The matching algorithm requires a full ranking of all potential peers to implement a matching. Since we only elicited a partial ranking, we randomly filled the preferences for each student to generate a full ranking. However, in most cases subjects were assigned a peer according to one of their first three preferences. Nonetheless, if groups were small, it could be the case that subjects were not assigned one of their most-preferred peers. This is especially the case for performance-based preferences. See also the discussion in section 3.3.1 below.

11. The survey asked students for their preferences for peers, socio-demographics and their social network. Moreover, in order to avoid deception, we told students in advance that they would run alone both times.

3.2.4 Procedures

We conducted the experiment in physical education lessons at three secondary schools in Germany.[12] All students from grades 7 to 10 (corresponding to age 12 to 16) of those schools were invited to participate in the experiment. Approximately two weeks prior to the experiment, teachers distributed parental consent forms. These forms contained a brief, very general description of the experiment. Only those students who handed in the parental consent before the study took place participated in the study.

The experiment started with a short explanation of the following lesson and a demonstration of the experimental task. A translation of this explanation as well as screenshots detailing the preference elicitation are presented in Appendix 3.I.

We informed students that their teacher would receive each student's times from both runs, but no information about the pairings during the second run.[13] The students themselves did not receive any information on their performance until the completion of the experiment.

Additionally, we stressed that both of their performances would be graded by their teacher – thus incentivizing both runs – and that the objective was to run as fast as possible in both runs.[14] Moreover, most students themselves were very interested in their own times. The introduction concluded with a short warm-up period. After this, the subjects were led to a location outside of the gym.

Students entered the gym individually, which ruled out any potential audience effects from classmates being present by design. Students completed the first suicide run and subsequently were handed a laptop to answer the survey. Answering the survey took place in a separate room.[15] After the completion of the survey, subjects returned the laptop to the experimenter and waited with the other students outside the gym. Upon completion of the survey by all students, they returned to the gym to receive further instructions for the second run. In particular, we reminded the students of the existence of the three matching rules, and announced which randomly assigned rule was implemented in their class as well as the resulting pairs from the matching process. Following these instructions, the entire group waited outside the

12. Physical education lessons in most German secondary schools last for two regular lessons of 45 minutes each, thus about 90 minutes in total. At the third school, lessons only lasted 60 minutes for most classes. In order to conduct the experiment in the same manner as at the other schools, we were allowed to extend the lessons by 10 to 15 minutes, which was sufficient to complete the experiment.

13. Of course, some teachers were present in the gym. In principle, they could observe the pairings and therefore reconstruct the resulting pairs. However, none of the teachers made notes about the pairings or asked for them.

14. In order for the teacher to grade the entire set of students, the students who did not participate in the study also had to run twice. Their times were recorded for the teacher only and were never stored by us.

15. At least one experimenter was present at all stages of the experiment to answer questions and limit communication between subjects to a minimum.

gym again. Pairs were called into the gym and both students participated in the second run simultaneously on neighboring tracks.

After all pairs had finished their second suicide run, the experiment concluded with a short statement by the experimenters thanking the students for their participation. The teacher received a list of students' times in both runs and students were informed about their performance. We then asked the teacher to evaluate the general atmosphere within the class.[16]

3.3 Data Description and Manipulation Check

We present summary statistics of the students in our sample in Table 3.1.[17] In total, 39 classes with an average class size of about 25 students participated in the experiment. On average, 73% of students within each class subsequently took part in the experiment.[18] This amounts to 627 students who participated in the treatments, with 66% being female.[19] Due to odd numbers of students within some matching groups, we randomly dropped one student in those groups to match students in pairs. Therefore, some students participated in the experiment but were only recorded once and are dropped for estimating the treatment effects in the next section. This procedure yields an estimation sample of 588 observations.

On average, female students took 27.57 seconds (SD of 2.50 seconds) in the first run. Their performance is quite stable across grades, with students from the seventh grade being somewhat slower. Male students' times improved with age: while male students in grade 7 took on average 25.33 seconds in the first run, their performance improved by about two seconds on average in grade 10. In the following, we therefore control for these effects by including gender-specific grade fixed effects in all of our regressions. Independent of their treatment assignment, males and females

16. Teachers indicated their agreement with three statements on a seven-point Likert scale: (1) "The class atmosphere is very good", (2) "Some students get excluded from the group", and (3) "Students stick together when it really matters".

17. We focus on the students in the three main treatments, namely Random, Name and Performance and do not include the students from the NoPeer treatment, which is discussed in Appendix 3.D.

18. We aimed to recruit all students from a class. However, due to numerous reasons this was not possible in every class. Normally, some students are missing on a given day due to sickness or other reasons, are injured and cannot participate in the lesson, are not allowed to take part in the study by their parents or do not want to participate. Additionally, some students simply forgot to hand in the parental consent. We do not have concerns of non-random selection into the study since students did not know in advance the exact day when the experiment was scheduled and most reasons for non-participation were rather exogenous (like injuries or sickness). Moreover, treatment randomization was at the class level within schools and therefore selection into treatments is not possible.

19. We have more females in our sample since one school in our sample – the smallest one – was a female-only school.

Table 3.1. Summary Statistics

	7th grade	8th grade	9th grade	10th grade	Total
Socio-Demographic Variables					
Age	12.77	13.80	14.77	15.83	14.52
	(0.48)	(0.45)	(0.39)	(0.53)	(1.22)
Female	0.60	0.60	0.66	0.72	0.66
	(0.49)	(0.49)	(0.48)	(0.45)	(0.48)
Times (in sec)					
Time 1 (Females)	28.03	27.06	27.31	27.83	27.57
	(2.75)	(2.06)	(2.28)	(2.71)	(2.50)
Time 2 (Females)	26.98	26.46	26.47	26.94	26.72
	(1.97)	(1.74)	(2.43)	(2.37)	(2.23)
Time 1 (Males)	25.33	24.23	23.71	23.27	24.09
	(1.93)	(1.99)	(2.03)	(2.18)	(2.16)
Time 2 (Males)	24.62	23.58	22.85	22.35	23.31
	(2.01)	(1.99)	(1.70)	(1.50)	(1.98)
Class-level Variables					
# Students in class	25.54	26.00	26.25	25.03	25.68
	(2.71)	(1.96)	(2.56)	(3.17)	(2.74)
Share of participating students	0.75	0.69	0.77	0.71	0.73
	(0.11)	(0.14)	(0.16)	(0.13)	(0.14)
Share of Students in Treatments					
RANDOM	0.32	0.46	0.34	0.32	0.35
	(0.47)	(0.50)	(0.47)	(0.47)	(0.48)
NAME	0.37	0.25	0.37	0.35	0.34
	(0.48)	(0.43)	(0.49)	(0.48)	(0.47)
PERFORMANCE	0.32	0.29	0.29	0.33	0.31
	(0.47)	(0.46)	(0.46)	(0.47)	(0.46)
Observations	123	124	182	198	627

Notes: Standard deviations are presented in parentheses. Note that some students only participated in the survey in cases in which they were allowed to participate in the study but were unable to take part in the regular physical education lesson, while some others only took part in the first run if there was an odd number of students in the matching group. See the text for details.

Table 3.2. Share of Name-based Preferences Being Friends

Name-based preference	1st	2nd	3rd	4th	5th	6th	Average
Share of peers being friends	0.89	0.79	0.73	0.60	0.49	0.41	0.65

Notes: This table presents the share of friends for each name-based preference (most-preferred peer to sixth most-preferred peer as well as pooled over all six preferences) as elicited in the survey.

improved their performance in the second run by .78 seconds and .85 seconds on average, respectively.

We randomized classes into treatment and check whether observable characteristics differ between our treatments in Appendix Table 3.A.1. There are no observable differences across treatments for most variables, except for a difference in the pre-treatment times in the first run. However, this gap results from the randomization of classes into treatments and can be explained entirely by variation in observables. Conditional on gender-specific grade fixed effects, school fixed effects and age, these differences disappear.

3.3.1 Preferences for Peers and Manipulation Check

Before turning to the results of the experiment, we briefly present the preferences for peers elicited in the survey. Furthermore, we show that our peer assignment based on those preferences indeed changed the actual match quality, which we define as the rank of the assigned peer in the elicited preference rankings. This means that students in the self-selected treatments had a higher probability of being matched with someone who they preferred more, i.e., who ranked higher in their name- or performance-based preferences. Hence, our experimental variation of taking the preferences into account should have an effect on the rank of the assigned peers within a subject's preferences (i.e., the quality of that match) in the respective treatment with self-selection.

We summarize the preferences for peers according to name- and performance-based preferences in Table 3.2 and Figure 3.2, respectively. Two findings emerge: first, most students nominated friends as their most-preferred peer; and second, while students on average preferred to run with a slightly faster peer, there is a strong heterogeneity in this preference. We analyze these preferences in further detail in Kiessling, Radbruch, and Schaube (2019).

Figure 3.3 shows the realized match quality for all three treatments with respect to the ranking of peers in the two sets of elicited preferences. The upper panel shows the realized match quality according to name-based preferences. We observe that some people were randomly matched to someone with whom they would liked to be paired in Random and Performance. As expected, this share is rather low. While the median peer in Name corresponds to the most-preferred peer according to the elicited name-based preferences, the median peer is not part of the elicited

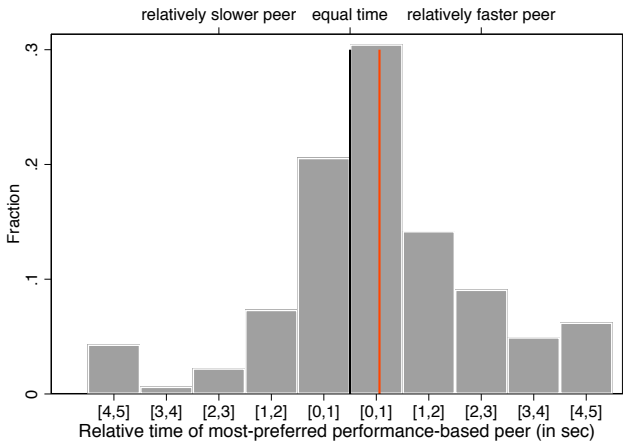


Figure 3.2. Most-preferred Performance-based Peer

Notes: The figure presents a histogram of the peer preferences over relative performance as elicited in the survey. Vertical lines indicate own time (black line; equals zero by definition) and the mean preference of all individuals (red line; 0.56 sec faster on average, where we used the midpoint of each interval to calculate the mean).

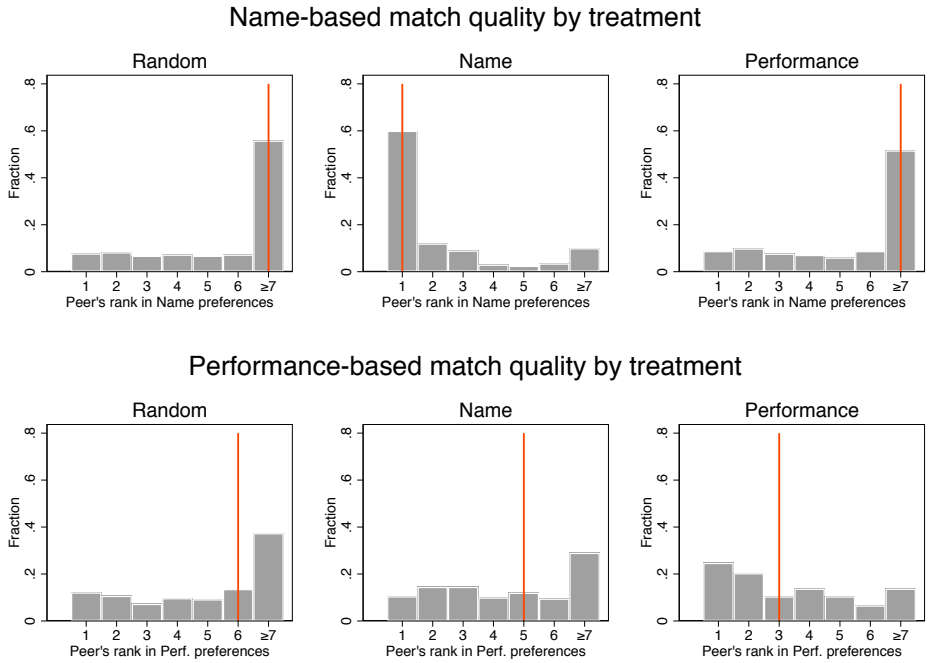


Figure 3.3. Match Quality Across Treatments

Notes: The figure presents a histogram of match qualities for each treatment measured by the rank of the realized peer in an individual's name- (upper panel) or performance-based preferences (lower panels). Vertical red lines denote median ranks.

preferences (i.e., not among the six most-preferred peers) for RANDOM and PERFORMANCE. A similar, albeit less pronounced picture arises when analyzing the match quality according to the preferences over relative performance as presented in the lower panel of Figure 3.3. We observe that students in PERFORMANCE were paired with more preferred peers according to their preferences relative to the other two treatments. However, subjects might have preferred other students or relative times that were not available to them, which mechanically affects the match quality. In Appendix 3.A, we check that once we take the mechanical effect into account, the median match quality in PERFORMANCE corresponds to the second most-preferred peer, i.e., we obtain a similarly pronounced pattern as in NAME.

3.4 Empirical Strategy

This section outlines our empirical framework. For this purpose, we first analyze the effect of being assigned to a particular peer assignment mechanism. In a second step, we decompose this change in performance into two effects: an indirect effect stemming from a change in the peer composition and a direct effect due to self-selection. Appendix 3.B derives these estimation equations from an economic model similar to a mediation analysis described in Heckman and Pinto (2015).

The random assignment of classes into treatments allows us to estimate the average effect of peer selection on performance. Let $D^d = 1$ with $d \in \{N, P\}$ denote treatment assignment to NAME and PERFORMANCE, respectively, and zero otherwise. We focus on percentage point improvements from the first to the second run, y_{igs}, of individual i in gender-specific grade g of school s as an outcome. Our baseline specification is then given by:

$$y_{igs} = \tau + \tau^N D_i^N + \tau^P D_i^P + \gamma X_i + \rho_s + \lambda_g + u_{igs} \tag{3.1}$$

The main parameters of interest are τ^N and τ^P, the effect of being assigned to one of our treatments relative to RANDOM. School fixed effects, ρ_s, and gender-specific grade fixed effects, λ_g, control for variation due to different schools (i.e., as a result of different locations and timing of the experiment) and variation specific to gender and grades.[20] Finally, X_i is a vector of predetermined characteristics such as age as well as personality characteristics and – in some specifications – class-level control variables, and u_{igs} is a mean zero error term clustered at the class level.

Any change in outcomes can be attributed to one of two main sources: first, different peer-assignment mechanisms may affect peer interactions directly; and second, self-selection may change the peer composition and therefore the difference between the student's and his or her peer's characteristics. To understand the source

20. See the section 3.3 for a discussion concerning why we include gender-specific grade fixed effects rather than gender and grade fixed effects separately.

of the average treatment effect, we decompose it into a direct effect of self-selection as well as a pure peer composition effect.[21] This takes into account the change in relative peer characteristics across treatments. We implement this decomposition using the following specification:

$$y_{igs} = \bar{\tau} + \underbrace{\bar{\tau}^N D_i^N + \bar{\tau}^P D_i^P}_{\substack{\text{Treatments} \\ \text{(direct effects)}}} + \underbrace{\beta\, \theta_i\left(D^N, D^P\right)}_{\text{Peer characteristics}} + \underbrace{\gamma X_i + \rho_s + \lambda_g}_{\text{Ind. characteristics and FE}} + u_{igs} \qquad (3.2)$$

We are interested in $\bar{\tau}_N$ and $\bar{\tau}_P$, the direct effects of our treatments relative to RANDOM. β denotes the influence of peer characteristics θ_i on the outcome. Changes in peer characteristics through our treatments are captured by changes in $\theta_i\left(D^N, D^P\right)$. In particular, we allow our effects to be mediated through several channels: a first set of channels capture the quality of the match measured by the rank of the peer in an individual's preferences[22], productivity differences measured by absolute differences of times in the first run, and (directed) friendship ties. We allow the effect of these to differ between the faster and slower student in a pair, given that previous research has shown that ranks affect peer interactions.[23]

While the existing literature to date has mainly concentrated on the influence of peers with respect to productivity differences and friendship ties on performance, our data allows us to go beyond this.[24] In particular, we allow for a second set of mediators based on the peer's personality and preference measures (i.e., Big Five, locus of control, competitiveness, risk attitudes, social comparison). Additionally, we also include the absolute difference in these personality measures to capture potential non-linear effects.

21. The direct effect mainly captures changes in performance due to being able to self-select a peer, which we interpret as an increase in autonomy (see section 3.5.5 for a discussion of the psychological underpinnings). We acknowledge that our definition of a direct effect also captures inputs that (i) differ across treatments, and (ii) are not measured in our rich set of potential mediators (match quality, friendship ties, productivity differences, ranks and personality differences). However, we show in robustness checks that in our setting this is of minor concern only.

22. We define two indicators to measure whether the assigned peer is nominated among the first three peers for name-based preferences or falls into the three highest ranked categories for performance-based preferences. Alternative specifications are shown in Appendix 3.E.

23. For example, beginning with Murphy and Weinhardt (2018), several studies document the importance of ranks for subsequent outcomes when peers interact with each other (Elsner and Isphording, 2017; Gill, Kissová, Lee, and Prowse, 2019). In a related manner, based on theoretical considerations, Cicala, Fryer, and Spenkuch (2018) show that individuals may select themselves into specific peer groups based on their rank within a prospective group, while Tincani (2017) sets up a model in which individuals have preferences over ranks and discusses how this can give rise to heterogeneous peer effects. Common across these studies is their emphasis on the importance of individual rank within groups for peer interactions.

24. Two exceptions include Chan and Lam (2015) and Golsteyn, Non, and Zölitz (2017), who study how peer personality traits affect one's own performance.

3.5 Results

Our experimental design allows to study the causal effect of different peer assignment mechanisms on individual performance. More specifically, we compare three treatments corresponding to random matching (RANDOM), matching with self-selected peers based on name-based peer preferences (NAME) and preferences over relative performance (PERFORMANCE). As outlined in section 3.2, the random assignment of peers constitutes a natural starting point for at least two reasons: first, the pure presence of any peer might already improve performance; and second, randomly assigned peers are used to document peer effects in a wide range of settings. We contrast this baseline condition with two treatments that assign peers based on elicited preferences, i.e., in which each subject endogenously chooses her peer.

Our empirical results start by documenting average treatment effects. As introduced in section 3.4, the average treatment effect can stem from two possible sources: if the (relative) characteristics of the peer affect performance and the treatments additionally induce a change in these characteristics, the altered peer composition might explain performance differences across treatments. Moreover, the ability to self-select a peer may directly influence the students' willingness to perform. Before we decompose each treatment effect into a *direct effect* of self-selection and an *indirect effect* due to changes in the peer composition, we establish two necessary conditions for the indirect effect to matter. First, we show that relative peer characteristics matter for individual outcomes. Second, we document that our treatments – which allow for self-selection – indeed change the relative characteristics of peers in the second run. We then decompose the average treatment effects into the two aforementioned channels. Our results conclude with an interpretation of the direct effect and a discussion of implications for peer assignment rules.

3.5.1 Average Effect of Self-selection on Performance

We analyze how average performance improvements differ between treatments. For this purpose, we use percentage point improvements as outcomes and therefore base our comparisons on the performance in the first run. This specification takes into account the notion that slower students (i.e., those with a slower time in the first run) can improve more easily by the same absolute value compared with faster students, as it is physically more difficult for the latter.

Figure 3.4 presents our first result. Subjects in RANDOM improve on average by 1.93 percentage points when paired with a random peer in the second run. However, their performance improves even more in NAME and PERFORMANCE by 3.22 and 3.58 percentage points, respectively. We present the corresponding estimates in Table 3.3. Columns (1)-(3) present the estimated percentage point improvements in time according to equation (3.1). Columns (4)-(6) additionally express the results in terms of times in the second run – and standardized times in column (7) – controlling

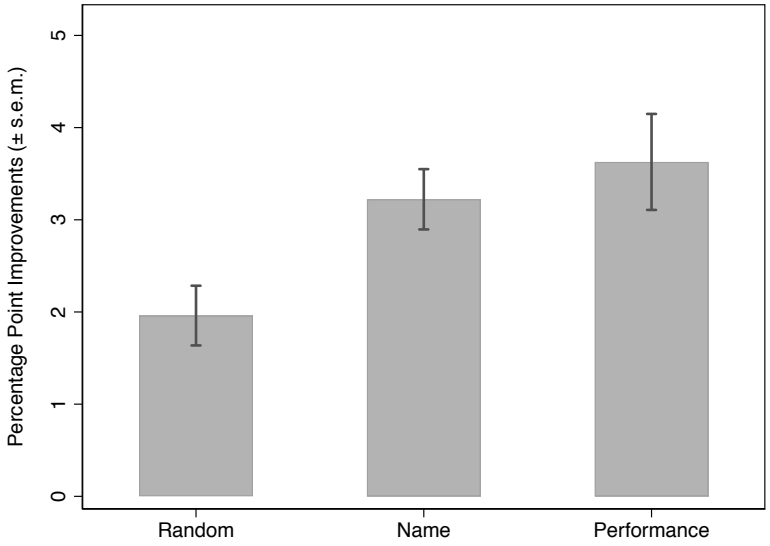


Figure 3.4. Average Performance Improvements

Notes: The figure presents percentage point improvements from the first to the second run with corresponding standard errors for the three treatments RANDOM, NAME, and PERFORMANCE corresponding to column (1) in Table 3.3. We control for gender, grade and school fixed effects as well as age and cluster standard errors at the class level.

Table 3.3. Average Treatment Effects

	(a) Percentage Point Imprv.			(b) Time (Second Run)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
NAME	1.26***	1.37***	1.84***	-0.38***	-0.38***	-0.48***	-0.14***
	(0.43)	(0.50)	(0.46)	(0.11)	(0.12)	(0.12)	(0.04)
PERFORMANCE	1.67**	1.69**	1.28**	-0.41***	-0.38***	-0.31**	-0.15***
	(0.62)	(0.65)	(0.60)	(0.14)	(0.14)	(0.14)	(0.05)
Time (First run)				0.69***	0.67***	0.71***	0.74***
				(0.04)	(0.04)	(0.05)	(0.04)
Class-level Controls	No	No	Yes	No	No	Yes	No
Own Characteristics	No	Yes	Yes	No	Yes	Yes	No
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	588	585	515	588	585	515	588
R^2	.056	.08	.096	.8	.81	.83	.8
p-value: NAME vs. PERF.	.51	.62	.38	.8	.98	.28	.8

Notes: This table presents least squares regressions according to equation (3.1) using percentage point improvements (panel (a)) and times of the second run controlling for times in the first run (panel (b)) as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Own and peer characteristics include the Big 5, locus of control, social comparison, competitiveness and risk attitudes. Class-level control variables in columns (3) and (6) include the share of participating students, three variables to capture the atmosphere within a class (missing for four classes), and indicators for the size of the matching group. Column (7) uses standardized times.

for times in the first run to confirm these effects in times rather than percentage point improvements. Assigning peers based on name-based preferences results in an additional 1.26 percentage point improvement in performance relative to the random assignment of peers. The coefficient for self-selected peers based on relative performance is 1.67 percentage points and thus somewhat larger, although it does not significantly differ from NAME (p-value= 0.51). These effects persist when controlling for students' own personal characteristics (column (2)) as well as if we additionally control for class-level variables capturing the atmosphere within a class (column (3)). Interestingly, the average treatment effects are about the same size as the improvement in RANDOM. On average, students are faster in the second run and this effect is nearly twice as large in PERFORMANCE and NAME compared to RANDOM. Our baseline effects correspond to additional time improvements of .38 to .41 seconds (cf. columns (4)-(6)) and account for 14% of a standard deviation in NAME and 15% in PERFORMANCE (cf. column (7)).[25,26]

3.5.2 Peer Characteristics Matter for Individual Improvements

Any decomposition of the average effect into a direct effect of self-selection and an indirect effect due to a change in the peer composition relies on two necessary conditions: first, peer characteristics need to be important for determining individual outcomes; and second, relative peer characteristics change when students can self-select their peers. We begin by providing evidence on the former condition, focusing on students in RANDOM. Therefore, we document the importance of peer characteristics by asking how much of the variation of performance improvements in RANDOM can be explained by variation in randomly assigned peer characteristics.

The intuition why peer characteristics may matter is that not all peers have the same effect on someone's performance. For example, friends who serve as a peer might influence us differently than other potential peers. Alternatively, the relative rank within a pair or productivity differences between peers may be driving individual outcomes. If some of these effects exist, then the variation in peer characteristics

25. Appendix 3.C presents additional robustness checks using biased-reduced linearization or group means to account for the limited number of clusters, specifications that control for outliers and reports the average treatment effects for different subgroups (by gender, grade, school). Our results are robust to all of these checks.

26. In Appendix 3.D, we document that the observed performance improvements in the three treatments described here are a result of the presence of peers and not due to learning. We present the results of an additional control treatment (NOPEER) and its implementation details. In the control treatment, subjects run twice without any peer and we find that they do not improve their time from the first to the second run; in fact, individual performance decreases. The improvements that we observe here can therefore be attributed to the presence of peers rather than learning or familiarity with the task.

can explain some of the variation in the performance improvements of subjects in the data and in particular when randomly assigning those characteristics in RANDOM.[27]

In order to show the relevance of peer characteristics, we decompose the coefficient of determination, R^2, into variation that is attributable to individual characteristics and peer characteristics.[28] Note that we cannot estimate partial models to obtain the fraction of variance explained by a set of predictors as an individual's and her peer's characteristics may be correlated (e.g., since both are from the same age group and age is related to performance, as documented in Table 3.1). We account for this interplay between different groups of explanatory variables by employing a variance decomposition based on Shapley values to calculate the marginal contribution of each group of variables (see Huettner and Sunder, 2012).

We base the variance decomposition on data from RANDOM only and estimate equation (3.2) to decompose R^2 into components attributable to individual as well as peer characteristics.[29] As Table 3.4 reports, we find that 20% of the total variation in percentage points improvements in individual performance can be attributed to characteristics of the peer, which corresponds to 78% of the explained variation. Consequently, only 6% of the total variation or 22% of the explained variation stems from individual characteristics.[30]

The decomposition therefore shows the importance of accounting for peer characteristics in general. Characteristics of peers are responsible for a large share of the explained variance. Hence, we need to take these peer characteristics into account for the analysis of our treatments.

3.5.3 Self-selection Changes the Peer Composition

In this section, we document that treatments that allow for self-selection change with whom someone interacts. Although relative peer characteristics are important

27. Note that only relative characteristics within a pair can help to explain differences between treatments. Since we randomize subjects into treatments, the overall distribution of peer characteristics across treatments and within classrooms remains constant. Our treatments only change with whom each student interacts within a class, and thus a peer's characteristics relative to one's own characteristics.

28. As peer characteristics, we include the rank within a pair itself as well as the rank interacted with match quality with respect to both sets of preferences, friendship indicators and productivity differences. We also include personality traits of a peer and absolute differences in personality traits between peers. This corresponds to the full specification that we also use in our decomposition (col. 5 of Table 3.6).

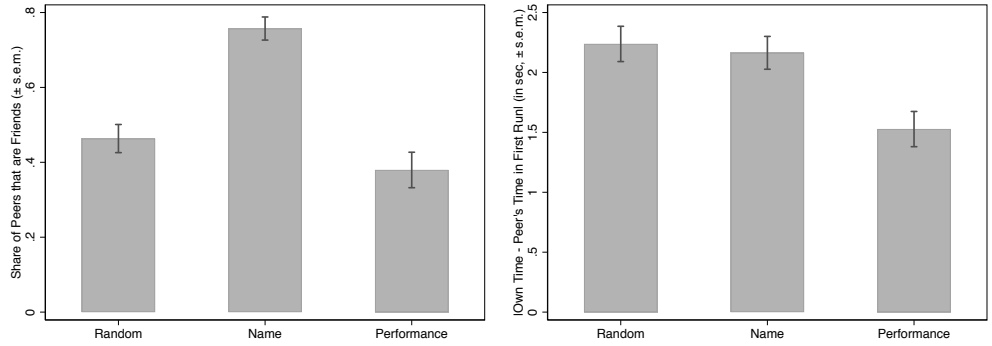
29. The corresponding estimates are delegated to column (1) of Appendix Table 3.E.5.

30. Note that we explain percentage point improvements from the first to the second run and hence much of the individual-level variation is already taken out of the dependent variable. When using time in the second run as an outcome variable, individual characteristics account for approximately 54% (67% when additionally controlling for time in the first run) of the explained variation ($R^2 = 0.70$ without time in the first run, $R^2 = 0.79$ with time in the first run), while peer characteristics explain the remainder of R^2. Nevertheless, the variation explained from peer characteristics remains sizable.

Table 3.4. Variance Decomposition of Performance Improvements in RANDOM

		Variation attributable to			
Explained variation (R^2)		Peer characteristics		Individual characteristics	
.26	*(100%)*	.2	*(78%)*	.06	*(22%)*

Notes: This table presents a decomposition of the coefficient of determination, R^2, using Shapley values and is based on equation (3.2) estimated on RANDOM only.



(a) SHARE OF PEERS BEING FRIENDS **(b)** ABSOLUTE DIFFERENCES IN PRODUCTIVITY

Figure 3.5. Changes in Peer Composition

Notes: Figure 3.6a presents the share of all students who nominated their assigned peer as a friend for each of the three treatments including standard errors. Figure 3.6b shows the average absolute within-pair difference in productivity (measured in times from the first run) and including standard errors for each treatment. We control for gender, grade and school fixed effects as well as age and cluster standard errors at the class level. We present the corresponding regressions and highlight additional compositional differences of the treatments in Appendix Table 3.A.2.

for understanding outcomes – as shown in the previous section – students also need to interact with systematically different peers when self-selecting them. A second necessary condition for the indirect effect is therefore that the relative peer characteristics have changed.

Figure 3.5 shows that our treatments indeed changed the peer composition with respect to two prime examples of peer characteristics, namely friendship ties and productivity differences within pairs. More specifically, Figure 3.6a shows that students are predominantly paired with friends in NAME (76% of all peers are friends), whereas the share of peers being friends in RANDOM and PERFORMANCE is 49% and 37%, respectively. As matching based on preferences over relative performance (PERFORMANCE) allows for targeting of other students with a similar or slightly higher productivity, the students' absolute time differences in the first run might change. Panel B of Figure 3.6b confirms this by showing that the average absolute difference in times from the first run is 1.53 seconds in PERFORMANCE, while it is

Table 3.5. Decomposition of Treatment Effects

	Direct Effects		Indirect Effects	
	PP imprv.	Std. Err.	PP imprv.	Std. Err.
NAME	1.24	0.50	0.13	0.24
PERFORMANCE	2.21	0.68	-0.52	0.23

Notes: The table presents the resulting direct and indirect effects from a decomposition according to equation (3.2) shown in column (5) of Table 3.6. Indirect effects are defined as the changes in percentage point improvements that are explained by changes in peer characteristics relative to RANDOM and comprises the combined effect of all peer characteristics in column (5) of Table 3.6.

larger than two seconds in the other two treatments (2.24 and 2.16 seconds in RANDOM and NAME). Even though students could mainly target peers along these two dimensions, we present how our treatments affect the peer composition along various other characteristics in Appendix Table 3.A.2. We find that targeting specific peers also results in systematically different peers in terms of their personality.

This establishes that self-selection changes with whom somebody interacts. The endogenously selected peers are neither equal to random peers nor to the average peer. Their characteristics differ with respect to several important dimensions.

3.5.4 Decomposition Into Direct and Indirect Effects of Self-selection

We now decompose the average treatment effects from Table 3.3 by taking changes in the peer composition explicitly into account. As outlined in section 3.4, the estimated average effects potentially comprise a direct effect as a result of self-selection and an indirect effect stemming from interacting with different peers. This is the case as our treatments have two features: on the one hand, our treatments change with whom someone interacts and those peer characteristics matter as documented above; and on the other, they change the selection procedure from exogenous assignment to the self-selection of peers. The indirect effect therefore captures changes in the relative characteristics of peers (e.g., the time differences between the student and peer in the first run) due to the altered peer composition induced by being able to select them. The direct effect captures the effect of the treatment due to a change in the selection rule. The previous two subsections documented that NAME and PERFORMANCE change the peer composition relative to RANDOM and established that those relative peer characteristics are important in determining individual outcomes. The decomposition analyzes the extent to which the average treatment effects are driven by these changes in the peer composition.

The results of the decomposition based on equation (3.2) are summarized in Table 3.5 and presented in further detail in Table 3.6. In Table 3.5, we use the whole set of characteristics to decompose the average treatment effects into the direct and

indirect effects. Therefore, the size of the direct effects equals the coefficients of the treatment indicators in column (5) of Panel A in Table 3.6. They correspond to 1.24 percentage points in NAME and 2.21 in PERFORMANCE.

The decomposition shows that even though peer characteristics are highly important in understanding the variation in outcomes, the indirect effects of self-selection in the two treatments are considerably low. They correspond to only 11% of the size of the direct effect in NAME and 24% in PERFORMANCE.[31] In NAME, we estimate a positive and insignificant indirect effect of .13 percentage point improvements (p-value = 0.59). This means that the altered peer characteristics have only a slightly positive effect on the students' performance. For PERFORMANCE, we find a significant indirect effect of -.52 percentage points (p-value = 0.03). Thus, the change in the peer composition even magnifies the direct effect as it negatively rather than positively affects performance.

Therefore, our decomposition shows that while self-selection of peers indeed changes the composition of peers, these changes cannot explain the average treatment effects; rather, the additional performance improvements in NAME and PERFORMANCE stem from a direct effect of self-selection.

We now analyze the detailed results of the decomposition in Table 3.6. Column (1) replicates the baseline estimates from column (2) of Table 3.3 for means of comparison. In columns (2)-(4), we include different sets of peer characteristics, before we include all of them in column (5). Turning to the separate columns, we find that the size of the treatment indicators only slightly differ across specifications. Nonetheless, some of the included peer characteristics influence the individual performance in the second run. Performance-based match quality has some predictive power for performance improvements in the restricted regression in column (2). However, the effects are insignificant when controlling for all peer characteristics in column (5). Overall, the quality of the match, i.e., how well a student's preferences were satisfied by the pairing in the second run, has little to no effect on their performance. We also observe that initially faster students within a pair reduce their performance when paired with a friend, while the relatively slower students do not adjust their performance differentially for friends as peers (column (3) and (5)). In column (4), we focus on productivity differences, since faster and slower students within a pair might be affected differentially. We also allow the effect of productivity differences, $|\Delta Time1|$, to differ by the rank within a pair. We find that differences in times of the first run have a significant effect on both faster and slower students within a pair.

31. The indirect effect in our decomposition is induced by the impact of peer characteristics and their change through self-selection.Therefore, it corresponds to the difference in the average effect for NAME and PERFORMANCE and the direct effect as the direct and indirect effect add up to the average effect. The indirect effect also corresponds to multiplying the coefficients for (relative) peer characteristics from column (5) with the change in the peer composition across treatments, as described in Appendix 3.B and Appendix Table 3.A.2.

Table 3.6. Decomposition of Treatment Effects

	Percentage Point Improvements					
	(1) Baseline	(2) Match Quality	(3) Friend-ship ties	(4) Time Difference	(5) All	(6) Class Controls
Direct Effects						
NAME	1.37***	1.23**	1.46***	1.35***	1.24**	1.46***
	(0.50)	(0.53)	(0.49)	(0.46)	(0.50)	(0.46)
PERFORMANCE	1.69**	1.78***	1.61**	1.84***	2.21***	1.73**
	(0.65)	(0.65)	(0.64)	(0.61)	(0.68)	(0.68)
Peer Characteristics						
Faster Student		0.00			0.52	0.69
× High match quality (NAME)		(0.39)			(0.43)	(0.45)
Slower Student		0.31			0.46	0.62
× High match quality (NAME)		(0.61)			(0.66)	(0.74)
Faster Student		1.17**			0.43	0.12
× High match quality (PERF.)		(0.52)			(0.53)	(0.59)
Slower Student		-2.07***			-0.71	-1.15
× High match quality (PERF.)		(0.61)			(0.66)	(0.73)
Faster Student			-0.77*		-1.15**	-1.03**
× Peer is Friend			(0.45)		(0.53)	(0.47)
Slower Student			-0.06		0.13	0.45
× Peer is Friend			(0.53)		(0.67)	(0.79)
Faster Student				-0.39***	-0.35**	-0.36**
× $\lvert\Delta Time\ 1\rvert$				(0.14)	(0.16)	(0.16)
Slower Student				1.03***	1.04***	0.84***
× $\lvert\Delta Time\ 1\rvert$				(0.21)	(0.20)	(0.19)
Slower Student in Pair		3.85***	2.20***	-0.17	-0.15	0.11
		(0.44)	(0.49)	(0.45)	(0.68)	(0.76)
Abs. Diff. in Personality	No	No	No	No	Yes	Yes
Peer Characteristics	No	No	No	No	Yes	Yes
Own Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Class-level Controls	No	No	No	No	No	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes	Yes	Yes
N	585	585	585	585	582	512
R^2	.08	.18	.15	.24	.29	.29
p-value: NAME vs. PERFORMANCE	.62	.41	.82	.43	.17	.72

Notes: This table presents least squares regressions according to equation (3.2) using percentage point improvements as the dependent variable. High match quality is an indicator that equals one if the partner was ranked within an individual's first three preferences. Personality characteristics include the Big Five, locus of control, social comparison, competitiveness, and risk attitudes. Appendix Table 3.E.6 presents the omitted coefficients of own and peer characteristics, and their absolute differences for our preferred specification in column (5). *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level.

Table 3.7. Variance Decomposition and the Role of Unobservables

Panel A: Variance decomposition							
		Variation attributable to					
Explained variation (R^2)		Treatments		Peer characteristics		Individual characteristics	
0.29	*(100%)*	0.03	*(12%)*	0.21	*(72%)*	0.05	*(16%)*

Panel B: Role of unobservables			
	Oster's δ		
	$R^2_{max} = 0.50$	$R^2_{max} = 0.75$	$R^2_{max} = 1.00$
NAME	2.54	1.19	0.78
PERFORMANCE	-7.05	-3.43	-2.27

Notes: Panel A decomposes the explained variance of specification (5) of Table 3.6 in components attributable to treatments, peer and individual characteristics similar to Table 3.4. Panel B quantifies the importance of unobservables relative to observables needed for zero direct effects according to Oster (2019).

While slower students within a pair benefit by a 1.03 percentage point improvement from running with a one second faster student, the relatively faster student's performance suffers from this productivity difference by .39 percentage points. In sum, the average performance of a pair thus improves with increasing differences in productivity.

We control for all of these characteristics jointly in column (5), where we also add a rich set of relative peer personality characteristics. The effect of friendship ties on the initially faster students as well as the effects on productivity differences persist. More importantly, the direct effects of both NAME and PERFORMANCE remain robust, showing a direct effect of self-selection on individual performance. In order to further probe the robustness of this finding, we additionally control for proxies of the class attitude in column (6). While the estimates slightly differ in magnitude, the results are generally robust. However, as we lose some observations, our preferred specification is column (5).

Our results, therefore, provide evidence for a direct effect of self-selection. In the remainder of this section, we provide further evidence for its robustness. First, in Panel A of Table 3.7 we replicate the variance decomposition of Table 3.4 for all three treatments and confirm the importance of the peer characteristics in terms of explaining the variation in outcomes. Second, in Panel B of Table 3.7 we address the possible concern that other characteristics for which we cannot account or control are driving the direct effect. Our results above remain relatively stable when adding different sets of peer controls, which is reassuring. A more formal approach to tackle this concern is to ask how important unobserved characteristics would have to be to explain our direct treatment effects (Altonji, Elder, and Taber, 2005; Oster,

2019). We follow Oster (2019) and calculate δ, a measurement for the relative importance of unobserved characteristics compared to observed characteristics. This measure describes how important unobserved variables would have to be relatively to observed ones to explain the direct effects, i.e., to drive down the direct effects to zero. Absolute values of δ larger than one indicate that these omitted variables have to be relatively more important than observed peer characteristics. Negative values indicate that those unobservable characteristics need to reverse the effect of observed covariates. We calculate these measures for three scenarios that differ in the maximum amount of variance that would theoretically be explained if all factors that might affect the outcomes were observed. More specifically, we calculate δ for R^2_{max} equal to 0.50, 0.75 and 1.00. In all but one extreme scenario the omitted peer characteristics are required to be more important than the observed peer characteristics. This suggests that such unobserved characteristics need to have a larger effect than productivity differences, friendship ties, match quality and all other controls – including personality traits – combined. Compared to other studies, our analysis already allows for more peer characteristics to influence subjects' behavior. Therefore, we allow for a very rich set of important characteristics and conclude that such unobserved characteristics are highly unlikely to drive the direct treatment effects.

In addition, we provide several robustness checks in the Appendix 3.E. Appendix Table 3.E.1 allows for different specifications of match quality by additionally considering the partner's match quality, an interaction between one's own and the partner's match quality, as well as feasible match quality. Appendix Table 3.E.2 considers different definitions of friendship ties apart from directed links (i.e., undirected, reciprocal, directed and reciprocal friendship ties). The results for all robustness checks remain qualitatively and quantitatively similar. Furthermore, we show in Table 3.E.3 and Appendix Figure 3.E.1 that the linear specification of productivity differences is not restrictive. Appendix Table 3.E.4 estimates the coefficients of peer characteristics on the subsample of students in RANDOM only and imposes these coefficients on the other treatments. Furthermore, Table 3.E.5 presents the robustness of the direct effects to using only those subjects in RANDOM who are matched in line with their preferences. These matches occurred by pure chance and not due to self-selection. All of these robustness checks support our conclusion.

Taken together, our analysis shows that self-selection improves individual performance directly and not due to a change in the peer composition. This means that subjects react to observationally similar peers differently once they have chosen them actively. Characteristics of peers are important in determining outcomes, but they do not explain the average treatment effects of self-selection, which are driven by the direct effect of self-selection. Although our treatments allowed for two different notions of self-selection, it is reassuring that the estimates of the direct effects are similar across treatments.

3.5.5 Interpretation of the Direct Effect

We interpret the direct effect as a positive effect of self-selection due to increased control or autonomy over the peer assignment mechanism. However, one might worry that knowledge of all three treatment conditions could lead students in RANDOM to react negatively due to disappointment that their preferences have not been taken into account.[32] If these disappointed students drove our findings, we would falsely attribute effects to self-selection even if students in NAME and PERFORMANCE do not react positively.[33] If the direct effect originated from disappointment, we would expect students in RANDOM to have less fun in the experimental task. Therefore, in column (1) of Appendix Table 3.F.1 we analyze the extent to which subjects across treatments had different perceptions regarding their fun in the second run. We find zero effects. The absence of direct effects in the fun dimension alleviates the potential concern that knowledge of all three treatments leads to disappointment when students are assigned to RANDOM.[34]

We therefore conclude that the direct effects in our experiment are due to positive effects of self-selection. More specifically, we argue that the opportunity to self-select key aspects of one's environment – in our experiment having autonomy over the peer selection – has a direct effect beyond the instrumental value of changing peer characteristics. Self-determination theory provides a credible explanation through which self-selection can impact performance directly. The theory identifies autonomy as a crucial determinant of motivation: individuals who can actively se-

32. This results from the fact that we elicited preferences for peers irrespective of the treatment and only announced the assignment rule after the survey, but before the second run.

33. At the same time, this also describes a feature of many real-world settings. Imagine that a person is randomly assigned a partner from a group of available people. Even if this person has not been asked explicitly with whom she would like to interact, she still has preferences about interacting with certain people. Therefore, disappointment could also play a role in these settings. This might be true for all settings that feature exogenous assignment and overrule the underlying preferences of the involved persons.

34. A related issue would be that the direct effect stems from a positive effect of subjects in treatments with self-selection as they may react reciprocal towards being treated kindly (see Aldashev, Kirchsteiger, and Sebald, 2017, for an analysis how reciprocity can influence treatment effects). If students prefer to be in one of the self-selection treatments (NAME or PERFORMANCE) rather than in RANDOM and they perceive their assignment as kind, reciprocal students could respond by increasing their performance. This in turn would imply that the direct effects of our treatments are due to reciprocity or some kind of experimenter demand effects. Then prosocial students should display a stronger (direct) effect than non-reciprocal students as they are are more likely to react reciprocally. We proxy prosociality by scoring higher on the agreeableness scale of the Big Five as it is significantly correlated with reciprocity and altruism (Becker, Deckers, Dohmen, Falk, and Kosse, 2012). Column (3) in Appendix Table 3.F.1 reports the interaction between the agreeableness score and treatment indicators. If the above motives are the underlying causes of the direct treatment effect, we should observe a positive and statistically significant interaction between agreeableness and the treatments. However, our results do not show this relationship. We interpret this finding as evidence against reciprocal motives driving our results.

lect parts of their environment – most importantly their tasks in work environments – display higher intrinsic motivation (Deci and Ryan, 1985, 2000).[35] Applying this explanation to our setting suggests that not the selected peer herself increases motivation, but the mere act of selecting her. However, we do not argue that this behavioral effects stems from self-selecting any aspect, but a relevant aspect of one's environment.

Self-determination theory and autonomy in particular have recently gained increasing attention from economists. Cassar and Meier (2018) review the economic literature on non-monetary aspects of work environments in the light of self-determination theory and highlight the importance of autonomy for various behavioral outcomes. A related argument to ours also underlies the findings of Bartling, Fehr, and Herz (2014) and Owens, Grossman, and Fackler (2014). Although they do not focus on the effect of autonomy on subsequent outcomes, their studies demonstrate that people have a willingness to pay for making decisions by themselves and maintaining autonomy. Similarly, a growing body of literature demonstrates that restricting subjects choice sets and therefore restricting their autonomy and freedom can negatively influence outcomes (e.g., Falk and Kosfeld, 2006). Therefore, our results add to this literature by highlighting the motivational benefits of autonomy and self-determination, and provide novel field evidence that having control positively affects outcomes.

3.5.6 The Limits of Reassignment Rules

Our results show that self-selected peers lead to substantially larger performance improvements than randomly assigned peers. In practice, however, policy makers frequently do not assign peers at random. Rather, they employ a variety of peer assignment rules to help or target specific individuals. Examples include schools employing tracking (e.g., Betts, 2011; Duflo, Dupas, and Kremer, 2011; Fu and Mehta, 2018; Garlick, 2018) or pairing high-performing students with low-performing ones (e.g., Carrell, Sacerdote, and West, 2013). While we have not conducted these treatments in our context, we can use our estimates to simulate the effect of such exogenous peer assignment rules and compare their effect to outcomes under self-selection.

For this purpose, we use our estimates obtained in section 3.5.4, using the whole set of peer characteristics (column (5) of Table 3.6). Based on these estimates, we simulate different (exogenous) assignment rules, calculate the resulting effects on performance, and compare them to performance improvements observed in our experiment. We first compare the improvements to the counterfactual of assigning the same peers in Name and Performance without the direct effect of self-selection.

35. Two other components of self-determination theory are relatedness and competence, referring to the need to care about something and the need to feel challenged, respectively. In our experiment, we hold these other components constant across treatments.

A comparison of other peer assignment rules with these results sheds light on the question of whether students are able to choose optimal peers. Second, we simulate the expected performance improvements under a random matching. Third, we use several assignment rules that base the assignment on one single and commonly employed peer characteristic, namely past performance. Our estimates obtained in section 3.5.4 suggest that pairs with a higher difference in initial performances will improve their performance on average. If this is the only characteristic of a peer that affects performance, aggregate performance would be maximized as long as the sum of productivity differences within a pair is maximized.[36] In order to compare the results of self-selection against exogenous assignment rules that promise the largest aggregate improvements, we consider two matching rules that maximize these productivity differences within pairs – Equidistance and High-to-Low – that keep the distance in ranks within the class constant or pair the best-performing student with the slowest student. Additionally, we look at the effect of tracking (i.e., pairing the best student with the second best, third with the fourth, etc.; Tracking). We compare the predicted performance improvements for those rules with our estimated performance improvements for the three assignment rules used in the experiment.[37]

Figure 3.6 presents the simulated average performance improvements of each assignment rule. The results show that no other peer assignment rule is able to reach similar performance improvements as those featuring self-selection. In fact, they are close to the results from our random matching, since students under those peer assignment rules do not benefit from the additional intrinsic value of self-selection. We observe that in the absence of a direct effect of self-selection, students do not experience additional improvements relative to randomly assigned peers. Compared to Equidistance and High-to-Low, students in Name (exog.) and Performance (exog.) perform worse indicating that they do not choose their peers optimally.

More surprisingly, the reassignment rules that maximize productivity differences in pairs – Equidistance and High-to-Low – do not improve average performance compared to the random assignment of peers. Although both rules increase the average productivity difference in pairs by construction and affect performance through this channel, those rules also change other characteristics of the peer. The lack of any additional improvement implies that these other changes in peer characteristics offset the positive effect of increased productivity differences. This highlights important consequences of peer effects that are multidimensional if one wants to enhance overall performance.

36. Given our specification, this is true for all peer-assignment rules that match each student from the bottom half of the productivity distribution with a student from the top half.

37. We provide details on the prediction of performance improvements and the peer assignment rules in Appendix 3.H.

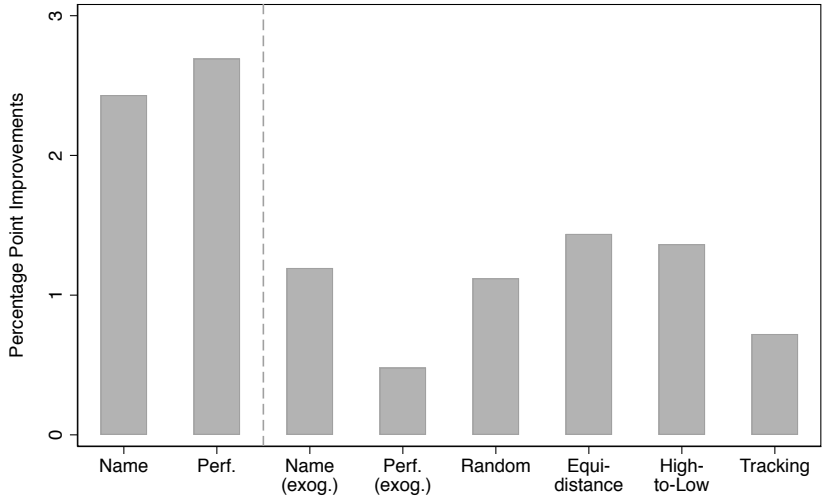


Figure 3.6. Simulation of Other Peer Assignment Rules

Notes: The figure presents predicted percentage point improvements for the two treatments (NAME, PERFORMANCE) with and without the effect of self-selection, the RANDOM-treatment as well as three simulated peer assignment rules (EQUIDISTANCE, HIGH-TO-LOW and TRACKING). We fix the personal characteristics and other covariates not at the pair level to 0, whereby effect sizes are therefore not directly comparable to treatment effects above. More details are provided in the text and Appendix 3.H.

This result suggests that reassignment rules based on specific characteristics may not work as intended given that other characteristics may affect performance at the same time. Thus, depending on the correlation structure between the characteristic used for the peer assignment rule and the omitted characteristics as well as their effect, the resulting outcomes may be either higher or lower than predicted. If peer effects are multidimensional, policy makers need to take all potential characteristics into account when reassigning students into peer groups. Consequently, designing optimal peer assignment rules might be more challenging than expected.[38] This insight further helps to understand why we observe a very small indirect effect in the decomposition of the treatment effects despite the fact that peer characteristics help to explain much of the variation in individual outcomes (cf. Table 3.6).

The simulations above suggest that self-selection of peers can be an attractive alternative compared to traditional peer assignment rules to increase individual performance. However, we want to stress that such peer assignments based on self-selection may also come at a cost. In particular, we show in Appendix Table 3.G.1 that students in PERFORMANCE experience significantly more pressure compared to the other two treatments, and individual ranks may be more perturbed between the

38. In general, designing optimal peer assignment rules requires an optimization taking into account all potential dimensions in which peers may exert effects. This creates a high-dimensional optimization problem that is highly difficult to solve.

two runs in NAME and RANDOM relative to PERFORMANCE. Hence, a policy maker might not only look at the resulting performances but also how different assignment rules affect the individuals' overall well-being.

3.6 Conclusion

Peer effects are an ever-present phenomenon discussed in a wide range of settings across the social sciences. For many situations, identifying the effect of an actively self-chosen peer is important beyond estimating peer effects in general. Our framed field experiment introduces a novel way to study the self-selection of peers in a controlled manner and is able to separate the impact of a specific peer on a subject's performance from the overall effect of self-selection. The results of our experiment provide evidence that self-selecting peers yields performance improvements of about 15% of a standard deviation relative to random assignment of peers. While peer characteristics affect the individual performance, they are not the origin of the estimated treatment effects. Rather, these improvements stem from a direct effect of self-selection. Based on self-determination theory (Deci and Ryan, 1985), we interpret this direct effect such that the ability to select one's own peer enhances a student's intrinsic motivation and subsequently increases individual performance.

Teachers or supervisors might be interested to leverage this direct effect of self-selection in addition to other forms of non-monetary incentives used in schools (Levitt et al., 2016) or workplaces (Cassar and Meier, 2018). They may allow students to choose their study group themselves or introduce flexible seating patterns in offices such that employees can self-select their seat mates, office partners or colleagues. Since our results suggest that self-selecting peers improves outcomes, the effectiveness of social comparison interventions in general may be improved if individuals are given the opportunity to select their relevant comparison themselves rather than being assigned an unspecific one.

One might be eager to infer that our results give rise to a trade-off between performance improvements as a result of self-selection per se and the exogenous assignment of performance-maximizing peers. However, our simulations show that exogenous reassigning rules, which try to lever peer effects in ability, have an impact close to zero in our case and are in general ambiguous in size and sign. This result relies on the existence of peer effects in multiple dimensions, which at least partially offset each other and in turn limit the effectiveness of exogenous reassignment rules. Hence, positive effects of peer self-selection might be performance-maximizing – even in the absence of subjects choosing "optimal" peers.

Our experimental design can easily be transferred to situations in which other production functions are used or where peer effects arise via other channels, e.g., implementing team production by reporting a function of both students' times to the teacher, or varying the task to allow for learning or skill complementaries as sources

of peer effects. In those settings, it is reasonable to assume that self-selection of peers may happen or can be implemented. For example, study groups at universities often form endogenously (Chen and Gong, 2018), researchers select their co-authors and workers in firms increasingly form self-managed work teams (Lazear and Shaw, 2007), and employees self-select with whom they work by referring others to their employer (Lazear and Oyer, 2012; Friebel, Heinz, Hoffman, and Zubanov, 2019).

In this paper, we highlight that self-selecting peers can serve as a complement to other established methods such as incentives and exogenous peer assignment policies aimed at increasing individual performance. However, further research on the interplay between endogenous group formation, social interactions and production environments remains imperative to understand how peer effects work.

Appendix 3.A Randomization and Manipulation Check

Table 3.A.1 presents the randomization check of our experiment. The residual of times in the first run are constructed from a regression of times of the first run on school and grade-specific fixed effects as well as age. As can be seen the difference in times in the first run can be explained by those observables and hence are an artifact of the block randomization as classrooms rather than individuals were randomly assigned to treatments.

Table 3.A.1. Randomization Check

	Random	Name	Diff.	Performance	Diff.
Socio-Demographics					
Age	14.43	14.55	0.13	14.58	0.15
	(1.18)	(1.24)	(0.12)	(1.24)	(0.12)
Female	0.73	0.62	-0.11*	0.61	-0.12*
	(0.45)	(0.49)	(0.04)	(0.49)	(0.05)
Doing sports regularly	0.82	0.82	0.00	0.90	0.08
	(0.39)	(0.38)	(0.04)	(0.31)	(0.04)
Times (in sec)					
Time (First Run)	26.81	26.08	-0.73*	26.19	-0.62*
	(2.96)	(2.93)	(0.28)	(2.78)	(0.28)
Residual of Time (First Run)	-0.02	-0.11	-0.09	0.08	0.10
	(2.31)	(2.35)	(0.22)	(2.24)	(0.22)
Class-level Variables					
# Students in class	26.01	25.39	-0.62*	25.61	-0.41
	(2.95)	(2.02)	(0.24)	(3.11)	(0.30)
Share of participating students	0.72	0.74	0.02	0.73	0.01
	(0.16)	(0.13)	(0.01)	(0.12)	(0.01)
Grade	8.68	8.76	0.08	8.75	0.07
	(1.07)	(1.12)	(0.11)	(1.13)	(0.11)
Observations	221	213	434	193	414

Notes: *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard deviations in parentheses in columns 1, 2 and 4; standard errors in column 3 and 5. Residuals of Time (First Run) are calculated as follows: We first regress all times from the first run on school, grade and gender fixed effects. We then use the residuals from this regression.

In section 3.3.1, we presented the resulting match qualities using the preferences as elicited in the survey. However, some subjects may prefer relative times, which are not available to them. For example, the fastest subject in the class might want to run with someone who is even faster, or a student wants to run with somebody else who is 1-2 seconds faster but by chance there is no one in the class with such a time. Similarly, subjects in Name may rank other students which were not present during the experiment or did not participate. We therefore present an alternative approach

to evaluate the match quality by taking the availability of peers into account. This implies that the quality of a match does not correspond directly to the elicited preferences; rather, based on these preferences all available subjects (i.e., the students participating in the study) are ranked. The quality of the match is then calculated based on this new ranking and results in a realized feasible match quality.

Consequently, we determine the feasible match quality by calculating how high a classmate is ranked in a list of available classmates. [39] In NAME, this can only increase the match quality. If someone nominates another student who is not available as her most-preferred peer and she received her second highest ranked choice, this means that she is matched with her most-preferred feasible peer. Similar arguments can increase the match quality for preferences over relative performance. However, the match quality in performance can also be lower. Suppose that a student ranks the category "1-2 seconds faster" highest and there are three students in that category. However, she is only matched with her second highest ranked category. There would have been three subjects whom she would have preferred more, generating a feasible match quality of 4. We present the corresponding histograms in Figure 3.A.1 and observe that the median of the feasible match quality is actually higher for both treatments relatively to the match qualities depicted in Figure 3.3.

As our treatments change the peer composition, they also change the relative characteristics of peers. In order to understand which characteristics change, we analyze how our treatments affect the peer composition in other dimensions apart from the match quality in Table 3.A.2.

39. We code peers who are not ranked among the first six preferences with a match quality of 7.

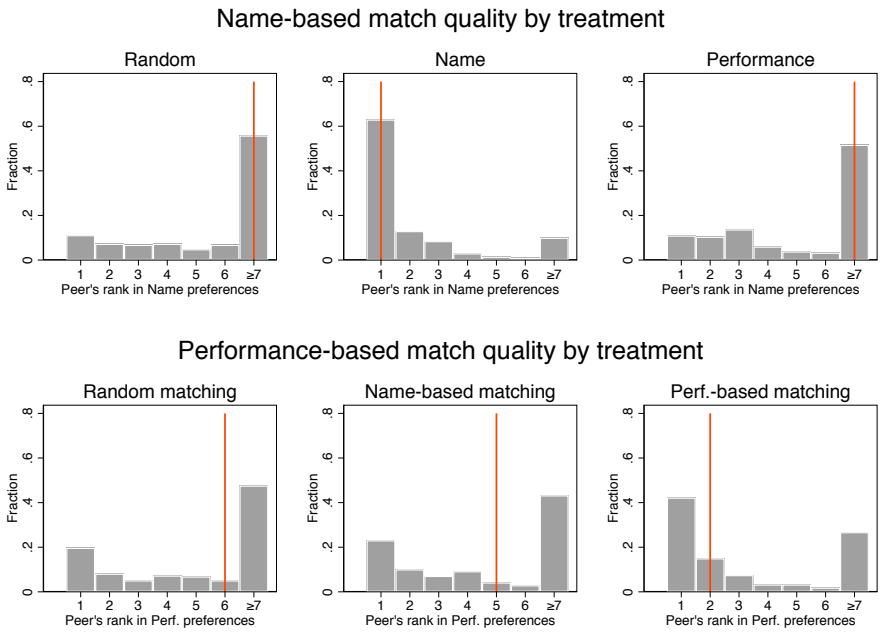


Figure 3.A.1. Feasible Match Quality Across Treatments

Notes: The figure presents a histogram of match qualities for each treatment evaluated according to either the students' name-based preferences (upper panel) or performance-based preferences (lower panel). Vertical lines denote median match qualities.

Table 3.A.2. Effects of Treatments on Peer Composition

	Match Qual. (name)	Match Qual. (time)	Friendship Ties	Time 1	
NAME	0.49***	0.07	0.32***	-0.08	
	(0.06)	(0.04)	(0.06)	(0.19)	
PERFORMANCE	-0.06	0.24***	-0.07	-0.70***	
	(0.06)	(0.04)	(0.07)	(0.21)	
Age (standardized)	-0.03	-0.12*	0.03		
	(0.04)	(0.07)	(0.08)		
N	588	588	294	294	
R^2	.34	.083	.2	.09	
p-value: NAME vs. PERF.	1.0e-11	.0002	1.3e-07	.0037	
Mean in RANDOM	.23	.3	.4	2.4	
	Extra-version	Agree-ableness	Conscien-tiousness	Neuroticism	Openness to Experience
NAME	-0.14	0.09	-0.15	0.11	-0.15
	(0.14)	(0.09)	(0.11)	(0.13)	(0.10)
PERFORMANCE	0.01	0.14	-0.20	0.28**	0.12
	(0.17)	(0.09)	(0.12)	(0.13)	(0.11)
N	292	292	292	292	292
R^2	.05	.058	.047	.039	.03
p-value: NAME vs. PERF.	.19	.53	.63	.19	.031
Mean in RANDOM	1.2	1	1.1	.98	1.1
	Locus of Control	Social Comparison	Compe-titiveness	Risk	
NAME	0.12	0.00	0.03	0.07	
	(0.11)	(0.10)	(0.13)	(0.11)	
PERFORMANCE	0.46***	-0.19**	0.12	0.05	
	(0.12)	(0.09)	(0.11)	(0.11)	
N	292	293	291	292	
R^2	.065	.033	.03	.019	
p-value: NAME vs. PERF.	.003	.079	.37	.76	
Mean in RANDOM	.98	1.1	1.1	1.1	

Notes: This table presents least squares regressions using absolute differences in pairs' characteristics except for match quality and friendship as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. All regressions control for gender, grade and school fixed effects as well as age in regressions with individual outcomes.

Appendix 3.B Econometric Framework

In this appendix, we outline how to interpret our estimates in light of a mediation analysis similar to Heckman and Pinto (2015). A key difference between their framework and ours is that we are interested in the direct effect of our treatments as well as indirect effects of a change in the production inputs, rather than only the latter.

In general, any observed change in outcomes of our experiment can be attributed to one of two main sources: first, different peer-assignment mechanisms may affect peer interactions directly; and second, self-selection changes the peers and therefore the difference between the student's and his or her peer's characteristics. We therefore decompose the average treatment effect into a direct effect of self-selection as well as a pure peer composition effect. This takes into account the change in relative peer characteristics across treatments.[40]

Consider the following potential outcomes framework. Let Y^P and Y^N and Y^R denote the counterfactual outcomes in the three treatments. Naturally, we only observe the outcome in one of the treatments:

$$Y = D^N Y^N + D^P Y^P + (1 - D^P)(1 - D^N) Y^R \tag{3.B.1}$$

Let θ_d be a vector characterizing a peer's relative characteristics in treatment $d \in \{R, N, P\}$.[41] Similar to the potential outcomes above, we can only observe the peer composition vector θ in one of the treatments and thus $\theta = D_P \theta_P + D_N \theta_N + (1 - D_P)(1 - D_N)\theta_R$ and define an intercept α analogously. The outcome in each of the treatments is therefore given by

$$Y_d = \alpha_d + \beta_d \theta + \gamma X + \epsilon_d \tag{3.B.2}$$

where we implicitly assume that we have a linear production function, which can be interpreted as a first-order approximation of a more complex non-linear function. The outcome depends on own characteristics X as well as treatment-specific effects of relative characteristics of the peer θ and a zero-mean error term ϵ_d, independent of X and θ.

40. Our treatments do not change the distribution of characteristics or skills within the class or of a particular subject; rather, the treatments change with whom from the distribution a subject interacts. Due to the random assignment, we assume independence of own characteristics and the treatment.

41. In our estimations, we include the following characteristics in θ_d: indicators whether the peer ranked high in the individual preference rankings, effects of absolute time differences for slower and faster students within pairs, the rank and presence of friendship ties within pairs, and absolute differences in personal characteristics (Big 5, locus of control, competitiveness, social comparison and risk attitudes).

Potentially, there are unobserved factors in θ. We therefore split θ in a vector with the observed inputs ($\bar{\theta}$) and unobserved inputs ($\tilde{\theta}$)[42] with corresponding effects $\bar{\beta}_d$ and $\tilde{\beta}_d$ and can rewrite equation (3.B.2) as follows:

$$Y_d = \alpha_d + \bar{\beta}_d\bar{\theta} + \tilde{\beta}_d\tilde{\theta} + \gamma X + \epsilon_d \tag{3.B.3}$$

$$= \tau_d + \bar{\beta}_d\bar{\theta} + \gamma X + \tilde{\epsilon}_d \tag{3.B.4}$$

where $\tau_d = \alpha_d + \tilde{\beta}_d \mathrm{E}[\tilde{\theta}]$ and $\tilde{\epsilon}_d = \epsilon_d + \tilde{\beta}_d(\tilde{\theta} - \mathrm{E}[\tilde{\theta}])$. We assume $\tilde{\epsilon}_d \overset{d}{=} \epsilon$,i.e., are equal in their distribution with a zero-mean. We can express the effect of $\bar{\theta}$ in Name and Performance relative to the effect in Random by rewriting $\beta_d = \beta + \Delta_{R,d}$. Accordingly, we rewrite the coefficients $\bar{\beta}_d$ of θ_i as the sum of the coefficients in Random denoted by β and the distance of the coefficients between treatment d and Random (denoted by $\Delta_{R,d}$).

$$Y_d = \tau_d + \bar{\beta}\,\bar{\theta} + \bar{\Delta}_{R,d}\bar{\theta} + \gamma X + \tilde{\epsilon}_d \tag{3.B.5}$$

$$= \hat{\tau}_d + \bar{\beta}\,\bar{\theta} + \gamma X + \tilde{\epsilon}_d \tag{3.B.6}$$

In what follows, we are interested in $\bar{\tau}_d = \mathrm{E}[\hat{\tau}_d - \hat{\tau}_R]$ ($d \in \{N, P\}$; $\hat{\tau}_d = \tau_d + \bar{\Delta}_{R,d}\bar{\theta}$) , i.e., the direct treatment effect of Name and Performance conditional on indirect effects from changes in the peer composition captured in $\bar{\theta}$. This direct effect subsumes the effect of the treatment itself ($\alpha_d - \alpha_R$), the changed impact of the same peer's observables ($\bar{\Delta}_{R,d}\bar{\theta}$), and changes in unmeasured inputs as well as their effect ($(\tilde{\beta} + \tilde{\Delta}_{R,d})\tilde{\theta}$). We interpret this direct effect in light of self-determination theory (Deci and Ryan, 1985) as an additional motivation due to being able to self-select a peer. This focus on the direct effect is a key difference compared with Heckman and Pinto (2015), who are mainly interested in the indirect effects of the mediating variables. The empirical specification of 3.B.6 is given by

$$y_{igs} = \bar{\tau} + \bar{\tau}^N D_i^N + \bar{\tau}^P D_i^P + \beta\theta_i + \gamma X_i + \rho_s + \lambda_g + u_{igs} \tag{3.B.7}$$

where we are interested in $\bar{\tau}_N$ and $\bar{\tau}_P$, the direct effects of our treatments relative to Random. Indirect effects are captured by $\beta\theta_i$, the effect of changed peer characteristics on the outcome y_{igs}.

42. Furthermore, we assume that unobserved and observed inputs are independent conditional on X and D.

Appendix 3.C Robustness Checks for Average Treatment Effects

In Table 3.C.1, we compare the clustered standard errors with clustered standard errors using a biased-reduced linearization to account for the limited number of clusters. Comparing the first two columns, we observe that the results are robust to this alternative specification of the standard errors. In column (3), we additionally check whether looking at matching group-specific group means – i.e., the average percentage point improvement for males and females in each class – affects the estimates. While the power is reduced due to the small number of observations, the treatment effects persist and the coefficients on the treatment effects are not significantly affected. Columns (4) and (5) analyze the sensitivity of our estimates with respect to outliers. We use two different strategies. First, we apply a 90% winsorization, which replaces all observations with either a time or a percentage point improvement below or above the threshold with the value at the threshold. We replace a time of improvement below the 5th percentile with the corresponding value of the 5th percentile and all observations above the 95th percentile with the 95th percentile. Second, we truncate the data and keep only those pairs where no time or no improvement falls into the bottom 5% or top 5%. Neither winsorization nor truncation significantly changes the estimated treatment effects.

Table 3.C.1. Robustness Checks

	Percentage Point Improvements				
	(1) Baseline	(2) BRL	(3) Group means	(4) Winsorization	(5) Truncation
NAME	1.26***	1.26**	1.13*	1.05***	0.95***
	(0.43)	(0.50)	(0.61)	(0.37)	(0.35)
PERFORMANCE	1.67**	1.67**	1.96***	1.51***	1.43***
	(0.62)	(0.72)	(0.62)	(0.51)	(0.43)
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes	Yes
N	588	588	70	588	496
R^2	.056	.056	.27	.072	.087
p-value: NAME vs. PERF.	.51	.55	.15	.37	.27

Notes: This table presents least squares regressions using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Column (1) presents the baseline specifications as used in Table 3.3. Columns (2) uses biased-reduced linearization (BRL) to account for the limited number of clusters. Column (3) uses matching group-specific means as the unit of observation. Finally, columns (4) and (5) apply a 90% winsorization and truncation, respectively.

We further analyze the robustness of our results by looking at different subsamples. We therefore split our sample first by grades in the upper panel of Table 3.C.2

Table 3.C.2. Robustness Checks – Subsample Analyses

	Percentage Point Improvements				
	(1) Baseline	(2) 7th grade	(3) 8th grade	(4) 9th grade	(5) 10th grade
NAME	1.26***	1.95***	2.60***	1.53**	1.08*
	(0.43)	(0.08)	(0.35)	(0.59)	(0.61)
PERFORMANCE	1.67**	2.78***	2.51***	2.53***	1.32
	(0.62)	(0.63)	(0.15)	(0.62)	(0.88)
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes	Yes
N	588	116	116	174	182
R^2	.056	.073	.064	.16	.039
p-value: NAME vs. PERF.	.51	.21	.82	.19	.82
	(6) Female	(7) Male	(8) School 1	(9) School 2	(10) School 3
NAME	1.26*	1.21***	1.36***	1.44**	2.09***
	(0.65)	(0.44)	(0.11)	(0.65)	(0.37)
PERFORMANCE	1.68**	1.63*	1.53***	2.29***	2.22*
	(0.77)	(0.85)	(0.05)	(0.55)	(1.12)
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes	Yes
N	390	198	148	274	166
R^2	.057	.065	.065	.1	.12
p-value: NAME vs. PERF.	.53	.62	.3	.14	.88

Notes: This table presents least squares regressions using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Column (1) presents the estimates using the whole sample as in Table 3.3. Columns (2)-(5) restrict the sample to one grade, columns (6) and (7) to each gender and columns (8)-(10) to one school.

and by schools as well as gender in the lower panel and estimate the treatment effects separately for those samples. The table shows the robustness of the estimated treatment effects as these effects persists for all subsamples with similar magnitude.

Appendix 3.D Control Treatment to Disentangle Peer Effects from Learning

Table 3.D.1 and Figure 3.D.1 present the estimated average treatment effects and the margins including an additional control treatment. The NoPeer treatment featured the same design as all other treatments. The only difference was that students participated in the running task twice without a peer. Moreover, we shortened the survey for this treatment by removing the questionnaires on personal characteristics. The control treatment was conducted to show that the observed performance improvements are not due to learning. If learning drives our effects, we should observe performance improvements in NoPeer, which is not the case. Even if this control treatment had yielded performance improvements, this would not affect any of our results. To see this, note that we are interested in a between treatment comparison of performance improvements. Learning effects between the runs should therefore be constant across treatments.

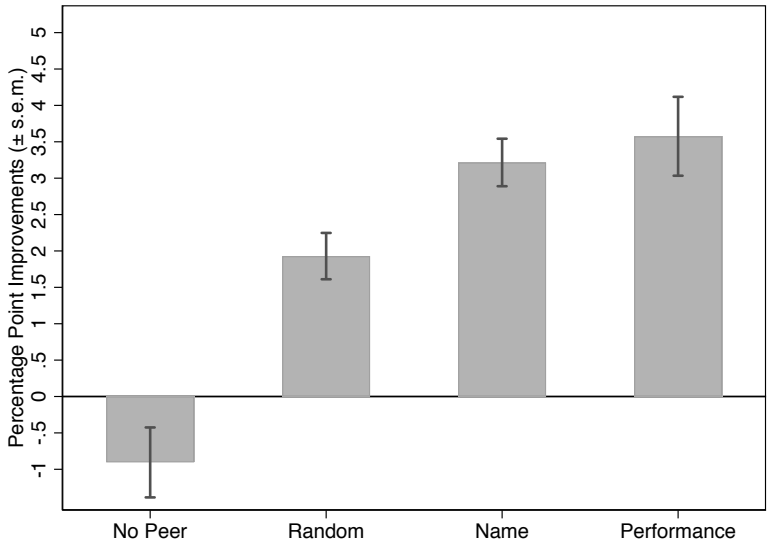


Figure 3.D.1. Average Treatment Effects

Notes: The figure presents percentage point improvements from the first to the second run with corresponding standard errors for the three treatments Random, Name, and Performance and an additional control treatment, where students run two times without a peer (NoPeer). See column (1) in Table 3.D.1 for the corresponding regression. We control for gender, grade and school fixed effects as well as age and cluster standard errors at the class level.

Table 3.D.1. Robustness Checks

	(a) PP. Imprv.	(b) Time (Second Run)	
	(1)	(2)	(3)
Name	1.29***	-0.37***	-0.14***
	(0.42)	(0.11)	(0.04)
Performance	1.65**	-0.40***	-0.15***
	(0.62)	(0.14)	(0.05)
NoPeer	-2.84***	0.82***	0.31***
	(0.61)	(0.16)	(0.06)
Controlling for Time (First Run)	No	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes
N	715	715	715
R^2	.14	.81	.81

Notes: This table presents least squares regressions using percentage point improvements (Panel (a)) or times from the second run (Panel (b)) as the dependent variables. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level.

Appendix 3.E Peer Composition Robustness Checks

We run several robustness checks for the results presented in Table 3.6. First, in Table 3.E.1 we use different specifications for match quality. We consider the partner's match quality, an interaction between one's own and the partner's match quality, and feasible match quality as defined in Appendix 3.A, and find that the estimates of our direct effects are qualitatively and quantitatively the same. Second, in Table 3.E.2, we show that our results do not depend on the precise definition of friendship ties. We check whether our results change when we define friendship ties as undirected or reciprocal rather than directed. As can be seen from the table, the coefficients on the direct effects as well as on other peer characteristics remain the same. Third, we control for differences in productivity in a more flexible way in Table 3.E.3 by allowing for quartic rather than linear effects of productivity differences in column (2) (see also Figure 3.E.1 comparing linear and quartic terms graphically). In addition, we allow for a second flexible specification using fixed effects for productivity differences. More specifically, we include an indicator for each one-second interval of productivity differences between subjects within a pair. This allows for a potential non-linear influence of productivity differences on our estimates. Comparing the estimates shows that neither the quartic functional form nor the fixed effect specification is restrictive. Fourth, we estimate the influence of peer characteristics (and individual characteristics) on the sample of RANDOM subjects only in Table 3.E.4 and use these coefficients to decompose the average effect. For this purpose, we first net out the effect of group variables such as school and gender-grade fixed effects (as well as individual characteristics) from both the outcome and independent variables such as peer characteristics according to the Frisch-Waugh-Lovell theorem using the whole sample. In a first version, we regress the outcome and peer characteristics on the fixed effects only. In a second version, we additionally net out the effect of individual characteristics from peer characteristics and the outcome. We use the residuals of those regressions to decompose the treatment effect. We then begin by estimating the influence of peer characteristics on the outcome using only subjects from RANDOM and the residualized outcome as well as peer characteristics (column (1) and (3)). In a second step, we restrict the influence of those peer characteristics and estimate the direct treatment effects (column (2) and (4)). Finally, Table 3.E.5 restricts the control group sample to subjects with a high match quality within RANDOM to show that the treatment effects persist for these subjects and the coefficients on peer compositional effects do not substantially change. Table 3.E.6 presents the ommited coefficients of own and peer characteristics, as well as their absolute differences, from column (5) Table 3.6 in the main text.

Table 3.E.1. Robustness Checks for Match Quality

	Percentage Point Improvements		
	(1) Partner's MQ	(2) Interaction	(3) Feasible
Direct Effects			
Name	1.15**	1.14*	1.19**
	(0.55)	(0.57)	(0.47)
Performance	2.23***	2.21***	2.05***
	(0.70)	(0.69)	(0.66)
Peer Characteristics			
High match quality (partner; Name)	0.28	0.18	
	(0.42)	(0.56)	
High match quality (partner; Perf.)	-0.07	0.21	
	(0.40)	(0.44)	
High match quality (own and partner; Name)		0.19	
		(0.84)	
High match quality (own and partner; Perf.)		-0.58	
		(0.94)	
Faster Student × High match quality (feasible; Name)			0.02
			(0.42)
Slower Student × High match quality (feasible; Name)			1.38*
			(0.79)
Faster Student × High match quality (feasible; Perf.)			0.83*
			(0.41)
Slower Student × High match quality (feasible; Perf.)			0.32
			(0.86)
Faster Student × Match Quality (name-based)	0.45	0.37	
	(0.40)	(0.44)	
Slower Student × Match Quality (name-based)	0.41	0.30	
	(0.65)	(0.75)	
Faster Student × Match Quality (perf.-based)	0.43	0.75	
	(0.51)	(0.65)	
Slower Student × Match Quality (perf.-based)	-0.71	-0.48	
	(0.65)	(0.61)	
Friendship Ties and Performance Differences	Yes	Yes	Yes
Abs. Diff. in Personality	Yes	Yes	Yes
Peer Characteristics	Yes	Yes	Yes
Own Characteristics	Yes	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes
N	582	582	582
R^2	.29	.29	.29
p-value: Name vs. Performance	.16	.18	.24

Notes: This table presents least squares regressions using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Own and peer characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. Column (1) adds the partner's match quality in addition to own match quality as in Table 3.6, while column (2) additionally controls for the interaction of own and partner's match quality. Finally, column (3) uses a different measure of match quality, (feasible match quality – see also Appendix 3.A), which acknowledges the fact that certain preferred peers may not be available.

Table 3.E.2. Different Definitions of Friendship Ties

	Percentage Point Improvements			
	(1) directed	(2) undirected	(3) reciprocal	(4) dir. & rec.
Direct Effects				
NAME	1.24**	1.20**	1.21**	1.14**
	(0.50)	(0.49)	(0.50)	(0.50)
PERFORMANCE	2.21***	2.13***	2.21***	2.19***
	(0.68)	(0.69)	(0.68)	(0.68)
Faster Student × Peer is friend	-1.15**			-1.67*
	(0.53)			(0.85)
Slower Student × Peer is friend	0.13			-0.38
	(0.67)			(0.83)
Faster Student × Peer is friend (undirected)		-1.63***		
		(0.58)		
Slower Student × Peer is friend (undirected)		0.16		
		(0.80)		
Faster Student × Peer is friend (reciprocal)			-0.56	0.76
			(0.59)	(0.94)
Slower Student × Peer is friend (reciprocal)			0.47	0.73
			(0.53)	(0.63)
Faster Student × $\lvert\Delta Time\ 1\rvert$	-0.35**	-0.34**	-0.34**	-0.34**
	(0.16)	(0.16)	(0.16)	(0.15)
Slower Student × $\lvert\Delta Time\ 1\rvert$	1.04***	1.04***	1.05***	1.05***
	(0.20)	(0.20)	(0.20)	(0.20)
Slower Student in Pair	-0.15	-0.47	0.05	-0.18
	(0.68)	(0.74)	(0.69)	(0.68)
Match quality and performance differences	Yes	Yes	Yes	Yes
Abs. Diff. in Personality	Yes	Yes	Yes	Yes
Peer Characteristics	Yes	Yes	Yes	Yes
Own Characteristics	Yes	Yes	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes
N	582	582	582	582
R^2	.29	.29	.29	.29

Notes: This table presents least squares regressions using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Own and peer characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. Column (1) presents the last specification of Table 3.6 for reference using directed friendship ties. Column (2) uses undirected friendship ties, column (3) reciprocal directed friendship ties, while column (4) allows for a differential effect of directed and reciprocal friendship ties.

Table 3.E.3. Robustness Checks for Absolute Time Differences

	Percentage Point Improvements		
	(1) Linear	(2) Quartic	(3) FEs
Direct Effects			
NAME	1.24**	1.28**	1.20**
	(0.50)	(0.49)	(0.52)
PERFORMANCE	2.21***	2.23***	2.25***
	(0.68)	(0.68)	(0.74)
Faster Student × $\lvert\Delta Time\ 1\rvert$	-0.35**	-2.70**	
	(0.16)	(1.26)	
Slower Student × $\lvert\Delta Time\ 1\rvert$	1.04***	1.27	
	(0.20)	(1.75)	
Slower Student in Pair	-0.15	-1.82*	
	(0.68)	(0.91)	
Faster Student × $\lvert\Delta Time\ 1\rvert^2$		0.90	
		(0.56)	
Slower Student × $\lvert\Delta Time\ 1\rvert^2$		-0.00	
		(0.97)	
Faster Student × $\lvert\Delta Time\ 1\rvert^3$		-0.12	
		(0.09)	
Slower Student × $\lvert\Delta Time\ 1\rvert^3$		-0.01	
		(0.18)	
Faster Student × $\lvert\Delta Time\ 1\rvert^4$		0.00	
		(0.00)	
Slower Student × $\lvert\Delta Time\ 1\rvert^4$		0.00	
		(0.01)	
Time Diff. FEs	No	No	Yes
Match Quality and Friendship Ties	Yes	Yes	Yes
Abs. Diff. in Personality	Yes	Yes	Yes
Peer Characteristics	Yes	Yes	Yes
Own Characteristics	Yes	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes
N	582	582	582
R^2	.29	.29	.3
p-value: NAME vs. PERFORMANCE	.17	.17	.14

Notes: This table presents least squares regressions using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Own and peer characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. Column (1) presents the last specification of Table 3.6 for reference. Column (2) includes quartic terms of time differences in the first run (also illustrated in Appendix Figure 3.E.1) and column (3) fixed effects for every one-second difference in productivity levels of the two students.

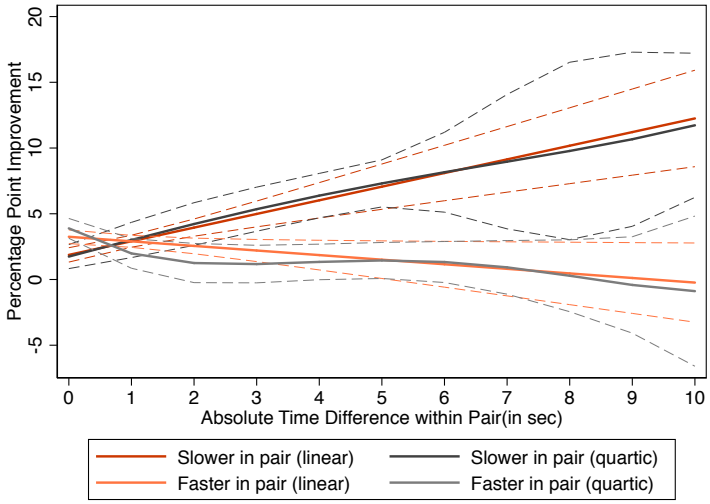


Figure 3.E.1. Robustness of Linear Specification in Time Differences

Notes: The figure presents marginal effects (solid lines) from a least squares regression using percentage point improvements as the dependent variable including 95% confidence intervals (dashed lines). It plots the linear specification (black lines) as used in the main text as well as a second specification using quartic polynomials (orange lines) of absolute time differences in the first run as regressors. We use the same set of controls as in column (5) of Table 3.6 and cluster standard errors at the class level. The corresponding regressions are presented in columns (1) and (2) of Appendix Table 3.E.3.

Table 3.E.4. Restricting Coefficients of Peer Characteristics

	Percentage Point Improvements			
	Fixing only FEs		Fixing FEs & own char.	
	(1) only Random	(2) all	(3) only Random	(4) all
Direct Effects				
Name		.77*		.79*
		(.46)		(.47)
Performance		1.67**		1.66**
		(.67)		(.67)
Peer Characteristics				
Faster Student × High match quality (Name)	.76	.76	.76	.76
	(.85)		(.78)	
Slower Student × High match quality (Name)	.26	.26	.38	.38
	(1.09)		(1.01)	
Faster Student × High match quality (Perf.)	.18	.18	-.15	-.15
	(1.11)		(1.13)	
Slower Student × High match quality (Perf.)	-.41	-.41	-.14	-.14
	(1.15)		(1.2)	
Faster Student × Peer is friend	-.14	-.14	-.19	-.19
	(.66)		(.59)	
Slower Student × Peer is friend	.03	.03	-.06	-.06
	(1.28)		(1.15)	
Faster Student × $\lvert\Delta Time\ 1\rvert$	-.51*	-.51	-.5	-.5
	(.3)		(.28)	
Slower Student × $\lvert\Delta Time\ 1\rvert$	.78**	.78	.84**	.84
	(.32)		(.3)	
Slower Student in Pair	.13	.13	-.1	-.1
	(.99)		(.87)	
Abs. Diff. in Personality	Yes	Yes	Yes	Yes
Peer Characteristics	Yes	Yes	Yes	Yes
Own Characteristics	No	No	Yes	Yes
N	204	582	204	582
R^2	.24		.22	

Notes: This table presents least squares regressions using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Own and peer characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. We use residualized dependent and independent variables, where we take out the variation of individual-specific variables. The first two columns take out the variation of the set of fixed effects, while the last two columns additionally take out variation of own characteristics. Columns (1) and (3) present least squares regressions in Random only, while columns (2) and (4) use all three treatments, but restrict the coefficients to equal the preceding columns.

Table 3.E.5. Only High Match Quality Sample As Comparison Group

	Percentage Point Improvements					
	(1) Random	(2) All	(3) Random & Name	(4) with Controls	(5) Random & Perf.	(6) with Controls
Direct Effects						
Name		1.24**	1.83***	1.93***		
		(0.50)	(0.55)	(0.47)		
Performance		2.21***			2.38***	1.75**
		(0.68)			(0.71)	(0.64)
Peer Characteristics						
Faster Student × Match Quality (name-based)	0.89	0.52				-0.47
	(0.95)	(0.43)				(1.28)
Slower Student × Match Quality (name-based)	0.15	0.46				-0.56
	(1.10)	(0.66)				(1.15)
Faster Student × Match Quality (perf.-based)	0.06	0.43		-0.51		
	(1.08)	(0.53)		(0.65)		
Slower Student × Match Quality (perf.-based)	-0.51	-0.71		-1.21		
	(1.22)	(0.66)		(0.86)		
Faster Student × Peer is friend	0.10	-1.15**		-1.53		-0.98
	(0.74)	(0.53)		(1.05)		(1.87)
Slower Student × Peer is friend	0.01	0.13		-1.18		-1.38
	(1.15)	(0.67)		(1.06)		(1.13)
Faster Student × $\lvert\Delta Time\ 1\rvert$	-0.54**	-0.35**		-0.72**		-0.07
	(0.25)	(0.16)		(0.29)		(0.51)
Slower Student × $\lvert\Delta Time\ 1\rvert$	0.73**	1.04***		1.25***		1.08**
	(0.32)	(0.20)		(0.38)		(0.47)
Slower Student in Pair	0.43	-0.15		-0.44		-0.97
	(1.15)	(0.68)		(1.70)		(1.47)
Abs. Diff. in Personality	Yes	Yes	No	Yes	No	Yes
Peer Characteristics	Yes	Yes	No	Yes	No	Yes
Own Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes	Yes	Yes
N	204	582	208	207	162	160
R^2	.28	.29	.16	.52	.16	.37

Notes: This table presents least squares regressions using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Own and peer characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. Column (1) and (2) present the last specification of Table 3.6 for Random and the full sample for reference. Columns (3) to (6) show that even if we restrict the comparison group to the sample of individuals in Random that received a peer with high match quality according to their name- (columns (3) and (4)) or performance-based preferences (columns (5) and (6)), respectively, the direct effects persist and the coefficients on peer compositional effects do not change much.

Table 3.E.6. Omitted Coefficients from Table 3.6 Column (5)

	Own characteristics	Peer characteristics	Abs. Diff in characteristics
Agreeableness	0.12	-0.11	0.29
	(0.22)	(0.20)	(0.29)
Conscientiousness	0.01	0.13	-0.13
	(0.21)	(0.17)	(0.23)
Extraversion	0.03	0.06	-0.51**
	(0.24)	(0.20)	(0.25)
Openness to Experience	-0.49**	-0.18	0.52
	(0.19)	(0.17)	(0.33)
Neuroticism	-0.16	-0.16	-0.65**
	(0.24)	(0.19)	(0.27)
Locus of Control	0.17	0.09	-0.15
	(0.20)	(0.19)	(0.31)
Social Comparison	0.32*	0.21	-0.21
	(0.18)	(0.16)	(0.31)
Competitiveness	-0.08	-0.37	0.35
	(0.30)	(0.23)	(0.21)
Risk Attitudes	0.04	0.06	1.32
	(0.18)	(0.17)	(1.70)

Notes: This table presents omitted coefficients from Table 3.6 in the main text. Columns (1) and (2) show the coefficients on own and peer characteristics, respectively. Column (3) presents the coefficients on the absolute differences in personality measures. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level.

Appendix 3.F Additional Material for Discussion of Direct Effects

Table 3.F.1 presents three regressions to support section 3.5.5's discussion of the psychological effect underlying the direct effects. First, we show that students in RANDOM are not disappointed by having a partner assigned. If they were disappointed, they should have less fun during the second run. As column (1) show this is not the case. Secobd, we do not find evidence that subjects with self-selected perceive winning in the second run as more important as we do not see a differential effect on fun between being faster or slower in the second run. Third, we show that prosocial students, that is individuals that score higher on agreeableness, do not show differentially direct effects. This is suggestive evidence against experimenter demand effects or other reciprocal motives driving the estimated direct effects.

Table 3.F.1. Potential Psychological Mechanisms for the Direct Effect

	Fun (std.).		PP. Imprv.
	(1)	(2)	(3)
Direct Effects			
Name	-0.01	0.01	1.24**
	(0.10)	(0.14)	(0.50)
Performance	-0.10	-0.07	2.20***
	(0.08)	(0.13)	(0.68)
Name × Slower Student in Pair (2nd Run)		-0.05	
		(0.18)	
Performance × Slower Student in Pair (2nd Run)		-0.07	
		(0.17)	
Name × Agreeableness			0.02
			(0.38)
Performance × Agreeableness			0.42
			(0.45)
Peer Characteristics			
Faster Student (2nd Run) × $\lvert\Delta Time\ 2\rvert$	-0.01	-0.01	
	(0.04)	(0.04)	
Slower Student (2nd Run) × $\lvert\Delta Time\ 2\rvert$	-0.14***	-0.14***	
	(0.04)	(0.04)	
Slower Student in Pair (2nd Run)	0.04	0.07	
	(0.18)	(0.20)	
Faster Student × $\lvert\Delta Time\ 1\rvert$			-0.35**
			(0.16)
Slower Student × $\lvert\Delta Time\ 1\rvert$			1.05***
			(0.20)
Slower Student in Pair			-0.20
			(0.66)
Match quality	Yes	Yes	Yes
Friendship indicators	Yes	Yes	Yes
Own Characteristics	Yes	Yes	Yes
Peer Characteristics	Yes	Yes	Yes
Abs. Diff. in Personality Characteristics	Yes	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes
N	582	582	582
R^2	.34	.34	.29
p-value: Name vs. Performance	.46	.63	.18

Notes: This table presents least squares regressions using a standardized measure of fun in the second run (columns (1) and (2)) or percentage point improvements (column (3)) as the dependent variable. Column (2) uses the full specification of Table 3.6 and additionally interacts the treatment indicators with one's own measure of agreeableness as a proxy of prosociality. Column (1) focuses on fun as an outcome variable that was elicited after the second run ("How much fun did you have during the second run? Please rate this on a scale from 1 – no fun at all – to 5 – a lot of fun.") and uses the full specification of Table 3.6 adapted using times and ranks from the second run. Column (2) additionally interacts treatment indicators with the final rank in the second run. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level.

Appendix 3.G Additional Material for Implications

Our treatments also have implications for individual ranks of students within a class since slower students improve more than faster ones. As ranks are important in determining subsequent outcomes (Elsner and Isphording, 2017; Murphy and Weinhardt, 2018; Gill et al., 2019), a policy maker has to take the distributional effects of peer assignment mechanisms into account.[43] Since low-ability students improve relatively more than high-ability students in NAME and RANDOM, these treatments yield potentially large changes of a student's rank within the class between the two runs. By contrast, PERFORMANCE will tend to preserve the ranking of the first run as improvements are distributed more equally relative to the two other treatments. We confirm this intuition in Table 3.G.1 in which we regress the absolute change in percentile scores from the first to the second run on treatment indicators. The outcome variable measures the average perturbation of ranks within in a class across the two runs. The results show that PERFORMANCE shuffles the ranks of students less in comparison to RANDOM and NAME. While in RANDOM students change their position by about 15 out of 100 ranks, we find significantly less changes in the percentile score in PERFORMANCE relative to RANDOM. This change corresponds to a 27% reduction in reshuffling. However, in NAME we do not find any effect compared to RANDOM.

As another side effect we consider the pressure students experienced during the second run due to their peer. We find that students in PERFORMANCE experience significantly more pressure than students in the other two treatments.

43. Suppose that a policy maker wants to establish a rank distribution (ranks based on times in the second run) that mirrors the ability distribution (ranks based on times in the first run) due to some underlying fairness ideal (e.g., she wants to shift the distribution holding constant individual ranks). In other words, she might want to implement a peer assignment mechanism that preserves individual ranks rather than shuffle them.

Table 3.G.1. Side Effects of Reassignment Rules

	Absolute Change in Percentile Scores		Pressure (std.)
	(1) within matching group	(2) within treatment	(3)
NAME	-0.01	-0.02	0.10
	(0.01)	(0.01)	(0.18)
PERFORMANCE	-0.04**	-0.04***	0.46**
	(0.02)	(0.01)	(0.15)
Gender/Grade/School FEs, Age	Yes	Yes	Yes
Other controls	No	No	Yes
N	588	588	161
R^2	.056	.051	.32
p-value: NAME vs. PERFORMANCE	.018	.085	.17
Mean in RANDOM	.15	.14	-.16

Notes: This table presents least squares regressions using absolute change in percentile scores or a standardized measure of pressure during the second run as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Absolute changes in percentile scores within matching groups are calculated based on the change of individual ranks of students in the their class and gender from the first to the second. Percentile scores within treatment are calculated for all students within the same treatment and gender (i.e., across classrooms). Other controls include the same controls as the mediation model in Table 3.6, where we use times and ranks from the second rather than the first run as the pressure variable has been elicited after the second run. Note that information on pressure was only elicited at one of the three schools.

Appendix 3.H Simulation of Matching Rules

We simulate three matching rules and predict their impact on performance improvements using our estimates from Table 3.6. In a first step, we create artificial pairs, based on the employed matching rules described below. In a second step, we then calculate the vector θ of differences for the artificial pairs as well as the matching quality of artificial peers. Finally, we use the estimated coefficients from the column (5) of Table 3.6 to predict the performance improvements we would observe for the artificial pairs. As peer-assignment rules only change θ, we are interested in the difference in the respective sums of the indirect effect and direct effect, that is between $\bar{\tau} + \beta\,\theta_i^{sim}$ and $\bar{\tau} + \beta\,\theta_i^{obs}$ from equation (3.2), where *sim* and *obs* denote simulated and observed pair characteristics, respectively. As we consider exogenous assignment rules, we assume that the direct effect of the simulated policies equals zero as in in RANDOM. We additionally fix the covariates X to 0 and leave out the fixed effects for the simulations and predictions. This means, we calculate the performance improvements for a particular baseline group for our treatments as well as the simulations. This enables us to compare our results of the simulations directly to the peer-assignment rules using self-selection implemented in the experiment, as we compare the performance improvements for the same group.

In addition to our three treatments, we simulate four types of peer assignment rules. First, we simulate two settings in which we assign the self-selected peers exogenously (NAME (EXOG.) and PERFORMANCE (EXOG.)). Hence, the resulting pairs are the same as in the self-selection treatment, but we exclude the direct effect of self-selection. Second, we implement an ability tracking assignment rule, TRACKING, in the spirit of the matching also employed in Gneezy and Rustichini (2004). Students are matched in pairs, starting with the two fastest students in a matching group and moving down the ranking subsequently. This rule minimizes the absolute distance in pairs. Third, we employ a peer assignment rule that fixes the distance in ranks for all pairs (EQUIDISTANCE). We rank all students in a matching group and match the first student with the one in the middle and so forth. More specifically, if G denotes the group size, the distance in ranks is $G/2 - 1$ for all pairs. This rule is one way to maximize the sum of absolute differences in pairs, but keeps the distance across pairs similarly. Fourth, we match the highest ranked student with the lowest one, the second highest ranked with the second lowest one and so forth (HIGH-TO-LOW). This is similar to Carrell, Sacerdote, and West (2013), who match low-ability students with those students from whom they would benefit the most (i.e., the fastest students). Again, this assignment rule maximizes the sum of absolute differences in pairs. Table 3.H.1 summarizes initial performance differences within pairs of the experimental treatments as well as the simulated assignment rules and the predicted performance improvements.

Table 3.H.1. Overview of Simulated Peer Assignment Rules

Peer assignment rule	Mean absolut productivity differences (in sec)	Predicted improvement (in pp.)	Description
Name	2.09	2.43	Self-selected peers based on names
Performance	1.41	2.69	Self-selected peers based on relative performance
Name (exog.)	2.09	1.19	Self-selected peers based on names without self-selection effect
Performance (exog.)	1.41	0.48	Self-selected peers based on relative performance without self-selection effect
Random	2.42	1.12	Randomly assigned peers
Equidistance	3.11	1.44	Same distance in ranks across pairs
High-to-Low	3.11	1.36	First to last, second to second to last etc.
Tracking	0.90	0.72	First to second, third to fourth etc.

Appendix 3.I Experimental Instructions and Protocol

The instructions below are translations of the German instructions for the experiment.

Introduction to the Experiment

Welcome everyone to today's physical education session. As you might have already noticed, today's session is going to be different. As you already know, you will take part in a scientific study. For that purpose, you received a parental consent form and handed it back to your teacher. If you have not handed it back to your teacher, you will not take part in the study.
The study is going to be conducted by the three of us: Lukas Kiessling, Sebastian Schaube and I am Jonas Radbruch. If you have any questions throughout the study, you can address us at any point in time.
The study comprises several parts. For the first part, we would like you to do a running task called suicide runs. My colleague will shortly demonstrate this exercise.
(The following verbal explanation was accompanied with physical demonstration of the exercise)
You start at the baseline of the volleyball court and run to to this first line. You touch it with your hand and run back to the baseline. You touch the baseline with your hand and run to the next line. Touch it again, back to the baseline; touch it, and then to the third line, back to the baseline, to the fourth line and then you return to the baseline.
Everyone of you will run alone and the goal is to be as fast as possible. After this run, we will hand you a computer to fill out a survey.
After all of you have ran and filled out the survey, you will run for a second time. This time at the same time as another student. During the survey we will ask you – among other questions – with whom you would like to run. You will receive detailed information about this later on.
The goal during both runs is to be as fast as possible. We will record your running times and hand it to your teacher. Your teacher will grade your performance during both runs.
Before we start with the study, we would like to remind you again that your participation is voluntary. If anyone does not want to take part in the study, then please inform us now.
Do you have any further questions? If this is not the case, please start with the warm-up, before we start with the experiment.
(Introduction was followed by short warm-up by students. After a short warm-up all students were asked to leave the gym and wait in an accompanying the hallway until they were called in the gym to take part in the first run. We asked students whether they understood the task and, if necessary, explained the task again. Directly afterwards,

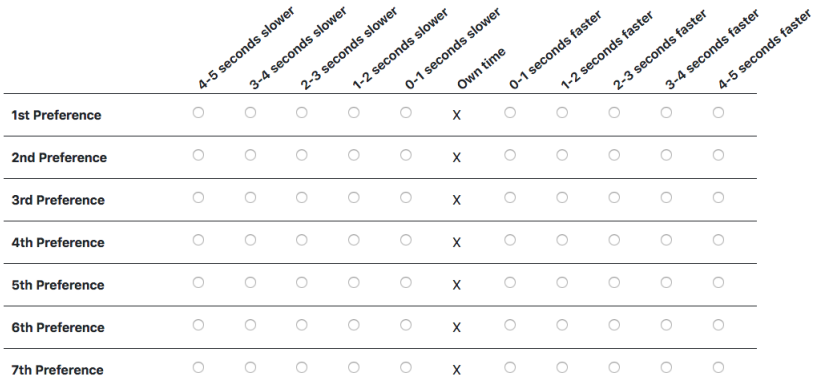


Figure 3.I.1. Performance-based Preferences

Figure 3.I.2. Name-based Preferences

they were asked to leave the gym and were led to a different room. There we asked them to complete the survey on a computer we handed them.)

Screenshots of the Preferences-elicitation During the Survey

(The following two screenshots, Figures 3.I.1 and 3.I.2, display translated elicitation screens for performance- and name-based preferences for peers.)

Introduction to the Second Run for the Whole Class

(Class was gathered for announcement)
We will shortly start with the second run. For this purpose a partner for you has been selected. In your class, the partner has been selected *randomly [based on your indication how fast you want your partner to be] [based on the classmates you nominated]*. We would like to remind you that the objective is to be as fast as possible and it is

only about your own time. Your teacher will receive a list with your performance, but no information about the pairs.
(The list with pairs was read out aloud to the students and students were accompanied to the waiting zone. Students were called into the gym one pair after the other. In the gym they were led to separate, but adjacent tracks. Each student was accompanied by one experimenter, who recorded their time as well their responses to four additional questions.)

Individual Introduction Directly Before the Second Run

The two of you will now run simultaneously. Your partner has been selected randomly [based on your indication how fast you want your partner to be] [based on the classmates you nominated]. .
(We then asked each subject to assess their relative performance in the first run) Please guess, who of you two was faster during the first run?

Post-run Questionnaire After the Second Run

(Directly after a pair participated in the second run, we asked each of the two subjects the following three questions in private)

(1) How much fun did you have during the second run? Please rate this on a scale from 1 – no fun at all – to 5 – a lot of fun

(2) If you were to run again, would you prefer to run alone or with a partner)

(3) How much pressure did you feel form your partner during the second run? Please rate this on a scale from 1 – no pressure at all – to 5 – a lot of pressure.

References

Ager, Philipp, Leonardo Bursztyn, and Hans-Joachim Voth. 2016. "Killer Incentives: Status Competition and Pilot Performance during World War II." NBER Working Paper Series. DOI: 10.3386/w22992. [110]

Agostinelli, Francesco. 2018. "Investing in Children's Skills: An Equilibrium Analysis of Social Interactions and Parental Investments." Working Paper. [113]

Aldashev, Gani, Georg Kirchsteiger, and Alexander Sebald. 2017. "Assignment Procedure Biases in Randomised Policy Experiments." *Economic Journal* 127 (602): 873–95. DOI: 10.1111/ecoj.12321. [135]

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005. "Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools." *Journal of Political Economy* 113 (1): 151–84. [133]

Aral, Sinan, and Christos Nicolaides. 2017. "Exercise Contagion in a Global Social Network." *Nature Communications* 8 (14753): [112]

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2009. "Social Connections and Incentives in the Workplace: Evidence From Personnel Data." *Econometrica* 77 (4): 1047–94. DOI: 10.3982/ECTA6496. [109]

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2010. "Social Incentives in the Workplace." *Review of Economic Studies* 77 (2): 417–58. DOI: 10.1111/j.1467-937X.2009.00574.x. eprint: /oup/backfile/content_public/journal/restud/77/2/10.1111_j.1467-937x.2009.00574.x/2/77-2-417.pdf. [109, 113]

Bartling, Björn, Ernst Fehr, and Holger Herz. 2014. "The intrinsic value of decision rights." *Econometrica* 82 (6): 2005–39. [112, 136]

Bartling, Björn, Ernst Fehr, and Klaus M. Schmidt. 2013. "Discretion, productivity, and work satisfaction." *Journal of Institutional and Theoretical Economics* 169 (1): 4–22. [113]

Becker, Anke, Thomas Deckers, Thomas Dohmen, Armin Falk, and Fabian Kosse. 2012. "The Relationship Between Economic Preferences and Psychological Personality Measures." *Annual Review of Economics* 4 (1): 453–78. DOI: 10.1146/annurev-economics-080511-110922. eprint: http://dx.doi.org/10.1146/annurev-economics-080511-110922. [135]

Belot, Michèle, and Jeroen van de Ven. 2011. "Friendships and Favouritism on the Schoolground – A Framed Field Experiment." *Economic Journal* 121 (557): 1228–51. DOI: 10.1111/j.1468-0297.2011.02461.x. [113]

Betts, Julian R. 2011. "The Economics of Tracking in Education." In. Edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Vol. 3, Handbook of the Economics of Education. Elsevier, 341–81. DOI: https://doi.org/10.1016/B978-0-444-53429-3.00007-7. [136]

Bó, Pedro Dal, Andrew Foster, and Louis Putterman. 2010. "Institutions and Behavior: Experimental Evidence on the Effects of Democracy." *American Economic Review* 100 (5): 2205–29. [113]

Booij, Adam S., Edwin Leuven, and Hessel Oosterbeek. 2017. "Ability Peer Effects in University: Evidence from a Randomized Experiment." *Review of Economic Studies* 84 (2): 547–78. DOI: 10.1093/restud/rdw045. eprint: /oup/backfile/content_public/journal/restud/84/2/10.1093_restud_rdw045/2/rdw045.pdf. [112]

Bradler, Christiane, Robert Dur, Susanne Neckermann, and Arjan Non. 2016. "Employee Recognition and Performance: A Field Experiment." *Management Science* 62 (11): 3085–99. [113]

Brandts, Jordi, David Cooper, and Roberto Weber. 2014. "Legitimacy, Communication, and Leadership in the Turnaround Game." *Management Science* 61 (11): 2627–45. [113]

Bursztyn, Leonardo, Florian Ederer, Bruno Ferman, and Noam Yuchtman. 2014. "Understanding Mechanisms Underlying Peer Effects: Evidence From a Field Experiment on Financial Decisions." *Econometrica* 82 (4): 1273–301. DOI: 10.3982/ECTA11991. [113]

Carrell, Scott, Bruce Sacerdote, and James West. 2013. "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation." *Econometrica* 81 (3): 855–82. URL: http://www.jstor.org/stable/23524165. [112, 136, 163]

Cassar, Lea, and Stephan Meier. 2018. "Nonmonetary Incentives and the Implications of Work as a Source of Meaning." *Journal of Economic Perspectives* 32 (3): 215–38. DOI: 10.1257/jep.32.3.215. [136, 139]

Chan, Tszkin Julian, and Chungsang Tom Lam. 2015. "Type of Peers Matters: A Study of Peer Effects of Friends Studymates and Seatmates on Academic Performance." URL: https://dl.dropboxusercontent.com/u/7436354/website/peer-effect-multiple-endogenous-network.pdf. [111, 112, 124]

Chen, Roy, and Jie Gong. 2018. "Can self selection create high-performing teams?" *Journal of Economic Behavior and Organization* 148: 20–33. DOI: https://doi.org/10.1016/j.jebo.2018.02.004. [112, 140]

Chevalier, Judith A., M. Keith Chen, Peter E. Rossi, and Emily Oehlsen. Forthcoming. "The value of flexible work: Evidence from uber drivers." *Journal of Political Economy*, [113]

Cicala, Steve, Roland Fryer, and Jörg Spenkuch. 2018. "Self-Selection and Comparative Advantage in Social Interactions." *Journal of the European Economic Association* 16 (4): URL: https://scholar.harvard.edu/files/fryer/files/ComparativeAdvantage.pdf. [111, 124]

Corgnet, Brice, Joaquín Gómez-Miñambres, and Roberto Hernán-González. 2015. "Goal Setting and Monetary Incentives: When Large Stakes Are Not Enough." *Management Science* 61 (12): 2926–44. DOI: 10.1287/mnsc.2014.2068. eprint: https://doi.org/10.1287/mnsc.2014.2068. [113]

Deci, Edward, and Richard Ryan. 1985. *Intrinsic motivation and self-determination in human behavior*. Springer Science & Business Media. [111, 112, 136, 139, 146]

Deci, Edward, and Richard Ryan. 2000. "The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior." *Psychological Inquiry* 11 (4): 227–68. [111, 136]

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 2011. "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences." *Journal of the European Economic Association* 9 (3): 522–50. DOI: 10.1111/j.1542-4774.2011.01015.x. [116]

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review* 101 (5): 1739–74. [112, 136]

Elsner, Benjamin, and Ingo Isphording. 2017. "A Big Fish in a Small Pond: Ability Rank and Human Capital Investment." *Journal of Labor Economics* 35 (3): 787–828. [124, 161]

Falk, Armin, and Michael Kosfeld. 2006. "The Hidden Costs of Control." *American Economic Review* 96 (5): 1611–30. DOI: 10.1257/aer.96.5.1611. [113, 136]

Friebel, Guido, Matthias Heinz, Mitchell Hoffman, and Nick Zubanov. 2019. "What Do Employee Referral Programs Do?" [140]

Fu, Chao, and Nirav Mehta. 2018. "Ability Tracking, School and Parental Effort, and Student Achievement: A Structural Model and Estimation." *Journal of Labor Economics* 36 (4): 923–79. DOI: 10.1086/697559. eprint: https://doi.org/10.1086/697559. [136]

Garlick, Robert. 2018. "Academic Peer Effects with Different Group Assignment Policies: Residential Tracking versus Random Assignment." *American Economic Journal: Applied Economics* 10 (3): 345–69. [112, 136]

Gibbons, Frederick, and Bram Buunk. 1999. "Individual Differences in Social Comparison: Development of a Scale of Social Comparison Orientation." *Journal of Personality and Social Psychology* 76 (1): 129–47. [116]

Gill, David, Zdenka Kissová, Jaesun Lee, and Victoria Prowse. 2019. "First-Place Loving and Last-Place Loathing: How Rank in the Distribution of Performance Affects Effort Provision." *Management Science* 65 (2): 494–507. [124, 161]

Gneezy, Uri, and Aldo Rustichini. 2004. "Gender and Competition at a Young Age." *American Economic Review* 94 (2): 377–81. DOI: 10.1257/0002828041301821. [115, 163]

Golsteyn, Bart, Arjan Non, and Ulf Zölitz. 2017. "The Impact of Peer Personality on Academic Achievement." URL: https://www.zora.uzh.ch/id/eprint/141964/1/econwp269.pdf. [112, 124]

Harrison, Glenn, and John List. 2004. "Field Experiments." *Journal of Economic Literature* 42 (4): 1009–55. DOI: 10.1257/0022051043004577. [110]

Heckman, James, and Rodrigo Pinto. 2015. "Econometric Mediation Analyses: Identifying the Sources of Treatment Effects from Experimentally Estimated Production Technologies with Unmeasured and Mismeasured Inputs." *Econometric Reviews* 34 (1-2): 6–31. DOI: 10.1080/07474938.2014.944466. eprint: http://dx.doi.org/10.1080/07474938.2014.944466. [123, 145, 146]

Herbst, Daniel, and Alexandre Mas. 2015. "Peer Effects on Worker Output in the Laboratory Generalize to the Field." *Science* 350 (6260): 545–49. DOI: 10.1126/science.aac9555. eprint: http://science.sciencemag.org/content/350/6260/545.full.pdf. [111]

Huettner, Frank, and Marco Sunder. 2012. "Axiomatic arguments for decomposing goodness of fit according to Shapley and Owen values." *Electronic Journal of Statistics* 6: 1239–50. [128]

Irving, Robert. 1985. "An Efficient Algorithm for the "Stable Roommates" Problem." *Journal of Algorithms* 6 (4): 577–95. [117]

Jackson, C. Kirabo, and Elias Bruegmann. 2009. "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers." *American Economic Journal: Applied Economics* 1 (4): 85–108. [113]

Kiessling, Lukas, Jonas Radbruch, and Sebastian Schaube. 2019. "Determinants of Peer Selection." Working Paper. [112, 121]

Koch, Alexander K., and Julia Nafziger. 2011. "Self-regulation through Goal Setting*." *Scandinavian Journal of Economics* 113 (1): 212–27. DOI: 10.1111/j.1467-9442.2010.01641.x. [113]

Kosfeld, Michael, and Susanne Neckermann. 2011. "Getting More Work for Nothing? Symbolic Awards and Worker Performance." *American Economic Journal: Microeconomics* 3 (3): 86–99. [113]

Lavy, Victor, and Edith Sand. Forthcoming. "The Effect of Social Networks on Students' Academic and Non-cognitive Behavioural Outcomes: Evidence from Conditional Random Assignment of Friends in School." *Economic Journal*, URL: https://onlinelibrary.wiley.com/doi/full/10.1111/ecoj.12582. [111]

Lazear, Edward P., and Paul Oyer. 2012. "Personnel Economics." In. *The Handbook of Organizational Economics*. Princeton University Press, 479–519. URL: http://www.jstor.org/stable/j.ctt1r2ggg.16. [140]

Lazear, Edward, and Kathryn Shaw. 2007. "Personnel Economics: The Economist's View of Human Resources." *Journal of Economic Perspectives* 21 (4): 91–114. DOI: 10.1257/jep.21.4.91. [140]

Levitt, Steven, John List, Susanne Neckermann, and Sally Sadoff. 2016. "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance." *American Economic Journal: Economic Policy* 8 (4): 183–219. DOI: 10.1257/pol.20130358. [113, 139]

Manski, Charles. 1993. "Identification of Endogenous Social Effects: The Reflection Problem." *Review of Economic Studies* 60 (3): 531–42. URL: http://www.jstor.org/stable/2298123. [111]

Mas, Alexandre, and Enrico Moretti. 2009. "Peers at Work." *American Economic Review* 99 (1): 112–45. DOI: 10.1257/aer.99.1.112. [109, 113]

Murphy, Richard, and Felix Weinhardt. 2018. "Top of the Class: The Importance of Ordinal Rank." [124, 161]

Oster, Emily. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics* 37 (2): 187–204. [133, 134]

Owens, David, Zachary Grossman, and Ryan Fackler. 2014. "The Control Premium: A Preference for Payoff Autonomy." *American Economic Journal: Microeconomics* 6 (4): 138–61. DOI: 10.1257/mic.6.4.138. [112, 136]

Rotter, Julian. 1966. "Generalized Expectancies for Internal Versus External Control of Reinforcement." *Psychological Monographs: General and Applied* 80 (1): 1–28. [116]

Sacerdote, Bruce. 2001. "Peer Effects with Random Assignment: Results for Dartmouth Roommates." *Quarterly Journal of Economics* 116 (2): 681–704. URL: http://EconPapers.repec.org/RePEc:oup:qjecon:v:116:y:2001:i:2:p:681-704.. [109]

Sacerdote, Bruce. 2011. "Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?" In. Edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Vol. 3, Handbook of the Economics of Education. Elsevier, 249–77. DOI: https://doi.org/10.1016/B978-0-444-53429-3.00004-1. [111]

Schneider, Simone, and Jürgen Schupp. 2011. "The Social Comparison Scale: Testing the Validity, Reliability, and Applicability of the IOWA-Netherlands Comparison Orientation Measure (INCOM) on the German Population." *DIW Data Documentation*, URL: https://www.diw.de/sixcms/detail.php?id=diw_01.c.368791.de. [116]

Sutter, Matthias, and Daniela Glätzle-Rützler. 2015. "Gender Differences in the Willingness to Compete Emerge Early in Life and Persist." *Management Science* 61 (10): 2339–54. DOI: 10.1287/mnsc.2014.1981. eprint: http://dx.doi.org/10.1287/mnsc.2014.1981. [115]

Sutter, Matthias, Stefan Haigner, and Martin G. Kocher. 2010. "Choosing the Carrot or the Stick? Endogenous Institutional Choice in Social Dilemma Situations." *Review of Economic Studies* 77 (4): 1540–66. URL: http://www.jstor.org/stable/40836657. [113]

Tincani, Michela. 2017. "Heterogeneous Peer Effects and Rank Concerns: Theory and Evidence." URL: http://www.homepages.ucl.ac.uk/~uctpmt1/Tincani_rank_240117.pdf. [111, 124]

Weinhardt, Michael, and Jürgen Schupp. 2011. "Multi-Itemskalen im SOEP Jugendfragebogen." *DIW Data Documentation*, URL: https://www.diw.de/sixcms/detail.php?id=diw_01.c.386630.de. [115]

Chapter 4

Interview Sequences and the Formation of Subjective Assessments*

Joint with Amelie Schiprowski

4.1 Introduction

Subjective assessments are commonly used to measure quality and performance in high-stakes situations. Examples include the grading of students, the evaluation of employees and the screening of applicants. In these and many other settings, subjective assessments can have long-lasting consequences for individual outcomes. It is therefore central to understand their formation process.

One context where subjective assessments are especially prevalent is the process of hiring or admitting candidates. In particular, candidates are usually assessed through personal interviews, which tend to happen at decisive stages. As in most assessment situations, interviewers rarely rate a single candidate in isolation. Instead, they conduct several interviews sequentially before determining assessments. This process can help the interviewer to learn about the underlying distribution of candidate quality, and thereby reduce uncertainty about the assessment standard. The learning process can, however, be disrupted if recently observed signals produce spillovers on the perception of the current signal. Such spillovers can arise from unconscious behavioral biases or heuristics, such as contrast effects in the perception of candidate quality (e.g., Simonson and Tversky, 1992; Bordalo, Gennaioli, and

* We thank Johannes Abeler, Steffen Altmann, Pedro Bordalo, Stefano DellaVigna, Thomas Dohmen, Armin Falk, Andreas Grunewald, Randi Hjalmarsson, Michael Kosfeld, Lena Janys, George Loewenstein, Andreas Klümper and Ulf Zoelitz for helpful comments. The paper further benefited from comments received at the University of Bonn, IZA, DIW Berlin, the CRC 224 Conference, the briq/IZA workshop on the Behavioral Economics of Education, the Colloquium on Personnel Economics 2019, the University of Cologne and the CESifo Conference on Behavioral Economics. We thank the study grant organization for the provision of the data and for numerous fruitful discussions.

Shleifer, 2019b). If interviewers are prone to such influences, they can induce firms and organizations to systematically hire or admit the wrong candidates.

In this paper, we study how a candidate's assessment outcome is influenced by the other candidates seen by the same interviewer, and how the influence varies with the relative timing between two interviews. We rely on novel administrative data from a study grant admission process with high stakes.[1] The process is organized through assessment center style admission workshops. Every workshop includes eight interviewers who each assess the quality, i.e., the fit with the selection critieria, of about twelve candidates. The interviewers do not face an admission quota. In total, we observe 9,420 assessments made by 815 interviewers at 102 admission workshops. Three main features make the setup ideal to study how candidates influence each others' assessments: first, candidates are quasi-randomly assigned to interviewers and time slots. Second, each candidate has a clearly defined reference group, as interviewers observe a closed sequence of candidates and make final assessments at the end of the workshop. Third, each candidate receives three as-good-as independent assessments, which facilitates the measurement of otherwise unobserved candidate quality.

Exploiting the quasi-random assignment and ordering of candidates, we estimate the causal effect of the other candidates' quality on an individual's assessment. We proxy a candidate's unobserved quality through an independent third party assessment (TPA). More formally, the TPA is defined as the sum of ratings that two other interviewers independently gave to the same candidate.[2] Results show that the same candidate is evaluated worse when assigned to an interview sequence with better candidates. Candidates observed previously as well as subsequently have a similiar negative influence. A striking exception to the pattern is the previous candidate, whose influence exceeds any other candidate's influence by a factor of about three. When conditioning on the average TPA of all other candidates, only the previous candidate shows an additional influence.

The previous candidate's additional influence translates into substantial changes in the current candidate's admission outcomes. If the TPA of the previous candidate increases by one standard deviation, the interviewer is about 5 percentage points less likely to vote in favor of admitting the current candidate (10% relative to the mean), and the rank of the current candidate decreases by about 0.25. As a result of the previous candidate's influence, interviewer assessments display a substantial autocorrelation: conditional on her mean assessment, an interviewer who votes in

1. The study grant can finance up to the entire living costs of accepted students during their university studies. It further offers access to a large network and a rich, cost-free program of academic courses and events.

2. Importantly, the other interviewers see the same candidate at different points in time and in different interview sequences.

favor of admitting a candidate observed in period $t-1$ is about 20% less likely to vote in favor of the candidate observed in period t.

We then investigate how the role of relative quality differences between subsequent candidates. We find that a marginal change in the previous candidate's TPA has a stronger effect if the current candidate has a similar TPA. The effect is most pronounced at the margin of being just better than the previous candidate. The pattern suggests that an exaggeration of small positive quality differences between subsequent candidates drive the effect.

To guide the discussion of potential mechanisms, we set up a simple conceptual framework where interviewers perform Bayesian updating about the quality of candidates after observing a sequence of noisy signals. The process of learning rationalizes why the other candidates' average quality influences the assessment of a single candidate. It can, however, not explain the additional influence of the previous candidate. Instead, the discussion suggests a sequential contrast effect as the main driver. More precisely, the empirical results point towards a non-linear version of the contrast effect, where absolute differences between current and previous quality have a decreasing marginal effect. Alternative mechanisms, in particular a gambler's fallacy where the previous candidate affects the interviewer's priors about the next candidate, are not in line with the empirical pattern.

In a further step, we explore how the effect varies with the sequencing of gender. Results show a striking asymmetry: while the gender of the previous candidate does not matter for female candidates, male candidates are not harmed by following a strong female candidate. This asymmetry in the previous candidate's influence has relevant implications for the 'gender assessment gap': males who follow a male candidate are 5% more likely to receive a yes vote than females; males who follow a female candidate are 20% more likely.

The findings of this paper imply that minor changes in relative candidate ordering can have major consequences on labor market careers. This has relevant implications for many hiring and admission situations, the economics job market meetings being only one of many settings where candidates are assessed through sequential interviews. Despite the strategic importance of hiring and admission decisions for firms and organizations, the underlying assessment process is still a black box (Oyer and Schaefer, 2011). But even beyond the context of hiring, little is known on the formation of subjective assessments, which are a central determinant of individual outcomes in many labor market contexts. A few studies show that extraneous factors can have an undesired influence on subjective assessments. For instance, evaluations of performance or quality have been found to be influenced by absolute order of appearance (Ginsburgh and Ours, 2003), the reference group (Calsamiglia and Loviglio, 2019) or narrow brackets among the evaluator (Simonsohn and Gino, 2013). We contribute by showing that sequential information processing, inherent to many assessment setups, can create distortive spillovers between assessments. This clearly questions the reliance on (single) subjective assessments and provides a

rationale for the success of technology-based screening devices (cf. Hoffman, Kahn, and Li, 2018).

As a second contribution, we identify a novel mechanism through which gender gaps in subjective assessments can emerge. Women have been shown to receive lower assessments for the same quality or performance in different setups, such as hiring and promotion decisions, teaching evaluations or referee reports (e.g., Neumark, Bank, and Van Nort, 1996; Rouse and Goldin, 2000; Mengel, Sauermann, and Zölitz, 2018). A larger presence of female evaluators has been found not to influence the assessment gap (Bagues, Sylos-Labini, and Zinovyeva, 2017). Recent studies try to uncover behavioral mechanisms underlying gender gaps. Bohren, Imas, and Rosenberg (2019) document the importance of biased initial beliefs about ability, while Sarsons (2019) identifies gender asymmetric belief updating about ability as an important mechanism. In this paper, we provide evidence for the role of gender asymmetries in the comparative assessment of quality or ability, which has, to our knowledge, not yet been identified in the literature. One potential reason for this is that a person's point of comparison is usually neither clearly defined nor quasi-randomly assigned. As our sequential setup overcomes this challenge, we can provide clean evidence that the comparative perception of signals can work in the favor of men and thereby widen the gender assessment gap.

More generally, the findings in this paper imply that individuals in groups do not only influence each other through social interactions, but also through the perception and evaluation by an observer. This relates to the literature on reference groups and peer effects (see for an overview Sacerdote, 2011; Sacerdote, 2014; Herbst and Mas, 2015). A few papers in this literature have exploited spatial arrangements to identify single workers as the relevant peers (e.g., Mas and Moretti, 2009). Similarly, we exploit the sequential arrangement of candidates to identify the previous candidate as the one with the strongest influence.

Finally, we add to the literature on path dependence in decision-making. Several studies have provided evidence that current decisions can be influenced by the characteristics and outcome of prior decisions. A positive path dependence has been found for jury decision making in criminal courts (Bindler and Hjalmarsson, 2018) and sport judges (Damisch, Mussweiler, and Plessner, 2006; Kramer, 2017). Studies documenting a negative path dependence are based on the contexts of speed dating (Bhargava and Fisman, 2014), asylum court judges, baseball umpires and loan officers (Chen, Moskowitz, and Shue, 2016), as well as financial investors (Hartzmark and Shue, 2018). Furthermore, Simonsohn and Loewenstein (2006) and Simonsohn (2006) show that movers' rental choices are influenced by prior experience of housing prices and commuting times. While Chen, Moskowitz, and Shue (2016) attribute their findings mainly to a gambler's fallacy, Bhargava and Fisman (2014), Hartzmark and Shue (2018), Simonsohn and Loewenstein (2006) and Simonsohn (2006) point to the existence of sequential contrast effects in their settings. More recently, Bordalo, Gennaioli, and Shleifer (2019b) provide a theoretical foundation

of sequential contrast effects. In their model, a contrast arises due to the comparison of alternatives to a retrieved norm from past experiences. In a companion paper, (Bordalo, Gennaioli, and Shleifer, 2019a), they reinterpret data on movers' rental choices in the light of their model. Our paper contributes to this literature in three main regards. First, we provide ample evidence of negative path dependence in subjective assessments, a yet unstudied context which is of key importance in many labor market setups. Second, we generalize from settings with sequential decisions, where decisions are made directly, to a setting where decisions are made at the end of a closed sequence. In this respect, we provide first evidence that the instantaneous error in perceptions, caused by the prior signal, is strong enough to persist up to a final decision where ex-post adjustments are relevant. Third, we document that the effect is driven by small differences in subsequent quality and point out its gender asymmetric pattern.

The remainder of the paper is structured as follows: Section 2 summarizes the institutional setting and background. Section 3 describes and summarizes the data. We study the influence of the interview sequence in section 4. Section 5 analyzes the additional influence of the previous candidate and presents the results in terms of the negative autocorrelation. Section 6 discusses the underlying mechanism. Section 7 explores the role of gender. Section 8 quantifies the results and discusses implications and section 9 concludes.

4.2 Institutional Setting

We exploit detailed information on interviews conducted during the admission workshops of a large, merit-based study grant program for university students in Germany. The program yields a large number of monetary and non-monetary benefits. The workshops therefore have high stakes from the candidates' perspective.[3] In the following, we describe the setup and schedule of the workshops. Additional institutional background on the study grant is provided in appendix 4.A.

Background. Admission workshops take place over the course of one weekend and resemble the structure of assessment centers. The admission committee is formed

3. During the sample period, the monetary scholarship ranged between 1,800 and about 10,000 euros per year, depending on parents' earnings. Given that there are (almost) no tuition fees at German universities, the scholarship can cover up to the entire living costs of a student. Non-monetary benefits include the access to cost-free summer schools and language classes, a strong signal on the CV, as well as networking opportunities. Students are admitted for the period of their entire university studies, subject to a positive interim evaluation.

by eight interviewers.[4] About 48 candidates participate at each workshop.[5] An employee of the organization that administers the scholarship is present throughout the course of the workshop.

Candidates are first year university students. They were pre-selected by their high school principals, who can nominate about 2% of a graduating cohort. Nominated students are invited to attend an admission workshop. Prior to participation, they have to hand in a written CV and their school transcripts. During the workshop, each candidate participates in two one-to-one interviews of 35 minutes and a group discussion round. Each task is assessed by a different interviewer, implying that every candidate receives three independent assessments – one per interview and one for the group discussion. The final decision is based on the sum of the three equally weighted assessments.

Interviewers are scholarship alumni working in diverse professions. They commonly participate at one admission workshop every one or two years. Interviewers do not receive any information about the candidates before the workshop, and vice versa. No interaction between the committee and candidates takes place afterwards. The workshop therefore constitutes a closed sequence of interaction.

The assignment of candidates to interviewers, and the assignment of time slots are quasi-randomized (cf. randomization checks in section 4.3).[6] Both candidates and interviewers receive an ID. A fixed schedule then matches candidate IDs to interviewer IDs and time slots. Candidates and interviewers do not know the assignment ex ante.[7]

Workshop Schedule. Table 4.1 sketches an interviewer's schedule during the admission workshop. Upon arrival on Friday night, interviewers receive a short briefing by an employee of the scholarship organization and prepare the interviews which they conduct on Saturday. For this purpose, they receive the candidates' CV, school records and a letter of recommendation written by the high school principal. On Saturday, each interviewer conducts six interviews and rates five group discussions. In the evening, interviewers receive the documents of the candidates they interview on Sunday and prepare the interviews. On Sunday, they conduct six interviews and

4. Technically, members of the committee do not only act interviewers, but also as passive observers of candidate behavior in a group task. As our focus is on the interview assessments, we refer to committee members as interviewers throughout the paper.

5. The baseline seminar schedule is designed for 48 candidates. Anticipating short notice cancellation, the program slightly over-books each workshop. If more or less than 48 candidates show up, the workshop follows a slightly adjusted schedule. We observe the actual schedule with the actual number of participants.

6. Randomization occurs conditional on gender, with the aim of ensuring a balanced gender composition in the group discussion.

7. Interviewers know their ID prior to the workshop, but they do not have any information on their assigned candidates, which renders the knowledge irrelevant.

Table 4.1. Stylized Interviewer Schedule

	Friday	**Saturday**	**Sunday**
Morning		**interviews** (≈ 3) + group discussions (≈ 3)	**interviews** (≈ 6) + group discussion (≈ 1)
Afternoon		**interviews** (≈ 3) + group discussions (≈ 2)	committee meeting
Evening	preparation	preparation	

rate one group discussion.[8] The detailed schedule including candidate assignments to interviewers and time slots is shown in Appendix Figure 4.A.1.

Assessment and Admission Decision. We focus here on the formation of assessments through one-to-one interviews. Interviewers are asked to evaluate candidates according to their intellectual abilities, ambition and motivation, communication skills, social engagement and broadness of interests, which comprise the program's selection criteria. They summarize their assessment on a scale from 1 to 10. A rating of 8 points and above implies a yes vote, i.e., an assessment in favor of accepting the candidate. 9 points are supposed to reflect a strong yes vote and 10 points are reserved for outstanding candidates. A candidate is accepted for the scholarship program if he or she receives at least two yes votes and a total of at least 23 points. Interviewers are informed about these rules at the start of the workshop. Moreover, the employee of the institution states explicitly that there is no admission quota and that the committee can in principle admit every or none of the candidates present at the workshop. Upon request, interviewers are informed that the average admission rate is around 25%, with large variation between workshops.

Interviewers are asked to determine their individual assessments after having seen all their assigned candidates. Importantly, they are not allowed to exchange with other interviewers about candidates before the final committee meeting. This rule is strictly enforced by the employee of the scholarship organization, who wants to ensure that every candidate receives the chance of being evaluated independently. Further, interviewers have a high intrinsic motivation for compliance, as they are alumni who have received many benefits from the program. In the final meeting, which takes place on Sunday afternoon, a list with candidate IDs is read out aloud

8. Every group discussion includes approximately six candidates and takes place six times over the workshop. In each round, one candidate has to give a short presentation on a self-chosen topic, and moderate a discussion. Interviewers do not interfere in the discussion at any time. Moreover, they do not have any information about the candidates they observe in the group discussions, except for their names, study major and visually observable characteristics such as gender. They base their rating on the candidate's presentation and her contributions to the discussion.

and every interviewer who has assessed the respective candidate states her rating.[9] Subsequently, ratings are aggregated and, after a short justification, candidates at or above the cut-off of 23 points are accepted for the scholarship. In exceptional cases, ratings for candidates at the margin to admission can be adjusted after a discussion by the committee. According to information by employees of the program, such adjustments happen usually in about two to three out of about 150 votes per workshop. We observe the final ratings of each candidate.[10]

4.3 Data and Measurement

In this section, we describe the data source, explain our baseline measure of candidate quality and assess the random assignment and ordering of candidates.

4.3.1 Data Description

Data Source & Sampling. The data cover the full population of participants at the admission process for recent high school graduates during the university year 2012/13. The data contain 102 admission workshops. We drop 9 candidates (0.002%) because we do not observe their assigned candidate ID. The final sample includes 9,420 interview ratings of 4,710 candidates, made by 815 different interviewers.[11]

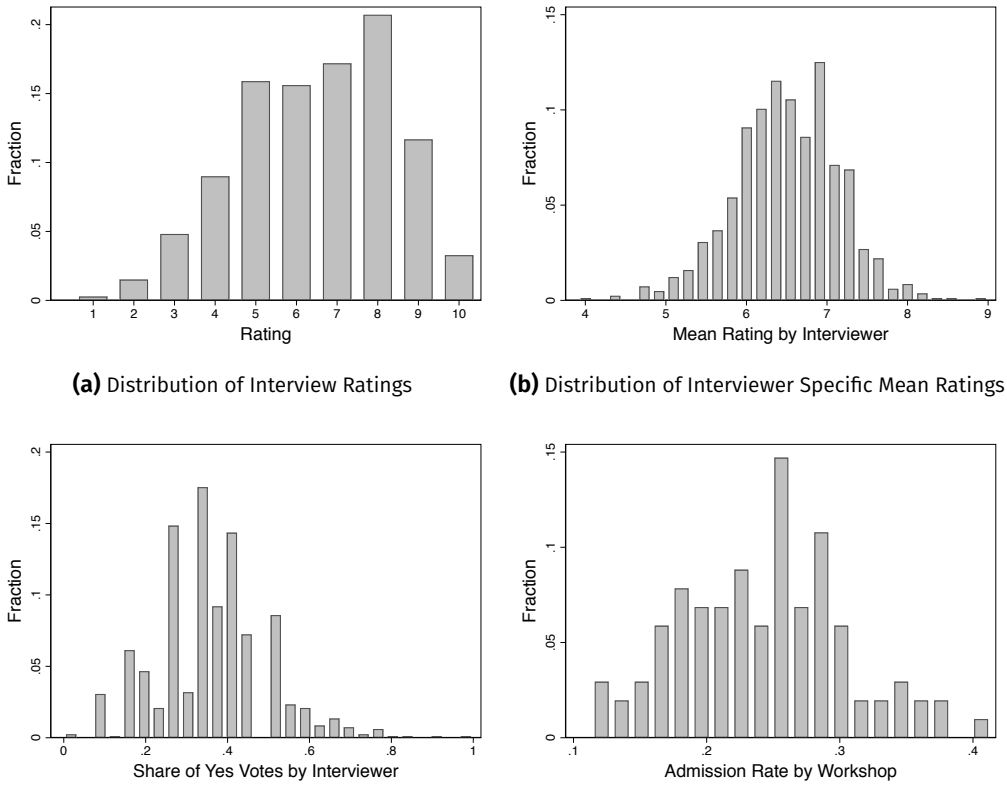
The data report for each candidate her interview and group discussion slots, as well as the resulting ratings and admission decision. In addition, they include candidate gender, age, field of study, high school grade, an indicator of migration background and an indicator of a non-academic parental background. We further observe basic characteristics of the interviewer: gender, field of study, age and prior workshop experience.

Summary Statistics. Figure 4.2a plots the sample distribution of individual interview ratings. Ratings range from 1 to 10, and the largest mass of ratings lies between 5 and 8 points. The average rating in the sample is 6.5 points, with a standard deviation of 1.9. For the empirical analysis, we standardize the overall rating distribution to have mean zero and standard deviation one. As shown in Figures 4.2b and 4.2c, there is substantial heterogeneity in average ratings and the share of yes votes across interviewers. In line with the institutional feature that interviewers do not face a quota, the interviewer-specific share of yes votes ranges from 0 to 100%,

9. In this process, it is not easily possible to trace the behavior of other interviewers, as assessments are collected with high frequency and not ordered with respect to interviewer IDs.

10. To ensure that the adjustment procedure does not influence our results, we run robustness checks where marginal candidates are excluded.

11. Three interviewers participate at two different workshops. We treat each workshop participation as independent.



(a) Distribution of Interview Ratings **(b)** Distribution of Interviewer Specific Mean Ratings

(c) Distribution of Interviewer Specific Yes Vote Shares **(d)** Distribution of Workshop Specific Admission Rates

Figure 4.1. Distribution of Assessments at the Individual and Aggregate Level

Notes: Panel (a) shows the distribution of interview ratings (N=9,402). Panel (b) shows the mean of ratings at the interviewer level (N=815). Panel (c) shows the share of yes votes (ratings of ≥ 8) at the interviewer level (N=815). Panel (d) shows the admission rate at the workshop level (N=102).

with a mean of 36 % and a standard deviation of 17%. This also translates into a wide range of workshop-specific admission rates from about 10% to about 40% of candidates. The average workshop has an admission rate of 24%, with a standard deviation of 6%.

Table 4.2 reports summary statistics on candidate and interviewer characteristics. The average interviewer age is 40, and close to half of interviewers are female. Interviewers come from various study backgrounds. About one third of interviewers participate in their first workshop, about one third has one or two prior workshop participations, and about one third has participated three or more times.

The average candidate age is 20. 54% of candidates female, 14% have a migration background and 25% a non-academic parental background. The average applicant achieved a GPA of 89% of the maximum high school final grade.

Table 4.2. Summary Statistics on Interviewer and Candidate Characteristics

	Interviewers		
	N	Mean	SD
Age	815	40.05	9.45
Female	815	0.47	0.50
Major: Humanities	815	0.29	0.46
Major: Social Sciences	815	0.20	0.40
Major: STEM	815	0.36	0.48
Major: Medicine	815	0.11	0.31
Major: Others	815	0.03	0.17
Experience: 0	815	0.30	0.46
Experience: 1	815	0.22	0.41
Experience: 2	815	0.17	0.38
Experience: 3+	815	0.31	0.46
Number of cases	815	11.56	0.85
	Candidates		
	N	Mean	SD
Female	4710	0.54	0.50
Age	4710	19.91	1.13
Migration Background	4710	0.14	0.35
Parents w/out Univ. Degree	4710	0.25	0.43
High School GPA (in %)	4710	89.26	5.94
Major: Humanities	4710	0.18	0.38
Major: Social Sciences	4710	0.21	0.41
Major: STEM	4710	0.37	0.48
Major: Medicine	4710	0.22	0.41
Major: Others	4710	0.02	0.14

4.3.2 Third Party Assessment As a Measure of Candidate Quality

Our aim is to analyze how one candidate's quality, as perceived by the interviewer, influences the assessment of another candidate in the same sequence. Given the institutional context, quality denotes how well a candidate meets the study grant selection criteria (see section 4.2).

As true quality is unobserved by design, we need to rely on an independent measurement of a candidates' fit with the criteria. This measurement needs to be independent of the interviewer and of the other candidates seen by that interviewer –most importantly the previous one.[12]

Our preferred measure of a candidate's quality is based on third party assessments (TPA) of that candidate's quality and fit with the selection criteria. More precisely, the third party assessment is the sum of the candidate's other two ratings at the workshop. These ratings are made independently by two of the other seven interviewers at the workshop. One of the other ratings is based on the candidate's second interview and the other on her performance in the group task.[13] The main idea behind this approach is twofold: first, interviewers use the same criteria when rating quality. Second, they measure these criteria with noise – but their noise terms are independent of each other. Below, we discuss these two advantages in more detail.

The first advantage of TPA as a measure of candidate quality is that all interviewers are supposed to rate the same dimensions of quality and ability. The correlation between the individual rating and the sum of the other two interviewers' ratings is 0.35.[14] Given that interviewers differ in their leniency and see the same candidate under different circumstances, we interpret this correlation as rather strong.[15]

The second advantage is that the other two interviewers' ratings are as good as independent of the interviewer's own assessment behavior. This is on the one hand due to the workshop schedule. Interviewers see the same candidate at very different points in time and the sets of candidates seen by two interviewers hardly overlap (see also workshop schedule in appendix figure 4.A.1).[16] In particular, two inter-

12. As interviewers potentially adjust individual ratings ex-post, all candidates seen by the same interviewer can influence each other. Therefore, we cannot simply use the correlation between two assessments by the same interviewer to study how one candidate influences a subsequent candidate's assessment.

13. Combining both ratings for the quality measure has the advantage of reducing noise. As a robustness check, we also run analyses using either only the other interview rating or only the group discussion rating to measure quality.

14. The two interview ratings are correlated by a factor of 0.33. As expected, the correlation with the group discussion rating, which is based on a different task, is smaller and amounts to about 0.23.

15. As one point of comparison, Card, DellaVigna, Funk, and Iriberri (2019) find a correlation of 0.25 between two referee reports of the same paper in four leading journals in economics.

16. Importantly, this implies that interviewer A's assessment of a given candidate is influenced by a different set of other candidate than interviewer B's assessment of that candidate.

viewers never see the same two candidates in the same relative order. Moreover, interviewers are not allowed to discuss candidates before ratings are joined in the final committee meeting. This rule is enforced by the employee of the scholarship organization who is present throughout the workshop (see section 4.2). Nevertheless, we cannot completely rule out by design that informal discussions on candidates take place. Furthermore, the independence assumption would be violated if a candidate's experience in her first interview influenced her behavior in the second interview or in the group discussion. In appendix table 4.B.1, we provide evidence against such spillovers between ratings of the same candidate. The idea is that under independence, the characteristics of an interviewer should influence this interviewer's own rating, but not the other two interviewers' ratings of the same candidate. For instance, female interviewers are on average more lenient. Hence, a female interviewer in the first interview should increase the rating of the first interview, but not the ratings in the other two interviews. If, instead, the experience of having a female interviewer strongly influences the candidate's behavior, we should observe a correlation with the other ratings as well. Similarly, we would observe such a correlation if interviewers systematically influence each others' assessments. Table 4.B.1 speaks against such spillovers. Estimates report that a candidate's rating made by a given interviewer is influenced by that interviewer's characteristics[17] On the contrary, the rating is not influenced by the characteristics of the other two interviewers who saw the same candidate. This holds for both interviews (columns 1 and 2) and the group discussion (column 3).

An alternative way to measure candidate quality is through pre-determined characteristics, in particular high school GPA. However, GPA is likely to be only a poor predictor of fit with the scholarship criteria, which go beyond grade performance. Furthermore, the fact that candidates are pre-selected based on their high school performance implies a low amount of variation in GPA. Indeed, Appendix Table 4.B.2 shows that individual assessments increase in high school GPA, but the power of all observed candidate characteristics to predict interview assessments is low (R-Squared ≈ 0.04). We nevertheless construct a quality measure based on observed candidate characteristics and use it to check the robustness of the qualitative results pattern we find.

4.3.3 Randomization Checks

Causal identification relies on the assumption that individuals are as good as randomly assigned to interviewers and as good as randomly ordered within an interview sequence. These conditions should be met by the institutional setup (see section 4.2). The only candidate characteristic taken into account for the assignment of candidate

17. More precisely, the rating increases in the interviewer's age and decreases in her experience. Further, female interviewers rate candidates better on average.

Table 4.3. Test of Quasi-Random Assignment

	TPA (1)	GPA (2)	Age (3)	Migrant (4)	Par. Non-Acad. (5)	STEM (6)
Interv. Leave-One-Out Mean	0.002	-0.002	-0.002	-0.003	0.001	-0.000
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)
Outcome Mean	12.55	89.26	19.91	0.14	0.25	0.37
N	9420	9420	9420	9420	9420	9420

Notes: Regressions control for gender and include workshop fixed effects. Following Guryan, Kroft, and Notowidigdo (2009), we control for the fact that an individual cannot be assigned to herself using the workshop leave-out mean of the respective variable. TPA = third party assessment of candidate quality (see section 4.3.2). $^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$. Standard errors are clustered at the workshop level (N=102).

IDs is gender, because the scholarship organization aims at having gender-balanced group discussions. We thus assume the random assignment and ordering conditional on own gender. In the following, we assess this central assumption.

Quasi-random assignment to interviewers implies that the characteristics of a candidate assigned to interviewer i are not systematically related to the characteristics of the other candidates assigned to i. We test this implication by regressing the third party assessment (TPA) of a candidate interviewed by i on the leave-out mean TPA of the other candidates interviewed by i, conditional on gender and workshop fixed effects.[18] The result is reported in column 1 of Table 4.3. In columns 2 to 5, we perform the same exercise using observed pre-determined candidate characteristics. In line with quasi-random assignment at the workshop level, we find no association between the individual characteristics and their leave-out mean at the interviewer level.[19]

To assess the assumption of quasi-random ordering, we test for the presence of an auto-correlation in candidate characteristics, conditional on own gender. Table 4.4 presents the results from a regression of the current candidate's characteristics on the previous candidate's characteristics.[20] It shows no indication of systematic ordering by TPA or other observed candidate characteristics.

In section 4.7, we will carry out additional analyses which relate a candidate's assessment to the gender of the previous candidate. The causal interpretation of these analyses will rely on the assumption that, conditional on own gender, the gen-

18. Following Guryan, Kroft, and Notowidigdo (2009), we control for the fact that an individual cannot be assigned to herself using the workshop leave-out mean of the respective variable.

19. In the appendix we additionally provide evidence that candidate and interviewer characteristics are not systematically related. For that purpose we show that interviewer characteristics do not predict candidate characteristics in appendix table 4.B.3

20. Again, following Guryan, Kroft, and Notowidigdo (2009), we control for the fact that an individual cannot follow herself using the leave-out mean of the other candidates assigned to the same interviewer.

Table 4.4. Test of Quasi-Random Ordering

	TPA (1)	GPA (2)	Age (3)	Migrant (4)	Par. Non-Acad. (5)	STEM (6)
t-1	-0.005	0.007	-0.001	0.011	0.002	0.006
	(0.011)	(0.011)	(0.011)	(0.012)	(0.011)	(0.011)
N	8605	8605	8605	8605	8605	8605

Notes: "t-1" refers to the previous candidate in the interview sequence. Regressions control for own gender and include workshop fixed effects. Following Guryan, Kroft, and Notowidigdo (2009), we control for the fact that an individual cannot follow herself using the leave-out mean of the other candidates assigned to the same interviewer. TPA = Third Party Assessment of candidate quality (see section 4.3.2). $^{*}p < 0.10$,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=102).

Table 4.5. Test of Random Quasi-Ordering with Respect to Gender

	TPA (1)	GPA (2)	Age (3)	Migrant (4)	Par. Non-Acad. (5)	STEM (6)
Female (t-1)	-0.007	0.217	-0.011	0.007	0.004	-0.009
	(0.021)	(0.145)	(0.024)	(0.008)	(0.010)	(0.013)
N	8605	8605	8605	8605	8605	8605

Notes: Regressions control for own gender and include workshop fixed effects. Following Guryan, Kroft, and Notowidigdo (2009), we control for the fact that an individual cannot follow herself using the leave-out mean of the other candidates assigned to the same interviewer. TPA = Third Party Assessment of candidate quality (see section 4.3.2). $^{*}p < 0.10$,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=102).

der of the previous candidate is as good as random. Table 4.5 tests this assumption. It shows that individuals who follow a female do not differ in their characteristics from individuals who follow a male.

4.4 Influence of the Interview Sequence

In this section, we estimate if and how much the assessment of a candidate is influenced by the other candidates seen by the same interviewer. In particular, we study how the influence of another candidate varies with the relative timing of her interview.

4.4.1 Empirical Specification

We estimate how the assessment of a candidate interviewed in period t is affected by a measure of candidate quality of the candidate who was interviewed by the same interviewer in period $t+k$. As described in section 4.3.2, we use the other two interviewers' assessment of the same candidate as a proxy of her quality. We

refer to this variable as the third party assessment (TPA). For each value of $k \in \{-11, ..., -1, 1, ..., 11\}$, we perform a separate estimation of the following regression model:[21]

$$Y_{i,t} = \beta_k \ TPA_{i,t+k} + \theta \ \overline{TPA}_{i,-\{t,t+k\}} + \pi \ TPA_{i,t} + X'_{i,t} \ \sigma + \eta_w + \epsilon_{i,t} \qquad (4.1)$$

The outcome variable $Y_{i,t}$ is the standardized rating made by interviewer i of the candidate interviewed in period t. $TPA_{i,t+k}$ is the standardized third party assessment of the candidate interviewed by interviewer i at time $t+k$. The coefficient of interest, β_k, measures the influence of $TPA_{i,t+k}$ on the rating of the candidate interviewed in t.

The standardized leave-two-out mean $\overline{TPA}_{i,-\{t,t-k\}}$ controls for the other candidates' average TPA, excluding the candidate interviewed at time t and the candidate interviewed at time $t+k$. $TPA_{i,t}$ denotes the candidate's own TPA, which is crucial as a control to avoid an exclusion bias due to removal of the candidate herself from the pool of candidates (see also Guryan, Kroft, and Notowidigdo, 2009; Caeyers and Fafchamps, 2016). $X_{i,t}$ includes observed characteristics of candidates and interviewers (see also summary statistics in table 4.2) as well as the absolute number of candidates observed by an interviewer. It also includes fixed effects for the candidate's absolute order in the interviewer's sequence. η_w controls for workshop fixed effects, as randomization occurs at the workshop level.

For each value of $k \in \{-11, ..., -1, 1, ..., 11\}$, we run a separate regression including the largest possible set of candidates, i.e., all candidates for whom period $t+k$ exists.[22]

4.4.2 Results

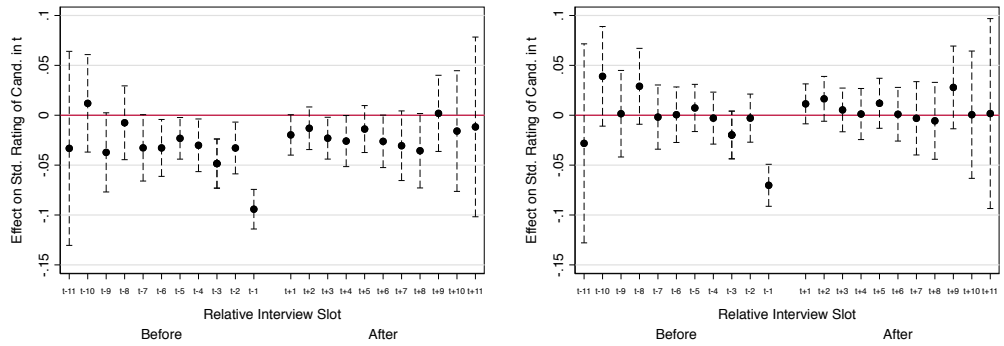
Panel (a) of Figure 4.2 plots the coefficients β_k from equation 4.1, resulting from separate regressions for each value of $k = \{-11, ..., -1, 1, ..., 11\}$. The outcome is the candidate's standardized interview rating. The corresponding coefficients and p-values are shown in Appendix Table 4.C.1.

The figure documents three main results. First, the rating of a candidate decreases in the quality (measured through TPA) of any other candidate seen by the same interviewer.[23] If another candidate's TPA increases by one standard deviation,

21. Recall that the institutional setting allows each of the other candidates within an interview sequence to matter equally, as ratings are set after the last interview took place. Therefore, both previously and subsequently observed candidates can have an influence.

22. For example, only candidates who are interviewed between $t = 1$ and $t = 4$ can have a candidate who appeared eight interview slots later. Vice versa, only candidates between $t = 9$ and $t = 12$ have a candidate who appeared eight interview slots earlier, etc.

23. The coefficients are jointly significant (Wald test, p-value <0.01), see also Appendix Table 4.C.1



(a) Conditional on Leave-Two-Out Mean **(b)** Conditional on Leave-One-Out Mean

Figure 4.2. Effect of Candidate Quality in $t + k$ on Std. Rating of Candidate in t

Notes: Panel (a) shows the coefficients β_k from equation 4.1, resulting from separate regressions for each value of $k = \{-11, ..., -1, 1, ..., 11\}$. The coefficients measure how the standardized TPA of the candidate interviewed in $t + k$ affects the standardized rating of the candidate in t. TPA = Third Party Assessment of candidate quality (see section 4.3.2). Panel (b) estimates the additional effect of the candidate interviewed in $t + k$, beyond her contribution to the leave-one-out mean. Dashed lines show 95% confidence intervals, with standard errors clustered at the workshop level. Appendix Table 4.C.1 reports the corresponding coefficients and p-values.

the candidate's rating decreases by about 2 to 3 % of a standard deviation. Second, candidates seen before ($k < 0$) have the same influence as candidates seen afterwards ($k > 0$) — except for the candidate in $t - 1$.[24] Interviewers thus adjust their ratings ex-post, equally taking into account previously and subsequently observed candidates. Third, the influence of the previous candidate strikingly stands out. If the previous candidate's TPA increases by one standard deviation, the individual rating decreases by almost 10% of a standard deviation. Thereby, the previous candidate's influence exceeds the influence of any other candidate by a factor of about three. Appendix Figure 4.C.1 shows a similar pattern when considering the candidate's probability of receiving a yes vote (rating ≥ 8 points) and her probability of admission.

In panel (b) we control for the average quality of the sequence using the leave-*one*-out TPA, $\overline{TPA}_{i,-t}$. The coefficient, therefore, only measures the additional influence of the candidate in $t + k$. Panel (b) reveals that only the previous candidate has an additional influence on the rating. This suggests the existence of two separate effects: a group effect, which can be captured by the effect of the other candidates' average quality measure, and an additional effect of the previous candidate. While the group effect is symmetric in time, the sequential effect is highly asymmetric: only the candidate observed before has an additional influence.

24. The average of the coefficients with $k < -1$ are statistically not different from the average of the coefficients with $k > 0$ (see also Appendix Table 4.C.1 for the corresponding p-values).

4.5 Additional Influence of the Previous Candidate

We now investigate the additional influence of the previous candidate in more detail. We begin by quantifying the influence on different admission outcomes. We also estimate the negative autocorrelation in assessments resulting from the previous candidate's influence. Moreover, we explore how the previous candidate's influence interacts with the current candidate's own quality.

4.5.1 Empirical Specifications

We study the influence of the previous candidate's quality through the lens of two alternative specifications.

Causal Effect. The first specification is similar to equation 4.1 for $k = -1$, but replaces the leave-two-out mean $TPA_{i,-\{t,t+k\}}$ by the leave-one-out mean $TPA_{i,-\{t,t+k\}}$. By conditioning on the leave-one-out mean, we estimate the additional influence that the previous candidate has beyond contributing to the average quality of the sequence.

$$Y_{i,t} \;=\; \gamma \;\; TPA_{i,t-1} + \kappa \;\; \overline{TPA}_{i,-t} + \lambda \;\; TPA_{i,t} + X'_{i,t} \;\; \phi + \tau_w + \xi_{i,t} \qquad (4.2)$$

The parameter of interest γ estimates how the previous candidate's third party assessment, $TPA_{i,t-1}$, affects interviewer i's assessment of the candidate interviewed in period t, conditional on the leave-one-out mean of the sequence, $\overline{TPA}_{i,-t}$. Therefore, γ measures the additional influence of the candidate in $t-1$, beyond contributing to the group mean. As above, the regression includes candidate and interviewer covariates (including the candidate's order), as well as workshop fixed effects.

Autocorrelation. As a complement, we estimate the autocorrelation in assessments made by the same interviewer. The autocorrelation measures how the assessments of two subsequent candidates influence each other. The immediate advantage of the autocorrelation is that it captures the interviewer's own perception. Given the results of section 4.4, the autocorrelation is likely to be driven by the previous candidate, who is the only candidate having an additional influence beyond its contribution to the group average. Yet, the autocorrelation contains also the influence of the candidate seen in t on the candidate seen in $t-1$. Note that this influence exists, but is not very strong (see section 4.4). In addition, the autocorrelation can include the influence of other prior interviews, to the degree that they have an effect on the perception of the candidate in $t-1$. We estimate the autocorrelation using the following specification:

$$Y_{i,t} \;=\; \delta \;\; Y_{i,t-1} + \alpha \;\; \overline{Y}_{i,-t} + X'_{i,t} \;\; \mu + \omega_w + \zeta_{it} \qquad (4.3)$$

$Y_{i,t}$ and $Y_{i,t-1}$ denote interviewer i's assessment of the candidates in t and $t-1$, respectively. The parameter of interest δ measures the excess autocorrelation between $Y_{i,t}$ and $Y_{i,t-1}$, which reflects the additional influence of the previous candidate, conditional on the average influence of all other candidates. To this end, we control for the interviewer's mean assessment, excluding the candidate in t (leave-one-out mean, $\overline{Y}_{i,-t}$). Note that the leave-one-out-mean assessment also controls for differences in interviewer leniency.[25] $\overline{Y}_{i,-t}$ always contains both the leave-one-out mean rating and the leave-one-out mean share of yes votes, to control for differences in both the average rating on the 1-10 scale and in the propensity to give a yes vote.

As above, the specification controls for workshop fixed effects (ω_w),[26] as well as interviewer and candidate covariates $X_{i,t}$ (including order and here also the measure of candidate quality).

4.5.2 Main Results

Causal Effect. Table 4.6 presents the estimates from equation 4.2. In columns 1 and 2, the outcome is the standardized rating of the candidate interviewed in period t. Column 1 includes only workshop fixed effects and column 2 adds candidate and interviewer characteristics, as well as order fixed effects. Estimates do not differ significantly between the two columns and imply that the average rating decreases by about 7% of a standard deviation in response to a one standard deviation increase in the previous candidate's third party assessment (TPA). This effect is about as large as the effect of a one standard deviation increase in the leave-one-out mean TPA of the sequence (leaving out the candidate in t). It is about one fourth as large as the influence of a one standard deviation increase in the candidate's own TPA. Column 3 shows that the probability that the interviewer gives a yes vote (rating of ≥ 8 points) decreases by 3.7 percentage points (10%) in response to a one standard deviation increase in the previous candidate's TPA. In comparison, the effect of a one standard deviation increase in the overall leave-one-out mean TPA amounts to 2.7 percentage points.

In column 4, the outcome is the candidate's relative rank in the interviewer's rating distribution. Results show that the sequential effect not only affects absolute ratings, but also relative rankings.[27] More precisely, a one standard deviation

25. An alternative strategy to control for leniency differences is the use of interviewer fixed effects. However, as first noted by Nickell (1981), fixed effects introduce a downward bias when auto-regressive models are estimated on finite panels (here: $\overline{T} = 12$). They are therefore not suited in our context.

26. Note that the use of workshop fixed effects in the context of an auto-regressive model also creates the potential for a 'Nickel bias'. However, T now amounts to $\approx 8 \times 12$ (the number of interviewer assessments per workshop), which makes the bias negligible.

27. Note that the mechanical effect of another candidate's quality on the ranking is captured by the leave-one-out mean TPA.

Table 4.6. Influence of the Previous Candidate

	Std. Rating		P(Yes Vote)	Rank	P(Best)	P(Admission)
	(1)	(2)	(3)	(4)	(5)	(6)
TPA (std.), t-1	-0.071***	-0.070***	-0.037***	-0.252***	-0.015***	-0.015***
	(0.010)	(0.011)	(0.005)	(0.037)	(0.003)	(0.004)
Leave-one-out Mean TPA (std.)	-0.096***	-0.088***	-0.027***	-0.323***	-0.019***	-0.010**
	(0.018)	(0.018)	(0.006)	(0.021)	(0.004)	(0.005)
TPA (std.), t	0.344***	0.324***	0.142***	1.146***	0.079***	0.275***
	(0.013)	(0.014)	(0.006)	(0.042)	(0.004)	(0.005)
Controls	No	Yes	Yes	Yes	Yes	Yes
Outcome Mean	0.00	0.00	0.36	6.32	0.15	0.24
R-Squared	0.13	0.15	0.12	0.17	0.07	0.42
N	8605	8605	8605	8605	8605	8605

Notes: TPA = Third Party Assessment of candidate quality (see section 4.3.2). All regressions include workshop fixed effects. Controls include candidate characteristics, interviewer characteristics and interview order fixed effects. $^{*}p < 0.10$,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=102).

increase in the previous candidate's TPA lowers a candidate's own rank in the interview sequence by 0.25 ranks on average (5%). Column 5 reports the effect on the probability to be ranked as the best candidate by an interviewer, i.e., to receive the highest relative rating in the interview sequence. Results show that a one standard deviation increase in the previous candidate's TPA decreases the probability to receive the best rating by about 1.5 percentage points (10%).

Finally, column 6 reports the effect on the committee's admission decision, which is based on the unweighted sum of all three ratings of a candidate.[28] The probability of admission decreases by 1.5 percentage points (7%) in response to a one standard deviation increase in the prior candidate's TPA, and by 1 percentage points in response to a one standard deviation increase in the leave-one-out TPA.[29]

Appendix Tables 4.D.1 to 4.D.4 show that the results presented in Table 4.6 are robust to alternative measures of quality,[30] to the inclusion of candidate fixed effects and interviewer fixed effects, and to the exclusion of candidates just below or above the admission cutoff.[31] Furthermore, Appendix Figure 4.D.1 plots estimates based on local linear regressions. They reveal a fairly linear effect pattern.

28. More precisely, a candidate is admitted when she receives in total 23 points and at least two yes votes, see also section 4.2.

29. Note that the influence of own TPA is dis-proportionally large in column 6 because TPA is the sum of the other two interviewers' ratings, which directly enter the admission decision.

30. When using a predicted quality measure based on GPA, age and study field, all coefficients are attenuated because the measure has low predictory power for the actual rating.

31. Ratings of candidates at the admission cutoff might have been adapted during the final committee meeting (see section 4.2).

Linear Autocorrelation. Table 4.7 reports the continuous and binary autocorrelation in interviewer assessments, based on equation 4.3. As above, column 1 includes only workshop fixed effects, column 2 adds further controls. The first two columns show that a one standard deviation increase in the rating given to the previous candidate is associated with a 6% standard deviation lower rating on average. Column 3 quantifies the autocorrelation in binary terms, which turns out to be substantial. If the previous candidate received a yes vote, the probability for the current candidate to receive a yes vote is 7.6 percentage points (about 20%) lower. Candidates who follow a candidate with a yes vote move down 0.4 ranks in the interviewer's ranking and are 4 percentage points less likely to receive the highest rating in the sequence (columns 4 and 5). As shown in column 6, the probability of admission is 3.7 percentage points (15%) lower.

In all columns (except for the ranking outcomes in columns 4 and 5), the interviewer's leave-one-out mean rating shows a positive coefficient, which reflects the role of interviewer leniency. Conditional on the leave-one-out mean rating, the leave-one-out share of yes votes shows a negative coefficient. The individual likelihood to receive a yes vote is thus lowered if the interviewer gives more yes votes to the other candidates.

Appendix Table 4.D.5 shows that the estimated autocorrelations are robust to the inclusion of candidate fixed effects. In line with the prediction of a downward bias that arises when estimating auto-regressive models on a finite panel (Nickell, 1981), coefficients become more negative when we control for interviewer leniency using interviewer fixed effects instead of leave-out-means (Table 4.D.6). Moreover, Appendix 4.D.2 provides additional analyses on the autocorrelation. Appendix Figure 4.D.2 documents that there is no significant autocorrelation in interviewer assessments beyond t-1. Table 4.D.7 shows that the probability of a yes vote does not decrease additionally in cases where an interviewer gives a yes vote to the candidates in both t-1 and t-2. Hence, decision 'streaks' do not reinforce the autocorrelation. In Appendix 4.D.2.1, we also test for heterogeneity of the autocorrelation with respect to candidate, interviewer and interview slot characteristics. Estimates show no evidence that ratings by more experienced interviewers are less auto-correlated. More generally, the autocorrelation shows a strikingly low amount of heterogeneity in interviewer and candidate characteristics. One exception is candidate gender, where the autocorrelation is twice as strong for females than for males. Section 4.7 will investigate the role of gender in more detail. With respect to interview slot characteristics, we find the autoccorrelation to be lower on the second interview day and if there was a break before the interview took place.

Non-Linear Autocorrelation. To test for non-linearity in the autocorrelation in ratings, Figure 4.3 plots the expected rating (panel a) and the probability of a yes vote (panel b) as a function of the previous candidate's rating in points, conditional on the control variables included in equation 4.3. The figures document that the

Table 4.7. Autocorrelation in Interviewer Assessments

	Rating (Std.)		P(Yes Vote)	Rank	P(Best)	P(Admission)
	(1)	(2)	(3)	(4)	(5)	(6)
Rating (t-1) (std.)	-0.057***	-0.063***				
	(0.010)	(0.010)				
Yes (t-1)			-0.076***	-0.422***	-0.040***	-0.037***
			(0.010)	(0.065)	(0.007)	(0.007)
Leave-one-out Mean Rating	0.254***	0.263***	0.061***	-0.682***	-0.035***	0.037***
	(0.033)	(0.032)	(0.015)	(0.092)	(0.009)	(0.008)
Leave-one-out Share Yes	-1.067***	-0.933***	-0.385***	-3.590***	-0.253***	-0.238***
	(0.139)	(0.128)	(0.092)	(0.427)	(0.044)	(0.042)
Controls	No	Yes	Yes	Yes	Yes	Yes
Outcome Mean	0.00	0.00	0.36	6.32	0.15	0.24
R-Squared	0.01	0.15	0.13	0.23	0.09	0.42
N	8605	8605	8605	8605	8605	8605

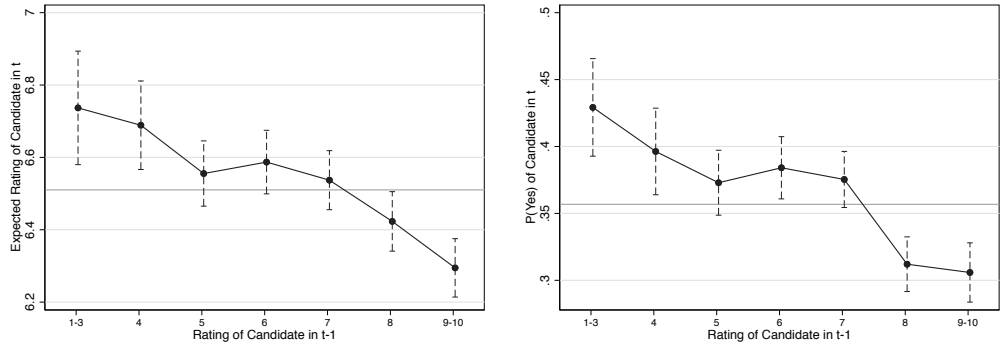
Notes: All regressions include workshop fixed effects. Controls include candidate characteristics (including TPA) interviewer characteristics and interview order fixed effects. 0.10,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=102).

autocorrelation operates mostly at the lower and the higher ends of the distribution of the previous candidate's ratings. If the interviewer rates the previous candidate with 4 points or less, the candidate in t receives an assessment above the average. The assessment of the candidate in t stays constant when the previous candidate's rating moves from 5 to 6 or 7 and close to the average. It, however, decreases further when the previous candidate receives a rating of 8 points or above. In particular, panel (b) shows a sharp jump at the yes-no cutoff: if the previous candidate received 7 (=weak no) instead of 8 (=weak yes) points, the probability that the candidate in t receives a yes vote is 6.5 percentage points higher.

4.5.3 Interaction between Candidate Quality in t and $t-1$

We now analyze how prior and current candidate quality interact. More specifically, we test if the previous candidate has a stronger or weaker influence when the two subsequent candidates are in the same range of quality, as measured through the third party assessment (TPA).

Figure 4.4 illustrates how the effect pattern differs between current candidates of high versus low TPA. The left panels show how the likelihood that a below-median TPA candidate receives a yes vote varies with the quartiles of the prior candidate's TPA, (panel a), and with the prior candidate's interview rating in points (panel c). Both panels show that changes in the current candidate's outcome operate mostly at the lower part of the prior candidate's quality distribution: when the previous candidate's TPA is in the third instead of the first quartile, the probability of a yes



(a) Rating **(b)** P(Yes)

Figure 4.3. Non-Linear Autocorrelation in Interviewer Ratings

Notes: The figures plot margins based on estimates of equation 4.3, controlling for workshop fixed effects, the interviewer's leave-one-out assessments, interviewer and candidate characteristics and interview order fixed effects. Ratings of 8 points and above imply a yes vote. The gray vertical line shows the outcome average. Dashed lines show 95% confidence intervals, with standard errors clustered at the workshop level. N=8605.

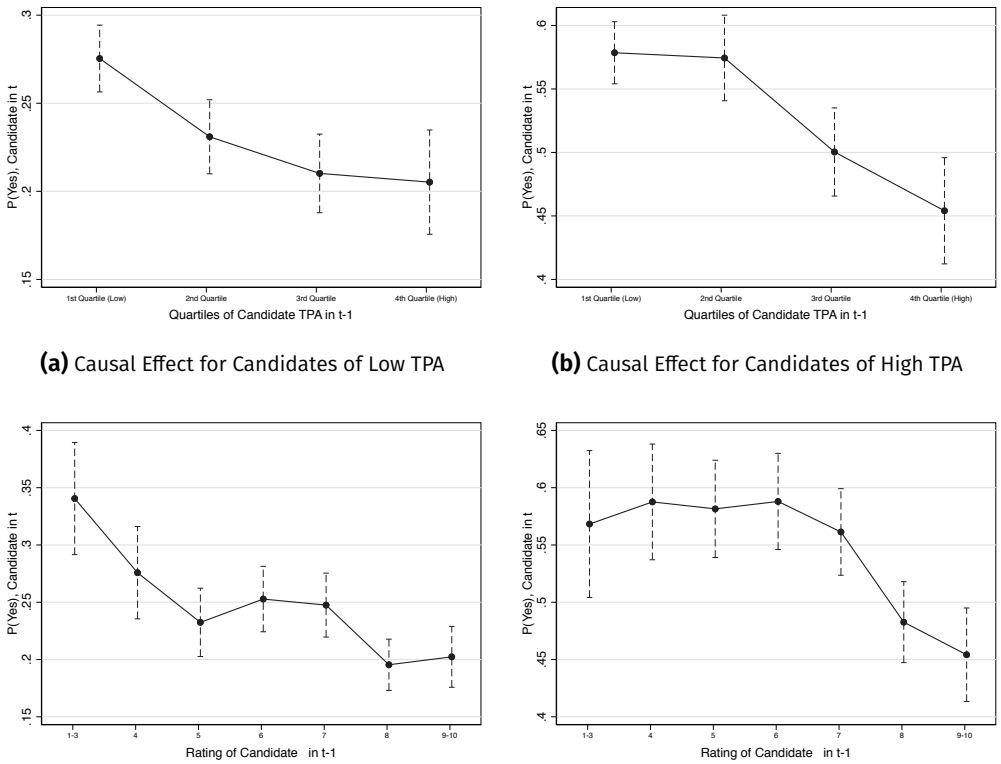
vote decreases by almost 30%, from about 0.28 to about 0.20 (panel a). In turn, an additional increase in the TPA of the previous candidate from the third to the fourth quartile has no effect. In line with this pattern, the autocorrelation also operates most strongly in lower parts of the previous candidate's rating distribution (panel c).

This pattern reverts for candidates above-median TPA: for these candidates, it does not make a difference if the previous candidate is in the first or the second quartile of the TPA distribution (panel b). In turn, the probability of a yes vote decreases by 20% from about 0.57 to about 0.45 when the previous candidate's TPA increases from the second to the fourth quartile. Similarly, panel (d) shows that changes in the prior candidate's rating only matter if they occur in the upper part of the rating distribution.

The pattern suggests that a marginal increase in the previous candidate's quality has a stronger effect if the current candidate is of similar quality. An alternative way to depict this phenomenon is to look at the difference in TPA between two subsequent candidates.[32] Results are shown in Figure 4.5. The x-axis denotes the categorical difference in points between the current and the previous candidates' TPAs; the y-axis reports the corresponding probability of a yes vote. The underlying regression includes dummies for each possible value of own TPA.

The figure reveals two main findings: first, it documents a striking asymmetry. Compared to the situation, where candidates do not differ in their TPA, candidates

32. This exercise is not possible for the autocorrelation, as the rating in t is the outcome and can therefore not be used to calculate a difference in perceptions.



(a) Causal Effect for Candidates of Low TPA

(b) Causal Effect for Candidates of High TPA

(c) Autocorrelation for Candidates of Low TPA

(d) Autocorrelation for Candidates of High TPA

Figure 4.4. Influence of the Previous Candidate, by Current Candidate's TPA

Notes: "Low TPA": third party assessment of quality $\leq$ median. "High TPA": third party assessment of quality $>$ median. Estimates result from two way-interacted regression models. The regression underlying panels a and b controls for workshop fixed effects, the leave-one-out TPA at the interviewer level, candidate characteristics (including TPA), interviewer characteristics and interview order fixed effects. The regression underlying panels c and d controls for workshop fixed effects, the interviewer's leave-one-out assessments, candidate characteristics (including TPA), interviewer characteristics and interview order fixed effects. 95% confidence intervals, with standard errors clustered at the workshop level.

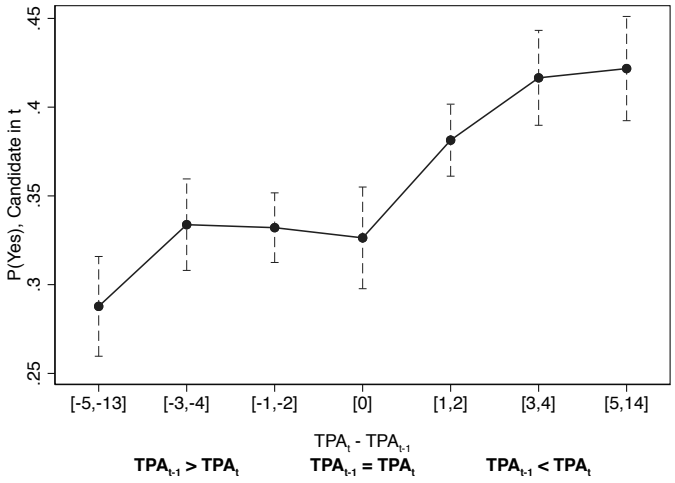


Figure 4.5. The Influence of Relative Quality

Notes: The graphs shows on the x-axis the difference in TPA between the candidate in t and the candidate $t-1$. The y-axis shows margins of the probability to receive a yes vote for the candidate in t. The underlying regression includes dummies for the candidate's own TPA. Further controls are the leave-one-out mean TPA, candidate characteristics, interviewer characteristics and order fixed effects.

strongly benefit from being slightly better than the previous candidate. However, there is no negative effect of being (slightly) worse than the previous candidate. Appendix table 4.D.11 shows the coefficients and formal test corresponding to this asymmetry. Second, and in line with the previous pictures, the effect of a positive difference in TPA shows a marginally decreasing pattern. The probability of a yes vote strongly increases at the margin of being slightly better (1-2 points higher TPA as previous candidate), but does not react strongly to additional increases in quality. Overall, the pattern suggests that an over-rating of small positive differences between subsequent candidates drive the influence of the previous candidate.

4.6 Discussion of Potential Mechanisms

This section discusses potential mechanisms underlying the influence of the other candidates, and of the previous candidate in particular. To fix ideas, we set up an illustrative theoretical framework of a decision making environment with signal extraction, which we more formally present in Appendix 4.E. We do not argue that any particular model is able to explain our results, but illustrate that the overall effect of the other candidates' quality can easily be rationalized by a model where an interviewer can learn about the distribution of quality, whereas the additional influence of the previous candidate can not.

In the framework, a rational risk-neutral interviewer votes on the admission of a closed sequence of candidates. The interviewer's aim is to accept candidates whose

quality exceeds a threshold. The interviewer forms beliefs about each candidate's quality based on noisy signals. Moreover, she infers the average of the quality distribution trough the observed signals. This average determines the interviewer's beliefs about the quality threshold. To allow the rating of a candidate interviewed in period t to be influenced by the quality of candidates interviewed both before and after t, signals are received sequentially, but belief updating and decision-making occur at the end of the sequence. This is a key difference to sequential decision making models, where updating and decision-making occur after each period.

In the appendix, we lay out that such a framework leads to a decision rule where a candidate is accepted if her signal exceeds a certain threshold. The threshold depends on the average quality of the other candidates observed by the same interviewer. Therefore, the framework can rationalize that the average quality of the other candidates has an influence. As this influence should not depend on the timing of another candidate's interview, the framework cannot rationalize why the previous candidate has an additional influence.[33] In the following, we discuss alternative mechanisms underlying this additional influence of the previous candidate.

Sequential (Bayesian) Updating. We first consider sequential updating about candidate quality, where interviewers form ratings immediately after observing each candidate. Under sequential updating, prior candidates of high quality increase the belief about the average quality and can therefore decrease the assessment of subsequent candidates. While this mechanism would produce a negative autocorrelation in ratings, it is unlikely to explain the additional influence of the previous candidate. First, candidates observed before and candidates observed afterwards matter similarly, which is not in line with immediate sequential updating (cf. Figure 4.2). Second, the ordering of prior candidates should be irrelevant for sequential updating. The quality of the previous candidate should not matter more than the quality of those candidates observed in other preceding periods. This is also not in line with the pattern presented in Figure 4.2, where only the previous candidate showed an additional influence.

Law of Small Numbers and Gambler's Fallacy. The belief in the law of small numbers (or representativeness heuristic) states that individuals erroneously believe small samples to be representative of the population. It is for example modeled via the belief that signals are not i.i.d., but drawn from an urn without replacement (e.g., Rabin, 2002; Benjamina, 2019). An immediate implication is the gambler's fallacy, which expresses the mistaken belief that a 'good draw' should follow a 'bad draw' and vice versa. Under the gambler's fallacy, interviewers underestimate the

33. Another argument which would produce an influence of the other candidates' average quality on the individual rating is a quota. The institutional setting excludes an explicit quota effect (cf. section 4.3). However, interviewers could still behave according to an implicit quota. This would equally produce an influence of the other candidates, but not an additional influence of the previous candidate.

probability that two candidates of similar quality follow each other. Therefore, they hold downward (upward) biased priors about the next candidate's quality after observing a strong (weak) candidate, which can produce a negative autocorrelation in assessments.

Several arguments rule out a major role of the gambler's fallacy to explain the additional influence of the previous candidate.[34] First, signals are received sequentially, but decisions, and thus the updating process, occur at the end of the sequence. This institutional feature is supported by the empirical result that subsequent candidates influence the assessment of prior candidates (cf. Figure 4.2). This motivates a model as presented in Appendix 4.E, where the prior belief on the quality of each candidate is the posterior mean of the distribution – without any particular role for the previous candidate.

Moreover, several predictions of the gambler's fallacy are not in line with our empirical findings on the previous candidate's influence. Most importantly, the gambler's fallacy works purely through the prior belief about the upcoming candidate's quality. This belief cannot take into account the upcoming candidate's actual quality. Therefore, relative quality differences between two subsequent candidates should not matter if interviewers would act under the gambler's fallacy. In opposition to this prediction, section 4.5.3 revealed that low (high) quality candidates are more affected by low (high) quality candidates. This empirical relevance of relative quality is not in line with a purely prior based explanation, but favors a perception-based argument. In addition, the gambler's fallacy predicts streaks of assessments to matter: having given two yes votes in a row should decrease the prior about the upcoming candidate more than having given one no vote and one yes vote. We find no evidence in this direction (see Appendix Table 4.D.7).

Sequential Contrast Effect. Instead, we argue that the results are most in line with the notion of a sequential contrast effect. Under the contrast effect, the perceived signal of a candidate interviewed in t is influenced by the previous candidate's perceived quality. We formalize this notion in Appendix 4.E.1.

The American Psychology Association defines a contrast effect as "the perception of an intensified or heightened difference between two stimuli or sensations when they are juxtaposed or when one immediately follows the other." (VandenBos, 2007). In our context, this means that contrast effects predict the perception of a current candidate to negatively depend on the perception of the previous candidate. However, as it is an unconscious bias, the interviewer believes that the perceived (biased) signal is entirely attributable to the candidate herself. Therefore, the interviewer infers candidate quality based on the biased perception.

34. We cannot exclude that the belief in the small numbers explains the influence of the other candidates' average quality. Instead of using the other signals to learn about the distribution, the interviewer would believe the other signals to be representative of the underlying distribution.

A recent contribution by Bordalo, Gennaioli, and Shleifer (2019b) provides a rationale why only the previous candidate's signal acts as an anchor. Their model offers an psychological foundation for contrast effects based on limited memory and salience of differences, i.e. comparison of current quality to a quality norm or anchor. Contrast effects arise as a current quality is compared to norm quality, which is formed by recalling similar past experiences. In our case, it is possible that the quality of the current candidate is contrasted against the previously observed signal, which is recalled due to similarity and –probably more important– proximity in time.

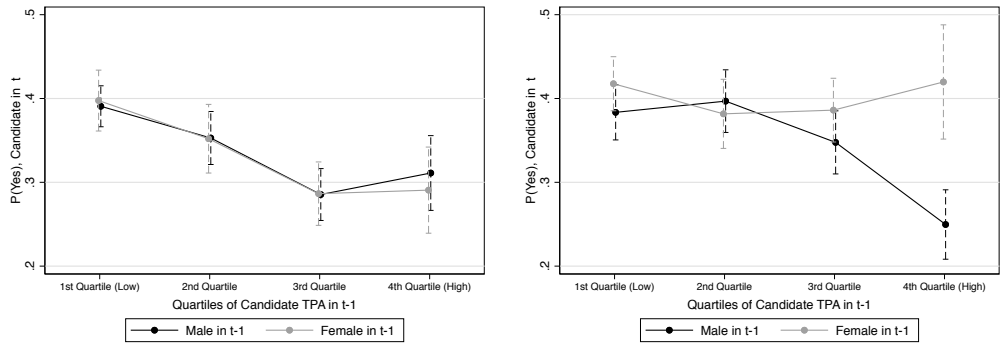
Our evidence is in line with the sequential contrast effect, where the previous candidate's signal influences the perception of the current candidate. Importantly, the results point to a non-linear version of the contrast effect, which is concentrated at the margin of being just better or worse than the previous candidate (see Figure 4.5).

4.7 The Role of Gender

In this section, we investigate how the influence of the previous candidate interacts with the sequencing of gender. We first study whether male and female candidates are differently affected by the quality of previous male versus female candidates. In a second step, we analyze to what extent gender asymmetric responses to the previous candidate interact with the gender gap in assessments. This question is of relevance because gender gaps in subjective assessments are commonly observed across many labor market settings (e.g., Neumark, Bank, and Van Nort, 1996; Rouse and Goldin, 2000; Mengel, Sauermann, and Zölitz, 2018). In the setting at hand, male candidates are about 10% more likely to receive a yes vote than female candidates (see table 4.8).

4.7.1 Influence of the Previous Candidate and the Sequencing of Gender

Figure 4.6 shows how the effect of the previous candidate's third party assessment (TPA) differs between gender sequences. Panel (a) plots the predicted probability of female candidates to receive a yes vote as a function of the previous candidate's TPA quartile, interacted with that candidate's gender. The figure shows that male and female candidates in t-1 have the same influence on the outcome of female candidates in t: the probability of a yes vote decreases from about 0.4 if the previous candidate ranks in the lowest quartile, to about 0.3 if the previous candidate ranks in the highest quartile. For male candidates (panel b), the pattern looks similar in cases where the previous candidate is male: the probability of a yes vote decreases from 0.4 if the previous male ranks in the lower two quartiles to 0.25 if he ranks in the highest quartile. On the contrary, there is no response to the previous candidate's TPA if the previous candidate is female: male candidates following a female candidate receive a yes vote with a probability of approximately 0.4, for any level of the prior female's



(a) Female Candidate in t **(b)** Male Candidate in t

Figure 4.6. Interaction between Prior Candidate Quality and the Gender Sequence

Notes: Estimates in panels a and b result from the same twoway-interacted regression model. 95% confidence intervals.

TPA. On average, male candidates who follow a female are therefore better off.[35] Appendix Tables 4.F.1 and 4.F.2 shows the corresponding linear regression results and the linear autocorrelation by gender sequence, which reveal the same pattern.

4.7.2 Implications for the Gender Gap in Assessments

The previous analysis has shown that the assessment of a male candidate does not decrease in the previous candidate's quality if the previous candidate is female. As a result, male candidates interviewed after a female are the only ones who do on average not suffer from following a strong candidate. As previous candidate quality is quasi-randomly assigned, this implies that male candidates are on average evaluated better than females candidates. We now analyze to which degree these results contribute to the observed gender gap in assessments.

We start by documenting the gender gap in average ratings and the probability of a yes vote in Table 4.8. Columns (1) and (3) show that males receive on average better assessments than females: the average rating of a male candidate is on average 6% of a standard deviation higher and male candidate are 3.4 percentage points more likely to receive a yes vote (10% relative to the female mean). Columns (2) and (4) decompose these gaps by the gender of the previous candidate. Results show that the previously observed difference in evaluations is driven almost entirely by male candidates following a female candidate. These male candidates receive on average 11% of a standard deviation higher ratings than females, while the same gap amounts to insignificant 1.6% of a standard deviation for males who follow a

35. Appendix Figure 4.F.1 shows that the pattern holds when computing TPA quartiles within gender.

male (column 2). When considering the probability of a yes vote (column 4), male candidates following a female are 5.9 percentage points (approximately 17%) more likely to receive a yes vote than female candidates. Male candidates following a male candidate are insignificant 0.9 percentage points (approximately 2.5%) more likely to receive a yes vote. This result also implies that males who follow a female receive significantly better assessments than males who follow a male. Appendix Table 4.F.3 reveals that this result is equally driven by male and female interviewers.

When considering the final admission decision (columns 5 and 6), males are on average 5.4 percentage points (about 20%) more likely to be admitted. As reported in column (6), the gap is 2 percentage points smaller for males following males compared to males following females. This difference is, however, no longer significant, as the other two assessments (and their respective gender gap) blur the effect on the single rating.

Table 4.8. Previous Candidate's Gender and the Gender Gap in Assessments

	Rating (Std.)		P(Yes Vote)		P(Admission)	
	(1)	(2)	(3)	(4)	(5)	(6)
Male (t)	0.062**		0.033***		0.054***	
	(0.026)		(0.012)		(0.013)	
Male (t) x Male (t-1)		0.016		0.009		0.044***
		(0.031)		(0.015)		(0.015)
Male (t) x Female (t-1)		0.108***		0.059***		0.064***
		(0.031)		(0.014)		(0.015)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
p-value: coeff equality		0.01		0.00		0.21
Outcome Mean	0.00	0.00	0.36	0.36	0.24	0.24
N	8605	8605	8605	8605	8605	8605

Notes: All regressions include workshop fixed effects. To avoid that covariates include variables that depend on gender, additional controls include only interviewer characteristics and order fixed effects. Standard errors are clustered at the workshop level (N=102). *0.10,** $p < 0.05$,*** $p < 0.01$.

4.8 Quantification & Implications

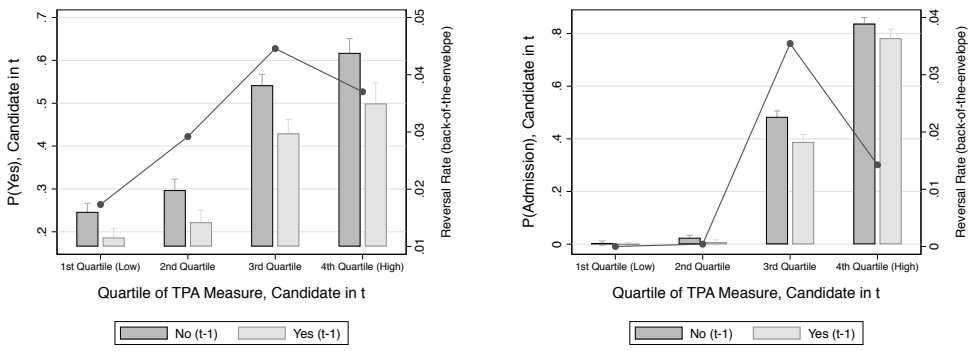
In a final step, we provide a back-of-the envelope quantification on the reversal of admission outcomes induced by the autocorrelation. The reversal rate tries to capture the amount of ratings and admission decisions which are reversed due to the autocorrelation in ratings. To compute the rate of reversed decisions, we follow the approach by Chen, Moskowitz, and Shue (2016) (see also Appendix 4.G for details).

Intuitively, the reversal rate varies over the distribution of candidate quality. The outcomes of very weak or very strong candidates are less likely to be reverted by the autocorrelation than the outcomes of candidates with more ambiguous admission prospects. We therefore calculate a separate reversal rate for each quartile of TPA.

Figure 4.7 illustrates the reversal pattern. In Panel (a), the outcome is the interviewer's yes vote. The share of reversals is about 2% for candidates from the lowest quartile and 3% for candidates from the second quartile. Candidates from the third quartile, who are in expectation at the margin of receiving a yes-vote, have a reversal rate of about 4.5%. The rate lowers back to about 3.5% for candidates in the highest quartile. The average reversal rate of yes votes is 3.5%.

When we consider the admission outcome (Panel b), the pattern looks similar. Here, the reversal rate is 0 for candidates from the two lowest quartiles. This is partly mechanical, given that the TPA is based on the other two interviewers' ratings. Candidates who receive bad ratings from the other two interviewers have close to zero chances of being admitted and the autocorrelation does not change this. Candidates in the third quartile, however, are at the margin to admission. As a result, their overall admission outcome is most affected. About 3.5% of these candidates would obtain a different outcome in absence of the autocorrelation. The same share is only about 1.5% for candidates in the highest quartile. The average reversal rate of admission decisions is 1.3%.

Appendix Figure 4.G.1 shows the pattern separately for female and male candidates. In line with our previous results that females are more affected by the autocorrelation, it shows that this effects translates into higher reversals. It also illustrates that the reversal rate of yes votes increases up to 7% for females who are at the margin of being admitted. In this group, 5.5% of admission decisions are reverted.



(a) Yes Vote **(b)** Admission

Figure 4.7. Influence on Admission Outcomes by Candidate Quality

Notes: Dashed lines show 95% confidence intervals.

4.9 Conclusion

This paper provides evidence on how a candidate's assessment outcome is influenced by the other candidates observed by the same interviewer. In line with a framework of Bayesian learning, we find the individual rating decreases in the average quality of the other candidates observed by the same interviewer. The previous candidate, however has a strong additional influence of the prior candidate. Our empirical results are most in line with the presence of a contrast effect in the perception of candidates. Further, we provide evidence that this influence is asymmetric with respect to gender: males are not affected by strong preceding female candidates and additionally receive higher ratings on average. This effect gives rise to a gender gap in assessments.

The findings in this paper help understanding how people make subjective assessments of individuals in the presence of others. They show that minor changes in candidate sorting and ordering can have major consequences on human capital formation.

This evidence carries two straightforward implications for the design of processes through which assessments are reached. First, the results document that it is crucial to minimize the overlap in the set and, importantly, in the ordering of candidates seen by the different evaluators. Second, the combination of subjective assessments with more objective screening devices, such as algorithm-based job testing technologies, might reduce the influence of human errors (e.g., Autor and Scarborough, 2008; Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan, 2017; Hoffman, Kahn, and Li, 2018). Up to now, it is, however, unclear how well these technologies perform when selecting from a high-ability segment of candidates. Ultimately, the determinants and properties of candidate selection under different screening technologies is an open research question.

Appendix 4.A Additional Material: Institutional Setting

4.A.1 Study Grant Program

Candidates at the admission workshops apply for a large merit-based study grant program in Germany. The program is prestigious and has a strong reputation for being highly competitive. It is administered by a foundation and mostly financed by the German ministry of education. Students in the program receive — at the time of our sample period — a lump-sum payment of at least 150 euros per month. Recipients can additionally receive up to 670 euros per month, depending on their parents' earnings.[1] Additional financial support is offered when spending a semester abroad. In addition, the program offers a large, cost-free course program including language classes abroad, summer schools and academic workshops. Finally, its benefits include substantial networking opportunities and a high signaling value. As a consequence of these financial and career-related benefits, the stakes for being accepted into the program are high.

The program offers several admission channels. Apart from being nominated by a high school principal, candidates can qualify for participation in an admission workshop by passing a written test or by being nominated by their university. In this paper, we concentrate on the admission channel of high-school graduates, who are nominated by their school principal. First, it constitutes the most important channel and comprised in 2012/2013 around 60% of all program admissions. Second, candidates who participate at later stages of their university studies are not as good as randomly matched to interviewers, but assigned according to their study major.

1. All German students are eligible for financial aid up to 670 euros per month, dependent on their parents' earnings. However, payments have to be repaid after graduation by students who are not receiving a merit-based scholarship.

4.A.2 Workshop Schedule

	Duration (minutes)	Type	Interviewer							
			A	B	C	D	E	F	G	H
Day 1	30	Group	1	7	13	19	25	31	37	43
	35	Interview 1	9	15	21	27	33	39	45	3
	35	Interview 1	46	4	10	16	22	28	34	40
	20	**Break**								
	30	Group	2	8	14	20	26	32	38	44
	35	Interview 1	35	41	47	5	11	17	23	29
	35	Interview 1	24	30	36	42	48	6	12	18
	60	**Lunch**								
	30	Group	3	9	15	21	27	33	39	45
	35	Interview 1	31	37	43	1	7	13	19	25
	30	Group	4	10	16	22	28	34	40	46
	20	**Break**								
	35	Interview 1	20	26	32	38	44	2	8	14
	30	Group	5	11	17	23	29	35	41	47
Day 2	35	Interview 2	43	1	7	13	19	25	31	37
	35	Interview 2	38	44	2	8	14	20	26	32
	20	**Break**								
	35	Interview 2	33	39	45	3	9	15	21	27
	30	Group	6	12	18	24	30	36	42	48
	35	Interview 2	28	34	40	46	4	10	16	22
	60	**Lunch**								
	35	Interview 2	23	29	35	41	47	5	11	17
	35	Interview 2	18	24	30	36	42	48	6	12

Figure 4.A.1. Illustration of Schedule

Notes: The time table shows which candidate was interviewed by interviewer A-H at the respective time slot. Candidates are identified by a running ID from 1-48. Group rounds consist of all candidates, who appear once in the group round for an interviewer. Each candidate has to present in his or her respective group time slot. Interviews are 35 minutes + 5 minutes break.

Appendix 4.B Additional Material: Data and Measurement

In this section, we provide additional material on the measurement of quality and randomization checks. Table 4.B.1 shows the influence interviewer characteristics on ratings. It provides evidence that an interviewer's characteristics only influence her own rating of a candidate, and does not have any spillovers on the ratings made by the other two interviewers of the same candidate. Table 4.B.2 presents results of an regression of individual ratings on candidate characteristics. Table 4.B.3 shows that there is no indication of systematic sorting to interviewers.

Table 4.B.1. Influence of Interviewer Characteristics on Assessments

	Rating (Std.) Interviewer 1	Rating (Std.) Interviewer 2	Rating (Std.) Interviewer 3
	(1)	(2)	(3)
Age (Interviewer 1)	0.003*	-0.001	0.002
	(0.002)	(0.002)	(0.002)
Female (Interviewer 1)	0.065**	-0.024	0.040
	(0.028)	(0.030)	(0.030)
Experience (Interviewer 1)	-0.022***	-0.005	-0.005
	(0.005)	(0.005)	(0.005)
Age (Interviewer 2)	0.002	0.007***	-0.002
	(0.002)	(0.002)	(0.002)
Female (Interviewer 2)	0.005	0.091***	0.021
	(0.030)	(0.032)	(0.029)
Experience (Interviewer 2)	-0.007	-0.037***	0.003
	(0.005)	(0.006)	(0.005)
Age (Interviewer 3)	-0.001	-0.001	0.005**
	(0.002)	(0.002)	(0.002)
Female (Interviewer 3)	0.020	0.032	0.030
	(0.032)	(0.032)	(0.035)
Experience (Interviewer 3)	0.006	-0.007	-0.024***
	(0.005)	(0.005)	(0.006)
p-value (joint significance int. 1)	0.00	0.59	0.31
p-value (joint significance int. 2)	0.50	0.00	0.52
p-value (joint significance int. 3)	0.69	0.15	0.00
N	4709	4710	4709

Notes: Experience is a continuous variable of prior workshop participations by an interviewer. All regressions include workshop fixed effects. $^{*}p < 0.10$,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=102).

Table 4.B.2. Influence of Candidate Covariates on Assessments

	Rating (Std.) (1)	Admission (2)
GPA Decile: 1	-0.170***	-0.039
	(0.051)	(0.027)
GPA Decile: 2	-0.110**	-0.006
	(0.052)	(0.027)
GPA Decile: 3	-0.052	-0.014
	(0.051)	(0.029)
GPA Decile: 4	-0.063	0.009
	(0.051)	(0.027)
GPA Decile: 6	0.037	0.010
	(0.042)	(0.026)
GPA Decile: 7	0.181***	0.087***
	(0.047)	(0.026)
GPA Decile: 8	0.145***	0.059**
	(0.049)	(0.027)
GPA Decile: 9	0.179***	0.088***
	(0.043)	(0.027)
GPA Decile: 10	0.271***	0.104***
	(0.045)	(0.027)
Female	-0.089***	-0.062***
	(0.025)	(0.013)
Age	0.075***	0.040***
	(0.012)	(0.007)
Migration Background	0.231***	0.115***
	(0.038)	(0.021)
Parents w/out Univ. Degree	-0.010	0.023*
	(0.029)	(0.013)
Major: Social Sciences	0.018	0.015
	(0.035)	(0.020)
Major: STEM	-0.110***	-0.058***
	(0.038)	(0.021)
Major: Medicine	-0.014	-0.001
	(0.035)	(0.020)
Major: Others	-0.186**	-0.102***
	(0.087)	(0.038)
Outcome Mean	-0.00	0.24
R-Squared (Within)	0.03	0.04
N	9420	4710

Notes: For study major, humanities is the baseline category. All regressions include workshop fixed effects. $^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$. Standard errors are clustered at the workshop level (N=102).

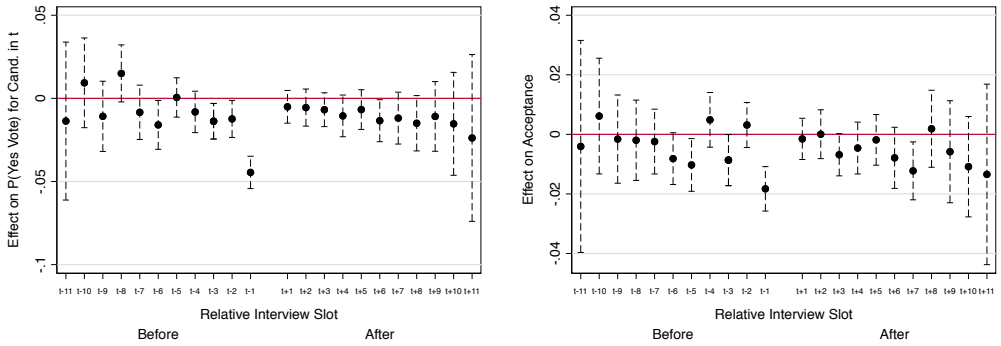
Table 4.B.3. Test of Quasi-Random Assignment to Interviewers

	Gender (1)	Age (2)	STEM (3)
Inteviewer Char.	-0.008	-0.001	-0.004
	(0.007)	(0.001)	(0.010)
Outcome Mean	0.54	19.91	0.37
N	9420	9420	9420

Notes: Regressions include workshop fixed effects. $^{*}p < 0.10$,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=102).

Appendix 4.C Additional Material: The Influence of the Interview Sequence

Figure 4.C.1 are analogous to Figure 4.2. They show estimates from regressions with alternative outcomes. Table 4.C.1 shows the coefficients and corresponding p-values for the coefficients plotted in Figures 4.2 and 4.C.1.



(a) P(Yes Vote) **(b)** P(Admission)

Figure 4.C.1. Effect of Candidate Quality in $t + k$ on Assessment of Candidate in t

Notes: The figure shows the coefficients β_k from equation 4.1, resulting from separate regressions for each value of $k = \{-11, ..., -1, 1, ..., 11\}$. The coefficients measure how the candidate in $t + k$ affects the standardized rating of the candidate in t. Dashed lines show 95% confidence intervals, with standard errors clustered at the workshop level.

Table 4.C.1. Coefficients and p-Values Corresponding to Figures 4.1 and 4.C.1

	Std. Rating			Std. Rating, Leave-One-out			P(Yes)			P(Admission)		
	Coeff.	p-value	p-value (adj.)	Coeff.	p-value	p-value (adj.)	Coeff.	p-value	p-value (adj.)	Coeff.	p-value	p-value (adj.)
t-11	-0.033	0.497	1.000	-0.028	0.580	1.000	-0.014	0.565	1.000	-0.004	0.816	1.000
t-10	0.0119	0.6222	1.0000	0.0390	0.1187	1.0000	0.0094	0.4847	1.0000	0.0061	0.5234	1.0000
t-9	-0.0372	0.0736	1.0000	0.0015	0.9461	1.0000	-0.0108	0.3483	1.0000	-0.0016	0.8322	1.0000
t-8	-0.008	0.712	1.000	0.029	0.164	1.000	0.015	0.162	1.000	-0.002	0.812	1.000
t-7	-0.0327	0.0645	1.0000	-0.0019	0.9189	1.0000	-0.0084	0.3333	1.0000	-0.0024	0.6723	1.0000
t-6	-0.0328	0.0276	0.6064	0.0005	0.9705	1.0000	-0.0159	0.0400	0.8793	-0.0082	0.0740	1.0000
t-5	-0.023	0.047	1.000	0.007	0.594	1.000	0.001	0.934	1.000	-0.010	0.033	0.729
t-4	-0.0301	0.0253	0.5565	-0.0029	0.8289	1.0000	-0.0081	0.1984	1.0000	0.0049	0.2887	1.0000
t-3	-0.0484	0.0001	0.0020	-0.0198	0.1006	1.0000	-0.0137	0.0106	0.2341	-0.0086	0.0462	1.0000
t-2	-0.033	0.012	0.265	-0.003	0.811	1.000	-0.012	0.029	0.644	0.003	0.410	1.000
t-1	-0.0942	0.0000	0.0000	-0.0702	0.0000	0.0000	-0.0445	0.0000	0.0000	-0.0183	0.0000	0.0000
t+1	-0.0197	0.0531	1.0000	0.0114	0.2574	1.0000	-0.0051	0.3055	1.0000	-0.0015	0.6625	1.0000
t+2	-0.0131	0.2240	1.0000	0.0164	0.1453	1.0000	-0.0055	0.3237	1.0000	0.0001	0.9902	1.0000
t+3	-0.0230	0.0291	0.6393	0.0053	0.6285	1.0000	-0.0068	0.1824	1.0000	-0.0068	0.0547	1.0000
t+4	-0.0259	0.0442	0.9731	0.0012	0.9248	1.0000	-0.0105	0.0943	1.0000	-0.0046	0.2935	1.0000
t+5	-0.0138	0.2417	1.0000	0.0120	0.3406	1.0000	-0.0067	0.2617	1.0000	-0.0018	0.6666	1.0000
t+6	-0.0262	0.0475	1.0000	0.0010	0.9403	1.0000	-0.0134	0.0339	0.7458	-0.0079	0.1260	1.0000
t+7	-0.0306	0.0809	1.0000	-0.0031	0.8676	1.0000	-0.0119	0.1296	1.0000	-0.0123	0.0120	0.2637
t+8	-0.0356	0.0568	1.0000	-0.0056	0.7718	1.0000	-0.0150	0.0727	1.0000	0.0019	0.7719	1.0000
t+9	0.0019	0.9211	1.0000	0.0279	0.1785	1.0000	-0.0109	0.3001	1.0000	-0.0058	0.4958	1.0000
t+10	-0.0159	0.5984	1.0000	0.0005	0.9864	1.0000	-0.0153	0.3200	1.0000	-0.0109	0.1954	1.0000
t+11	-0.0117	0.7907	1.0000	0.0017	0.9713	1.0000	-0.0238	0.3319	1.0000	-0.0134	0.3645	1.0000
Joint test		0.00			0.00			0.00			0.00	
$t-1=t+1$		0.00			0.00			0.00			0.00	
Before vs. after		0.46			0.63			0.22			0.36	

Notes: Table shows the coefficients and p-values corresponding to Figure 4.2 Panel (a)-(d). p-values are adjusted using Bonferroni. At the bottom we report tests on the joint significance and equality of $t-1$ and $t+1$. The last line reports a test on the equality of the average coefficient before and the average coefficient after the candidate.

Table 4.D.1. Robustness Checks: Alternative Quality Measures

	Std. Rating	P(Yes Vote)	Rank	P(Best)	P(Admission)
	(1)	(2)	(3)	(4)	(5)
Panel A: Qality measured through group discussion rating only					
Leave-one-out Mean Rating group (std.)	-0.041**	-0.014**	-0.175***	-0.010**	-0.001
	(0.019)	(0.007)	(0.020)	(0.004)	(0.005)
TPA (std.), t-1	-0.061***	-0.030***	-0.212***	-0.015***	-0.017***
	(0.010)	(0.005)	(0.036)	(0.004)	(0.005)
Rating Group Disc. (std.)	0.204***	0.086***	0.719***	0.049***	0.191***
	(0.011)	(0.005)	(0.034)	(0.004)	(0.006)
Panel B: TPA measured through other interview rating only					
Leave-one-out Mean Rating oth. int. (std.)	-0.084***	-0.024***	-0.290***	-0.018***	-0.009**
	(0.014)	(0.005)	(0.021)	(0.003)	(0.004)
TPA (std.), t-1	-0.047***	-0.027***	-0.177***	-0.009**	-0.008*
	(0.010)	(0.005)	(0.036)	(0.004)	(0.004)
Rating other int. (std.)	0.284***	0.128***	1.004***	0.069***	0.225***
	(0.017)	(0.007)	(0.052)	(0.005)	(0.005)
Panel C: TPA measured through prediction based on GPA, age and major					
Leave-one-out Mean TPA (std.)	-0.034**	-0.011	-0.145***	-0.015***	-0.006
	(0.017)	(0.007)	(0.022)	(0.004)	(0.006)
TPA (std.), t-1	-0.030***	-0.017***	-0.099**	-0.009**	-0.014***
	(0.011)	(0.006)	(0.038)	(0.004)	(0.005)
TPA (std.)	0.144***	0.059***	0.480***	0.027***	0.059***
	(0.013)	(0.006)	(0.041)	(0.005)	(0.006)
Outcome Mean	0.00	0.36	6.32	0.15	0.24
N	8605	8605	8605	8605	8605

Notes: All regressions include workshop fixed effects. Controls include candidate characteristics, interviewer characteristics and interview order fixed effects. $^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$. Standard errors are clustered at the workshop level (N=102).

Appendix 4.D Additional Material: Influence of the Previous Candidate

4.D.1 Additional Material for Causal Analysis

This section provides several robustness checks for the influence of the previous candidate's quality (section 4.5). Table 4.D.1 reproduces the results from Table 4.6 using alternative measures of quality. Table 4.D.2 excludes marginal candidates from the analysis. Tables 4.D.3 and 4.D.4 reports results from regressions with candidate fixed effects and interviewer fixed effects, respectively. Figure 4.D.1 plots estimates from local linear regressions.

Table 4.D.2. Robustness Checks: Exclusion of Marginal Candidates

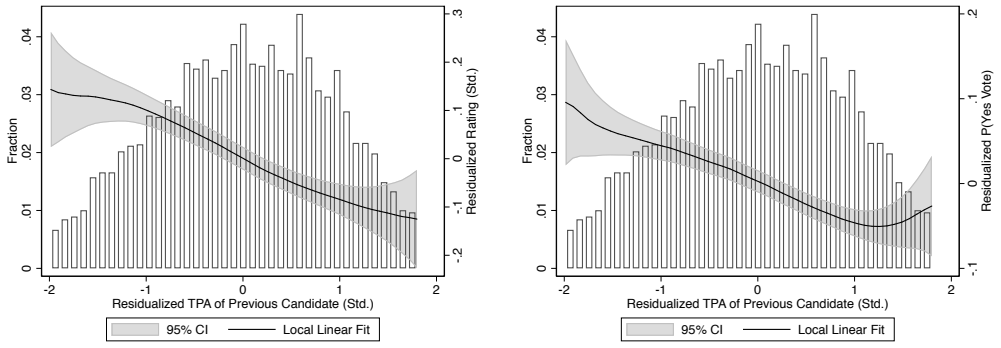
	Std. Rating	P(Yes Vote)	Rank	P(Best)	P(Admission)
	(1)	(2)	(3)	(4)	(5)
Leave-one-out Mean TPA (std.)	-0.086***	-0.025***	-0.305***	-0.020***	-0.011**
	(0.021)	(0.007)	(0.034)	(0.005)	(0.005)
TPA (std.), t-1	-0.079***	-0.040***	-0.267***	-0.016***	-0.019***
	(0.013)	(0.006)	(0.041)	(0.004)	(0.004)
TPA (std.), t	0.289***	0.127***	1.034***	0.082***	0.237***
	(0.017)	(0.006)	(0.050)	(0.005)	(0.007)
Controls	Yes	Yes	Yes	Yes	Yes
Outcome Mean	-0.12	0.30	5.98	0.13	0.16
N	6159	6159	6159	6159	6159

Notes: All regressions exclude marginal candidates in t and $t-1$ (candidates with 23 points in total). Further controls include interviewer and candidate characteristics and interview order fixed effects. TPA = Third Party Assessment of candidate quality (see section 4.3.2). $^{*}p < 0.10$,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=102).

Table 4.D.3. Robustness Checks: Estimation with Candidate Fixed Effects

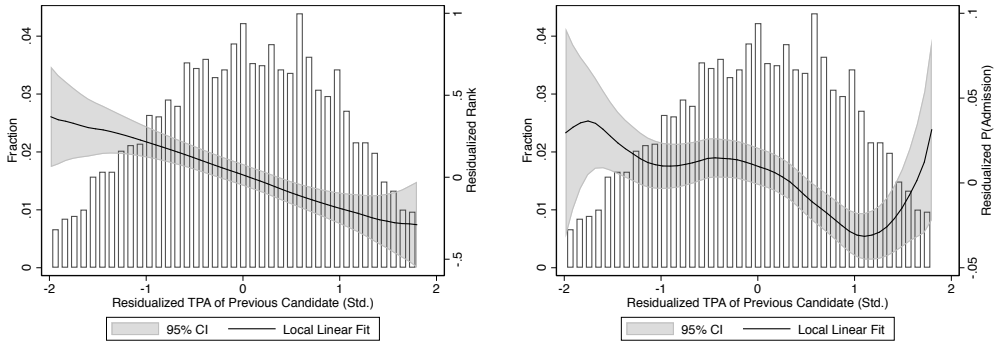
	Std. Rating	P(Yes Vote)	Rank	P(Best)
	(1)	(2)	(3)	(4)
Leave-one-out Mean TPA (std.)	-0.064***	-0.025***	-0.303***	-0.014***
	(0.020)	(0.008)	(0.036)	(0.005)
TPA (std.), t-1	-0.058***	-0.025***	-0.267***	-0.015***
	(0.014)	(0.007)	(0.045)	(0.005)
Controls	Yes	Yes	Yes	Yes
Outcome Mean	0.00	0.36	6.32	0.15
N	8605	8605	8605	8605

Notes: All regressions include candidate fixed effects. As the admission outcome does not vary on the candidate level, this outcome is omitted from the table. Further controls include interviewer characteristics and interview order fixed effects. TPA = Third Party Assessment of candidate quality (see section 4.3.2). $^{*}p < 0.10$,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=102).



(a) Rating

(b) P (Yes Vote)



(c) Rank

(d) P (Admission)

Figure 4.D.1. Local Linear Regressions

Notes: Residuals stem from regressions of the respective variable on workshop fixed effects, leave-two-out mean quality in the sequence, candidate characteristics (including quality), interviewer characteristics and interview order fixed effects. Top and bottom 2% are excluded. Shaded areas show 95% confidence intervals.

Table 4.D.4. Robustness Checks: Estimation with Interviewer Fixed Effects

	Std. Rating	P(Yes Vote)	Rank	P(Best)	P(Admission)
	(1)	(2)	(3)	(4)	(5)
TPA (std.), t-1	-0.070***	-0.038***	-0.253***	-0.016***	-0.015***
	(0.011)	(0.005)	(0.038)	(0.003)	(0.004)
TPA (std.), t	0.350***	0.150***	1.233***	0.083***	0.278***
	(0.013)	(0.006)	(0.045)	(0.004)	(0.006)
Controls	Yes	Yes	Yes	Yes	Yes
Outcome Mean	0.00	0.36	6.32	0.15	0.24
N	8605	8605	8605	8605	8605

Notes: Leave-one-out mean quality is omitted due to co-linearity with the interviewer fixed effects. All regressions include candidate characteristics and interview order fixed effects. TPA = Third Party Assessment of candidate quality (see section 4.3.2). $^{*}p < 0.10$,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=102).

4.D.2 Additional Material for Autocorrelation

This section provides several robustness checks for the autocorrelation (section 4.5). Tables 4.D.5 and 4.D.6 report results from regressions with candidate fixed effects and interviewer fixed effects, respectively. Figure 4.D.2 plots the correlation between an interviewers rating in t and her rating in $t-6, \ldots, t-2$. Table 4.D.7 tests whether streaks have an additional influence.

Table 4.D.5. Robustness Checks: Estimation with Candidate Fixed Effects

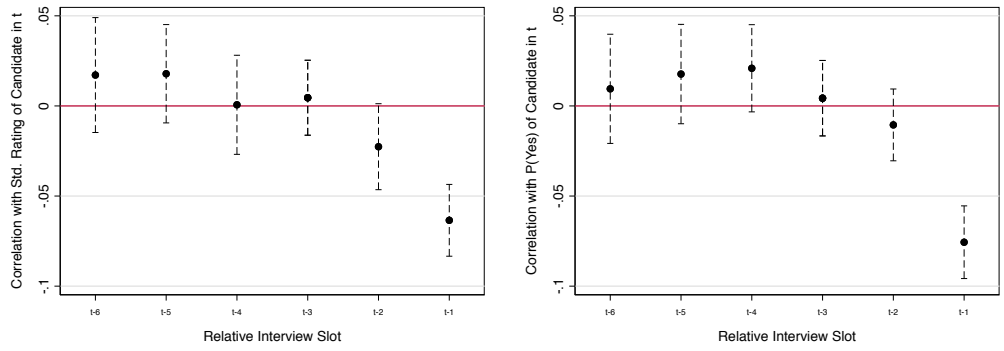
	Rating (Std.)	P(Yes Vote)	Rank	P(Best)
	(1)	(2)	(3)	(4)
Rating (t-1) (std.)	-0.060***			
	(0.013)			
Yes (t-1)		-0.073***	-0.425***	-0.041***
		(0.012)	(0.079)	(0.010)
Leave-one-out Mean Rating	0.316***	0.062***	-0.560***	-0.033***
	(0.032)	(0.015)	(0.097)	(0.010)
Leave-one-out Share Yes	-0.731***	-0.152*	-2.731***	-0.140***
	(0.127)	(0.083)	(0.440)	(0.049)
Controls	Yes	Yes	Yes	Yes
Outcome Mean	0.00	0.36	6.32	0.15
N	8605	8605	8605	8605

Notes: All regressions include candidate fixed effects. As the admission outcome does not vary on the candidate level, this outcome is omitted from the table. Further controls include interviewer characteristics and interview order fixed effects. $^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$. Standard errors are clustered at the workshop level (N=102).

Table 4.D.6. Robustness Checks: Estimation with Interviewer Fixed Effects

	Rating (Std.)	P(Yes Vote)	Rank	P(Best)	P(Admission)
	(1)	(2)	(3)	(4)	(5)
Rating (t-1) (std.)	-0.143***				
	(0.010)				
Yes (t-1)		-0.159***	-0.054**	-0.009	-0.069***
		(0.010)	(0.025)	(0.006)	(0.007)
Leave-one-out Mean Rating			-15.774***	-0.697***	
			(0.244)	(0.032)	
Leave-one-out Share Yes			-12.073***	-2.483***	
			(0.758)	(0.173)	
Controls	Yes	Yes	Yes	Yes	Yes
Outcome Mean	0.00	0.36	6.32	0.15	0.24
N	8605	8605	8605	8605	8605

Notes: All regressions control for interviewer leniency using interviewer fixed effects instead of leave-one-out mean outcomes. Further controls include candidate characteristics and interview order fixed effects. $^{*}p < 0.10$,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=102).



(a) Rating **(b)** P (Yes Vote)

Figure 4.D.2. Autocorrelation Beyond $t-1$

Notes: Each coefficient results from a separate regression, where the assessment of the candidate in t is related to the assessment of the candidate in $t-k$, $k \in \{-11, ..., -6\}$. All regressions include workshop fixed effects and the interviewer's leave-one-out mean in ratings and yes votes. Further controls include candidate characteristics (including TPA), interviewer characteristics and interview order fixed effects. 95% confidence intervals, with standard errors clustered at the workshop level.

Table 4.D.7. Test for Additional Influence of Streaks

	P(Yes Vote)	
	(1)	(2)
Yes (t-1)=1	-0.080***	
	(0.012)	
Yes (t-1) and (t-2)	0.019	
	(0.019)	
TPA High (t-1)=1		-0.072***
		(0.014)
TPA High (t-1) and (t-2)		-0.010
		(0.033)
Controls	Yes	Yes
N	8605	9420

Notes: Column 1 tests whether the probability of a yes vote changes when the interviewer gives the two preceding, instead of the one preceding candidate a yes vote. Column 2 tests whether the probability of a yes vote changes when the two preceding, instead of the one preceding candidate is in the highest TPA quartile. All regressions include workshop fixed effects, candidate characteristics, interviewer characteristics (including TPA) and interview order fixed effects. In column 1, the regression additionally includes the interviewer's leave-one-out mean rating and share of yes votes. $^{*}p < 0.10$,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=102).

4.D.2.1 Further Heterogeneity

In the following, we test for additional sources of heterogeneity. The dimensions we look at are interviewer characteristics, the time slot of the interview and candidate characteristics.

Interviewer Characteristics. We start by analyzing whether the autocorrelation interacts with the interviewer's characteristics, most importantly experience. In our context, experience is defined by the number of prior workshop participations. Experienced interviewers could be more familiar with the assessment task, and therefore more able to judge a candidate independently of the previous candidate.

Column 1 of table 4.D.8 does not support the idea that the previous candidate's influence is mitigated by experience in interviewing for the program. While the point estimate of the autocorrelation is slightly weaker for interviewers with three or more prior workshops, the difference is not significant. Having participated once or twice has no influence at all. One potential reason is that participations are too distant in time, given that interviewers tend to participate only once every one or two years. Moreover, experience in the form of pure participation (without reflection, feedback and supervision) may not mitigate the effect, as interviewers are likely aware of the autocorrelation.

Columns 2 and 3 additionally show that the autocorrelation is about 3.4 percentage points stronger for interviewers aged above the median, and does not differ between male and female interviewers.

Time Slot of Interview. We further study how the effect varies between interview time slots. Column 1 of table 4.D.9 shows that the autocorrelation is 3.4 percentage points weaker if the interview took place on the second day of the workshop. A possible interpretation is that interviewers are less influenced by the previous candidate as they have seen already more candidates overall. Another possibility is that interviewers are better able to recall second day interviews when they finalize their ratings at the end of the sequence. This might mitigate previous candidate's influence on their immediate perception.

In column 2, the autocorrelation is interacted with the incidence of a break before the interview in t. We distinguish between long breaks (more than 50 minutes) and short breaks (between 20 to 50 minutes). Our results show that the autocorrelation amounts to -10.8% if no break occurs between two interviews (i.e., a maximum of 5 minutes for interviewers to take notes). The autocorrelation strongly decreases to approximately half of the magnitude if either a short or a long break took place between the interviews. Breaks thus appear to render the influence of the previous candidate less important.

Candidate Characteristics. Furthermore, we assess whether the autocorrelation differs by observable characteristics of the candidate interviewed in period t. Table 4.D.10 reveals that there is no heterogeneity along the characteristics observed

Table 4.D.8. Heterogeneity in the Autocorrelation: Interviewer Characteristics

	P(Yes Vote)		
	(1)	(2)	(3)
Yes (t-1)	-0.078***	-0.060***	-0.078***
	(0.018)	(0.014)	(0.012)
Experience: 1 x Yes (t-1)	-0.009		
	(0.026)		
Experience: 2 x Yes (t-1)	0.001		
	(0.032)		
Experience: 3+ x Yes (t-1)	0.014		
	(0.023)		
Age > Median x Yes (t-1)		-0.034*	
		(0.020)	
Female x Yes (t-1)			0.005
			(0.021)
Controls	Yes	Yes	Yes
Outcome Mean	0.36	0.36	0.36
N	8605	8605	8605

Notes: All regressions include workshop fixed effects and control for the interviewer's leave-out mean of ratings and yes votes. The leave out-means are interacted with the dimension of heterogeneity. Controls are candidate and interviewer characteristics and interview order fixed effects. 0.10,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=102).

in the data, with the exception of gender: the autocorrelation amounts to only 4.5 percentage points for male candidates, and is 6 percentage points higher for female candidates. The next section explores the role of gender in more depth.

Table 4.D.9. Heterogeneity in the Autocorrelation: Characteristics of the Interview Slot

	P(Yes Vote)	
	(1)	(2)
Yes (t-1)	-0.095***	-0.107***
	(0.014)	(0.019)
Day 2 x Yes (t-1)	0.035*	
	(0.020)	
Short Break x Yes (t-1)		0.044
		(0.027)
Long Break x Yes (t-1)		0.050*
		(0.027)
Controls	Yes	Yes
Outcome Mean	0.36	0.36
N	8605	8605

Notes: All regressions include workshop fixed effects and control for the interviewer's leave-out mean of ratings and yes votes. The leave out-means are interacted with the dimension of heterogeneity. Controls are candidate and interviewer characteristics and interview order fixed effects. 0.10,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=102).

Table 4.D.10. Heterogeneity in the Autocorrelation: Candidate Characteristics

	P(Yes Vote)					
	(1)	(2)	(3)	(4)	(5)	(6)
Yes (t-1)	-0.045***	-0.071***	-0.078***	-0.075***	-0.082***	-0.088***
	(0.014)	(0.012)	(0.014)	(0.011)	(0.013)	(0.013)
Female x Yes (t-1)	-0.060***					
	(0.021)					
Age > Median x Yes (t-1)		-0.018				
		(0.022)				
GPA > Median x Yes (t-1)			0.005			
			(0.023)			
Migration Background x Yes (t-1)				-0.006		
				(0.030)		
Parents w/out Univ. Degree x Yes (t-1)					0.027	
					(0.025)	
Major: STEM x Yes (t-1)						0.033
						(0.024)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Outcome Mean	0.36	0.36	0.36	0.36	0.36	0.36
N	8605	8605	8605	8605	8605	8605

Notes: All regressions include workshop fixed effects and control for the interviewer's leave-out mean of ratings and yes votes. The leave out-means are interacted with the dimension of heterogeneity. Controls are candidate and interviewer characteristics and interview order fixed effects. 0.10,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=102).

4.D.3 Additional Material for Interaction between Candidate in t and $t-1$

Table 4.D.11. Effect of Being Better or Worse Than the Previous Candidate

	Std. Rating	P(Yes Vote)
	(1)	(2)
Candidate in t is better	0.157***	0.086***
	(0.032)	(0.017)
Candidate in t is worse	0.004	-0.004
	(0.036)	(0.018)
Controls	Yes	Yes
p-value (|better|=|worse|)	0.00	0.01
Outcome Mean	-0.00	0.36
N	9420	9420

Notes: Controls include leave-one-out TPA, own TPA, interviewer and candidate characteristics and interview order fixed effects. TPA = Third Party Assessment of candidate quality (see section 4.3.2). $^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$. Standard errors are clustered at the workshop level (N=102).

Appendix 4.E Conceptual Framework

We lay out a framework where a rational risk-neutral interviewer votes on the admission of a closed sequence of candidates. The interviewer's aim is to accept candidates whose quality exceeds a threshold. The interviewer forms beliefs about each candidate's quality based on noisy signals. Moreover, she infers the average of the quality distribution trough the observed signals. This average determines the interviewer's beliefs about the quality threshold. To allow the rating of a candidate interviewed in period t to be influenced by the quality of candidates interviewed both before and after t, signals are received sequentially, but decisions are made at the end of the sequence. This is a key difference to sequential decision making models, where updating and decisions occur after each period.

Setup. Suppose that a candidate observed by interviewer i at time t has quality $q_{i,t} \sim \mathcal{N}(\theta_0, \sigma_0^2)$. The risk-neutral interviewer observes a noisy signal of quality, $\tilde{q}_{i,t} = q_{i,t} + \epsilon_{i,t}$ where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

The interviewer votes on the admission decision of each observed candidate. Each decision is supposedly independent, as the interviewer does not face a quota. Therefore, the interviewer evaluates the two alternatives of voting in favor or against admitting a candidate. Admission yields a value $V_{accept} = \mathbb{E}(q_{i,t} - \underline{q}_i)$, while the value of a rejection is $V_{reject} = 0$

In the expression of V_{accept}, $\underline{q}_i$ is a predefined quality threshold. It can be expressed as $\underline{q}_i = \alpha_i * \mathbb{E}(q)$, where $\alpha_i > 1$ is an interviewer-specific term capturing, for example, differences in leniency between interviewers. The interviewer gives a yes vote to a candidate if $V_{accept} > V_{reject}$, i.e.:

$$\mathbb{E}(q_{i,t}) \; > \; \alpha_i \mathbb{E}(q) \tag{4.E.1}$$

This decision rule implies that an interviewer votes in favor of a candidate, if her posterior belief about the candidate's quality exceeds the threshold. The threshold depends on the expected quality of all candidates. Therefore, the interviewer has to form posterior beliefs about the quality of the candidate and about the average quality of all candidates.

As postulated above and in line with the institutional framework, we assume that the interviewer updates her belief about the mean quality after observing all (uncorrelated) signals $\tilde{q}_{i1}, \tilde{q}_{i2}, ..., \tilde{q}_{iT}$. Let $\overline{\tilde{q}_i}$ be the average of all signals. Following Bayes' rule, we can express the updated belief about $\mathbb{E}(q)$ (see DeGroot, 1970):

$$\theta_1 \; = \; \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_0^2 T}\theta_0 + \frac{T\sigma_0^2}{\sigma_\epsilon^2 + \sigma_0^2 T}\overline{\tilde{q}_i} \tag{4.E.2}$$

In a second step, the interviewer forms posterior beliefs about the quality of each individual candidate, given her posterior belief about the average. The belief about the quality of a candidate is thus a precision weighted average of the signal and the posterior belief about the average:

$$q_{posterior,i,t} = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_0^2}\theta_1 + \frac{\sigma_0^2}{\sigma_\epsilon^2 + \sigma_0^2}\tilde{q}_{i,t} \tag{4.E.3}$$

Decision Rule. Plugging the two posterior beliefs into equation 4.E.1 yields the following decision rule:

$$\frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_0^2}\theta_1 + \frac{\sigma_0^2}{\sigma_\epsilon^2 + \sigma_0^2}\tilde{q}_{i,t} > \alpha_i\theta_1 \tag{4.E.4}$$

The decision rule shows that an interviewer votes in favor of a candidate if her posterior belief about the quality of this candidate exceeds the threshold, which depends on the posterior of the mean quality. In this rule, a candidate's signal acts in two counteracting ways. On the one hand, it affects the posterior belief about the candidate's individual quality. On the other hand, it affects the threshold, as it increases the posterior belief about the average quality. To solve the model, we assume the first effect dominates, i.e. the threshold reacts less than the posterior belief about individual quality. Formally, this assumption is satisfied iff $\frac{2\sigma_\epsilon^2+\sigma_0^2 T}{\sigma_\epsilon^2+\sigma_0^2} > \alpha_i$. For $T \geq 2$ it is sufficient that $\alpha_i < 2$. While we cannot test this condition formally, it is plausible that this condition is fulfilled in our data, as the left hand side is increasing in T and T is relatively large in our data. We can then derive a threshold for the signal of each candidate:

$$\tilde{q}_{i,t} > [\alpha_i - \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_0^2}]\Big[\frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_0^2 T}\theta_0 + \frac{(T-1)\sigma_0^2}{\sigma_\epsilon^2 + \sigma_0^2 T}\overline{\tilde{q}_{i,-t}}\Big]$$
$$\frac{\sigma_\epsilon^2 + \sigma_0^2}{\sigma_0^2}\Big[1 - \frac{\alpha_i(\sigma_\epsilon^2 + \sigma_0^2)}{\sigma_\epsilon^2 + \sigma_0^2 T} + \frac{\sigma_\epsilon^2}{(\sigma_\epsilon^2 + \sigma_0^2 T)}\Big]^{-1}$$

A candidate is therefore accepted if her signal $\tilde{q}_{i,t}$ exceeds the threshold $\underline{q_{i,t}}(\alpha_i, \sigma_\epsilon^2, \sigma_0^2, \overline{\tilde{q}_{i,-t}}, T, \theta_0)$. The threshold increases in α_i, reflecting that interviewers with a higher leniency have lower thresholds. It further increases in the average signal, which implies that the individual probability of a yes vote decreases in the (average) signals of the other candidates observed by the same interviewer.

As a direct consequence of the threshold rule, the average quality of the other candidates observed by the same interviewer can affect the rating of the single candidate. Besides, it is easy to see that the partial derivatives of the threshold with

respect to the signals of any other candidate do not depend on the timing of a particular candidate's interview. Therefore, the threshold rule cannot rationalize why the previous candidate has a stronger (additional) impact compared to the other candidates.

4.E.1 Sequential Contrast Effect

We postulate that interviewers with a sequential contrast effect, perceive the signal to be:

$$\tilde{q}^{C}_{i,t} \;=\; r(\tilde{q}_{i,t}, \tilde{q}^{C}_{i,t-1})$$

where $r(\tilde{q}_{i,t}, \tilde{q}^{C}_{i,t-1})$ is a function that increases in $\tilde{q}_{i,t}$ and decreases in $\tilde{q}_{i,t-1}$. Among the most simple parameterizations of this negative relationship is a linear formulation, such as $\tilde{q}^{C}_{i,t} = \tilde{q}_{i,t} - \gamma(\tilde{q}^{C}_{i,t-1} - \tilde{q}_{i,t,})$, with $\gamma \in [0,1)$. [1] Here, the perception of a candidate's signal is influenced by the perception of the previous candidate's signal, at a strength captured by γ.

The biased perception of signals changes 4.E.4 to:

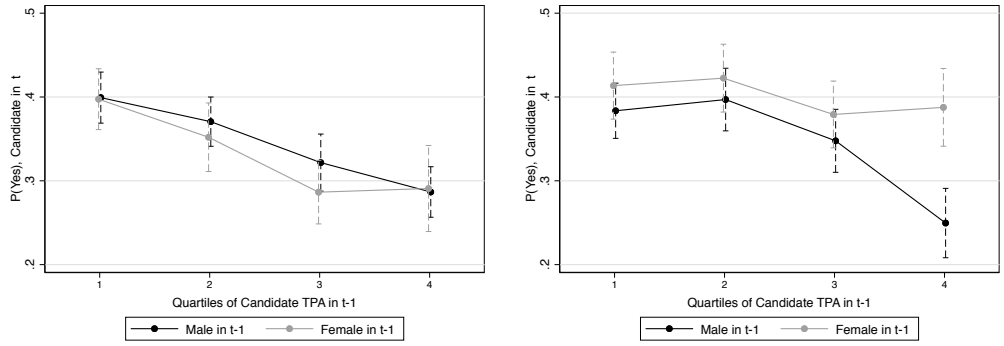
$$\tilde{q}_{i,t} * (1+\gamma) \;>\; [\alpha_i - \frac{\sigma^2_\epsilon}{\sigma^2_\epsilon + \sigma^2_0}]\Big[\frac{\sigma^2_\epsilon}{\sigma^2_\epsilon + \sigma^2_0 T}\theta_0 + \frac{(T-1)\sigma^2_0}{\sigma^2_\epsilon + \sigma^2_0 T}\overline{\tilde{q}^{C}_{i,-t}}\Big]$$
$$\frac{\sigma^2_\epsilon + \sigma^2_0}{\sigma^2_0}\Big[1 - \frac{\alpha_i(\sigma^2_\epsilon + \sigma^2_0)}{\sigma^2_\epsilon + \sigma^2_0 T} + \frac{\sigma^2_\epsilon}{(\sigma^2_\epsilon + \sigma^2_0 T)}\Big]^{-1} + \gamma\tilde{q}^{C}_{i,t-1}$$

The threshold therefore increases if any other candidates' quality increases. However, it increases more strongly with respect to the (perceived) quality of the previous candidate. More formally, the partial derivatives of the threshold with respect to perceived quality of a candidate in $t \in [1, ..., t-2, t+1, ...T]$ are equal, but lower than the partial derivative with respect to the perceived quality of the candidate in $t-1$. The threshold increases more strongly in the quality of the previous candidate, which can explain the additional influence of the previous candidate on the likelihood to receive a yes vote.

1. A more sophisticated version of the contrast effect could potentially include a non-linear specification, in line with the findings in section 4.5.3.

Appendix 4.F Additional Material: The Role of Gender

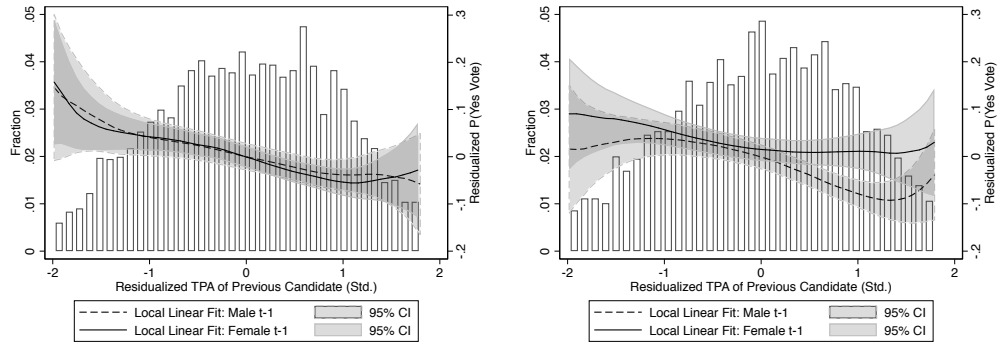
We provide robustness checks for the results presented in section 4.7. Figure 4.F.1 plots the same results as Figure 4.6, but TPA quartiles are computed within each gender respectively, to account for the overall lower ratings of females. Figure 4.F.2 provides graphical evidence from a local linear regression of residualized TPA of the previous candidates TPA on the respective outcomes. Table 4.F.1 provides linear estimates of the effect. Lastly, Table 4.F.3 provides evidence that the gender of the interviewer does not have play a role for the gender gap.



(a) Female Candidate in t **(b)** Male Candidate in t

Figure 4.F.1. Prior Candidate Quality and the Gender Sequence

Notes: Estimates in panels a and b result from the same twoway-interacted regression model. Quality quartiles are computed within gender. 95% confidence intervals.



(a) Female Candidate in t **(b)** Male Candidate in t

Figure 4.F.2. Prior Candidate Quality and the Gender Sequence

Notes: Residuals stem from regressions of the respective variable on workshop fixed effects, leave-two-out mean TPA in the sequence, candidate characteristics (including TPA), interviewer characteristics and interview order fixed effects. Top and bottom 2% are excluded. Shaded areas show 95% confidence intervals.

Table 4.F.1. Interaction between Prior Candidate Quality and the Gender Sequence: Linear Specification, Causal Effect

	Rating (Std.)	P(Yes Vote)	P(Admission)
	(1)	(2)	(3)
Male × Female (t-1)=0 × TPA (std.), t-1	-0.073***	-0.044***	-0.020***
	(0.019)	(0.010)	(0.007)
Male × Female (t-1)=1 × TPA (std.), t-1	-0.019	-0.009	-0.012
	(0.024)	(0.012)	(0.008)
Female × Female (t-1)=0 × TPA (std.), t-1	-0.071***	-0.045***	-0.015*
	(0.023)	(0.011)	(0.008)
Female × Female (t-1)=1 × TPA (std.), t-1	-0.092***	-0.042***	-0.014**
	(0.017)	(0.008)	(0.007)
Controls	Yes	Yes	Yes
p-value: Male (t) coeffs equal	0.07	0.02	0.45
p-value: Female (t) coeffs equal	0.44	0.85	0.91
N	8605	8605	8605

Notes: All regressions include workshop fixed effects. Further controls include candidate characteristics, interviewer characteristics and interview order fixed effects, as well as own TPA and leave-one-out mean TPA interacted with gender. $^{*}p < 0.10$,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=102).

Table 4.F.2. Interaction between Prior Candidate Rating and the Gender Sequence: Linear Specification, Autocorrelation

	Rating (Std.)	P(Yes Vote)	P(Admission)
	(1)	(2)	(3)
Male × Female (t-1)=0 × Rating (t-1) (std.)	-0.088***		
	(0.021)		
Male × Female (t-1)=1 × Rating (t-1) (std.)	0.021		
	(0.024)		
Female × Female (t-1)=0 × Rating (t-1) (std.)	-0.119***		
	(0.023)		
Female × Female (t-1)=1 × Rating (t-1) (std.)	-0.063***		
	(0.017)		
Male × Female (t-1)=0 × Yes (t-1)		-0.059***	-0.042***
		(0.021)	(0.014)
Male × Female (t-1)=1 × Yes (t-1)		-0.025	-0.025
		(0.026)	(0.017)
Female × Female (t-1)=0 × Yes (t-1)		-0.100***	-0.045**
		(0.025)	(0.018)
Female × Female (t-1)=1 × Yes (t-1)		-0.107***	-0.035***
		(0.017)	(0.013)
Controls	Yes	Yes	Yes
p-value: Male (t) coeffs equal	0.00	0.36	0.47
p-value: Female (t) coeffs equal	0.05	0.80	0.64
N	8605	8605	8605

Notes: All regressions include workshop fixed effects. Further controls include candidate characteristics, interviewer characteristics and interview order fixed effects, as well as own TPA and leave-one-out mean rating and share of yes votes. $^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$. Standard errors are clustered at the workshop level (N=102).

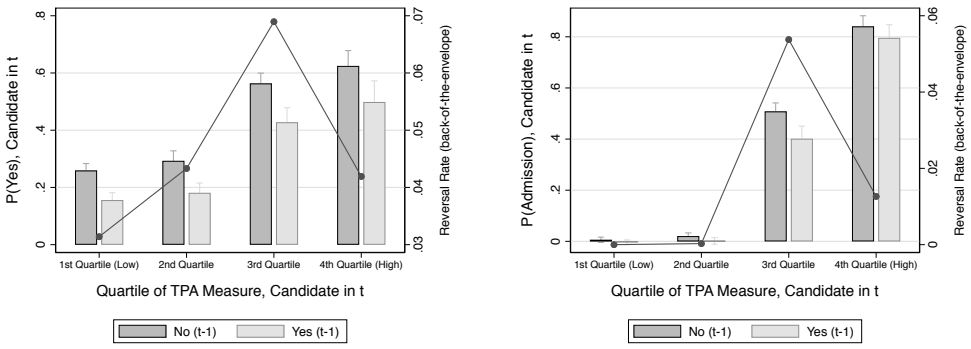
Table 4.F.3. Previous Candidate's Gender, Own Gender and Interviewer Gender

	Rating (Std.)	P(Yes Vote)	P(Admission)
	(1)	(2)	(3)
Male (t) x Male (t-1)	-0.014	-0.009	0.030
	(0.043)	(0.020)	(0.019)
Male (t) x Female (t-1)	0.146***	0.056***	0.078***
	(0.046)	(0.020)	(0.021)
Male (t) x Male (t-1) x Male Interviewer	0.059	0.035	0.028
	(0.056)	(0.026)	(0.020)
Male (t) x Female (t-1) x Male Interviewer	-0.072	0.005	-0.028
	(0.060)	(0.028)	(0.026)
Controls	Yes	Yes	Yes
Outcome Mean	0.00	0.36	0.24
N	8605	8605	8605

Notes: All regression include workshop fixed effects. Additional controls include interviewer characteristics and order fixed effects. Standard errors are clustered at the workshop level (N=102). *0.10,** $p < 0.05$,*** $p < 0.01$.

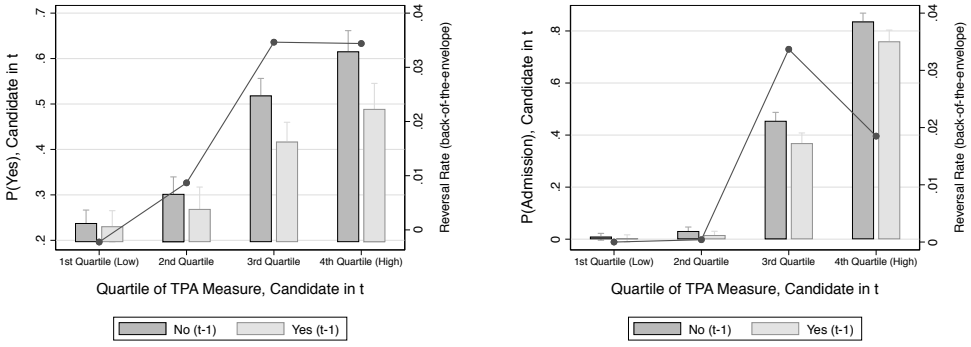
Appendix 4.G Additional Material: Quantification & Implications

In this section we describe how we compute the reversal rate based on Chen, Moskowitz, and Shue (2016). Following their approach, we derive the share of reverted decisions from a simple regression $Y_t = \beta_0 + \beta_1 Y_{t_1} + \epsilon_t$. Taking expectations, $E(Y) = \frac{\beta_0}{1-\beta_1}$. Assuming that the rate of positive decisions, $P(Y = 1)$, would would be equal in absence of the bias, reversal can be due to two situations. If the previous candidate received a no vote, the negative autocorrelation increases the current candidate's probability of a yes vote by $\beta_0 - P(Y = 1)$, i.e. her empirical probability to receive a yes vote minus her (assumed) counterfactual probability in the absence of the bias. If the previous candidate received a yes vote, the current candidate is not likely enough to receive a yes vote by $P(Y = 1) - (\beta_0 + \beta_1)$, i.e. the counterfactual probability of a yes vote minus the empirical probability. The expected number of reversals is therefore the weighted instance of the two cases $(\beta_0 - P(Y = 1))P(Y_{t-1} = 0) + (P(Y = 1) - (\beta_0 + \beta_1))P(Y_{t-1} = 1)$. Substituting $P(Y = 1) = \frac{\beta_0}{1-\beta_1}$, the rate of affirmative decisions becomes $R = -2\beta_1 P(Y = 1)(1 - P(Y = 1))$.



(a) Yes Vote, Female Candidates

(b) Admission, Female Candidates



(c) Yes Vote, Male Candidates

(d) Admission, Male Candidates

Figure 4.G.1. Influence on Admission Outcomes by Candidate Quality & Gender

Notes: Dashed lines show 95% confidence intervals.

References

Autor, David H., and David Scarborough. 2008. "Does Job Testing Harm Minority Workers? Evidence from Retail Establishments." *Quarterly Journal of Economics* 123(1): 219–77. [203]

Bagues, Manuel, Mauro Sylos-Labini, and Natalia Zinovyeva. 2017. "Does the gender composition of scientific committees matter?" *American Economic Review* 107 (4): 1207–38. DOI: 10.1257/aer.20151211. [176]

Benjamina, Daniel J. 2019. "Errors in probabilistic reasoning and judgment biases." *Handbook of Behavioral Economics*, 69. [197]

Bhargava, Saurabh, and Ray Fisman. 2014. "Contrast Effects in Sequential Decisions: Evidence from Speed Dating." *Review of Economics and Statistics* 96 (3): 444–57. DOI: 10.1162/REST. eprint: arXiv:1011.1669v3. [176]

Bindler, Anna, and Randi Hjalmarsson. 2018. "Path Dependency in Jury Decision Making." [176]

Bohren, J. Aislinn, Alex Imas, and Michael Rosenberg. 2019. "The Dynamics of Discrimination: Theory and Evidence." *American Economic Review* forthcoming: [176]

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2019a. "Memory and Reference Prices: An Application to Rental Choice." *AEA Papers and Proceedings* 109 (May): 572–76. [177]

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2019b. "Memory, Attention, and Choice." *Working Paper*, [173, 176, 199]

Caeyers, Bet, and Marcel Fafchamps. 2016. "Exclusion Bias in the Estimation of Peer Effects." (22565): [187]

Calsamiglia, Caterina, and Annalisa Loviglio. 2019. "Grading on a curve: When having good peers is not good." *Economics of Education Review* 73: 101916. [175]

Card, David, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri. 2019. "Are Referees and Editors in Economics Gender Neutral?*." *Quarterly Journal of Economics*, (11): [183]

Chen, Daniel, Tobias J Moskowitz, and Kelly Shue. 2016. "Decision Making under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires." *Quarterly Journal of Economics* 131 (3): 1181–242. [176, 201, 230]

Damisch, Lysann, Thomas Mussweiler, and Henning Plessner. 2006. "Olympic Medals as Fruits of Comparison? Assimilation and Contrast in Sequential Performance Judgments." *Journal of Experimental Psychology: Applied* 12 (3): 166–78. DOI: 10.1037/1076-898X.12.3.166. [176]

DeGroot, Morris H. 1970. *Optimal statistical decisions*. New York, NY [u.a]: McGraw-Hill. XVI, 489. [223]

Ginsburgh, Victor A, and Jan C van Ours. 2003. "Expert Opinion and Compensation: Evidence from a Musical Competition." *American Economic Review* 93 (1): 289–96. DOI: 10.1257/000282803321455296. [175]

Guryan, Jonathan, Kory Kroft, and Matthew J. Notowidigdo. 2009. "Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments." *AEJ:Applied Economics* 1 (4): 34–68. [185–187]

Hartzmark, Samuel M., and Kelly Shue. 2018. "A Tough Act to Follow: Contrast Effects in Financial Markets." *Journal of Finance* 73(4): 1567–613. [176]

Herbst, Daniel, and Alexandre Mas. 2015. "Peer Effects on Worker Output in the Laboratory Generalize to the Field." *Science* 350 (6260): 545–49. DOI: 10.1126/science.aac9555. eprint: http://science.sciencemag.org/content/350/6260/545.full.pdf. [176]

Hoffman, Mitchell, Lisa Kahn, and Danielle Li. 2018. "Discretion in hiring." *Quarterly Journal of Economics* 133 (2): 765–800. [176, 203]

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics* 133 (1): 237–93. [203]

Kramer, Robin S. S. 2017. "Sequential effects in Olympic synchronized diving scores." *Royal Society Open Science* 4 (1): 160812. DOI: 10.1098/rsos.160812. [176]

Mas, Alexandre, and Enrico Moretti. 2009. "Peers at Work." *American Economic Review* 99 (1): 112–45. DOI: 10.1257/aer.99.1.112. [176]

Mengel, Friederike, Jan Sauermann, and Ulf Zölitz. 2018. "Gender Bias in Teaching Evaluations." *Journal of the European Economic Association* 17 (2): 535–66. [176, 199]

Neumark, David, Roy J. Bank, and Kyle D. Van Nort. 1996. "Sex Discrimination in Restaurant Hiring: An Audit Study." *Quarterly Journal of Economics* 111 (3): 915–41. URL: https://EconPapers.repec.org/RePEc:oup:qjecon:v:111:y:1996:i:3:p:915-941.. [176, 199]

Nickell, Stephen. 1981. "Biases in Dynamic Models with Fixed Effects." *Econometrica* 49 (6): 1417–26. [190, 192]

Oyer, Paul, and Scott Schaefer. 2011. *Personnel Economics: Hiring and Incentives*. Vol. 4, PART B, 1769–823. DOI: 10.1016/S0169-7218(11)02418-X. [175]

Rabin, Matthew. 2002. "Inference by Believers in the Law of Small Numbers." *Quarterly Journal of Economics*, (Vol. 117, No. 3): 775–816. [197]

Rouse, Cecilia, and Claudia Goldin. 2000. "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians." *American Economic Review* 90 (4): 715–41. URL: https://EconPapers.repec.org/RePEc:aea:aecrev:v:90:y:2000:i:4:p:715-741. [176, 199]

Sacerdote, Bruce. 2011. "Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?" In. Edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Vol. 3, Handbook of the Economics of Education. Elsevier, 249–77. DOI: https://doi.org/10.1016/B978-0-444-53429-3.00004-1. [176]

Sacerdote, Bruce. 2014. "Experimental and Quasi-Experimental Analysis of Peer Effects: Two Steps Forward?" *Annual Review of Economics* 6 (1): 253–72. DOI: 10.1146/annurev-economics-071813-104217. [176]

Sarsons, Heather. 2019. "Interpreting Signals in the Labor Market: Evidence from Medical Referrals." *Working Paper*, [176]

Simonsohn, Uri. 2006. "New Yorkers Commute More Everywhere: Contrast Effects in the Field." *Review of Economics and Statistics* 88 (1): 1–9. DOI: 10.1162/rest.2006.88.1.1. [176]

Simonsohn, Uri, and Fransesca Gino. 2013. "Daily Horizons: Evidence of Narrow Bracketing in Judgment from 10 years of MBA-admission Interviews." *Pyschological Science* 24 (2): 219–24. [175]

Simonsohn, Uri, and George Loewenstein. 2006. "Mistake #37: The Effect of Previously Encountered Prices on Current Housing Demand." *Economic Journal* 116 (508): 175–99. [176]

Simonson, Itamar, and Amos Tversky. 1992. "Choice in Context: Tradeoff Contrast and Extremeness Aversion." *Journal of Marketing Research* 29 (3): 281–95. [173]

VandenBos, Gary R. 2007. *APA dictionary of psychology*. American Psychological Association. [198]