I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11919

Competing on the Holodeck: The Effect of Virtual Peers and Heterogeneity in Dynamic Tournaments

Frederik Graff
Christian Grund
Christine Harbring

OCTOBER 2018

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11919

Competing on the Holodeck: The Effect of Virtual Peers and Heterogeneity in Dynamic Tournaments

Frederik Graff
RWTH Aachen University

Christian Grund
RWTH Aachen University and IZA

Christine Harbring
RWTH Aachen University and IZA

OCTOBER 2018

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Competing on the Holodeck: The Effect of Virtual Peers and Heterogeneity in Dynamic Tournaments

We propose experiments in virtual reality (VR) as a new approach to examining behavior in an economic context, e.g., heterogeneity in dynamic tournaments. We simulate a realistic working situation in a highly immersive environment. Implementing a tournament in VR, we are able to mitigate the reflection problem, which usually undermines research on dynamic interaction. Moreover, VR allows us to ceteris paribus control for the performance of the virtual peer (humanoid avatar), and thus to get an understanding of the reaction of the subject to the avatar in a really dynamic setting, as the subject is constantly able to observe the avatar's performance. Based on a first experimental phase, we match our subjects with an avatar yielding a specific output. We observe that the subjects' performance is highest in a homogeneous tournament, i.e., when they compete against an avatar achieving the same output as they did in the preceding phase. Interestingly, these results are particularly driven by peer effects rather than by tournament incentives. We extensively track the behavior of subjects and the particular situation and, e.g., examine the role of intermediate score differences and the degree of the subjects' movements.

JEL Classification: C91, D9, J33, M52

Keywords: reflection problem, tournament, peer effects, real effort, virtual reality, avatar, heterogeneity

Corresponding author:
Christian Grund
School of Business and Economics
RWTH Aachen University
Templergraben 64
52056 Aachen
Germany
E-mail: christian.grund@hrm.rwth-aachen.de

Competing on the Holodeck

The effect of virtual peers and heterogeneity in dynamic tournaments

1. Introduction

When Michael Phelps quit the stage of professional sports in 2016 with 23 Olympic gold medals, 26 world championships, and a lot more of national and international titles and dozens of awards, he was not only the most successful swimmer of all times but also the most decorated sportsman to have ever taken part in the Olympic Games. Since his international debut in 2000, he has set 39 world records (individual and relay), of which three individual records still hold. Most of them have been broken by himself (Fina 2018). Unsurprisingly, Michael set most of his records, i.e. 31, in final heats where he competed against the seven best swimmers of the tournament. All of them had passed qualifying heats previously, and it seems plausible that each swimmer had to give their very best to beat the others. Thus, we can assume that the swimmers in a final heat of the Olympic Games or World Championships are as homogeneous in ability as is possible. Accordingly, the alltime most successful tennis player, Roger Federer, referred to a question with regard to his motivation and to his longtime opponent and current world no. 1 player, Rafael Nadal, said: “He will forever be my ultimate opponent. He was the one who helped me to improve the most and to be a better player. […] Rafa’s presence was an extra motivation” (Archer 2017).

Homogeneity in ability , however, differs across context specific competitions among individuals or teams. What is it like, for example, when a firm’s employees who are heterogeneous in ability compete for a promotion? What role does intermediate information about performance play? Moreover, is it the tournament pay scheme or the pure peer’s presence that drives incentives?

In this paper we study the effect of ex ante heterogeneity on tournament incentives, also taking intermediate heterogeneity in terms of score differences into account. As the individual behavioral motives in human interactions are always confounded with spontaneous reactions to others’ behavior, studying the causality of individual behavior in interactions is complicated as long as the parties mutually react towards each other. Manski (1993) was the first to describe this challenge and coined the notion “reflection problem”. The reflection problem means that in experiments where participants interact, it can never be disentangled whether one participant *i* influences participant *j* or vice versa.

We realize the experiment in a virtual reality (VR) environment because VR enables us to manipulate both heterogeneity and peer observability systematically. With respect to the reflection problem, this means that the (computer-generated) avatar never reacts to the subject. Thus, the traditionally unclear causality of mutual influence (the reflection) can be solved. Furthermore, VR provides us with a number of additional measures (e.g. subjects' movements) which can be used as proxies of effort and thoroughness complementing the traditional output measures.

The scientific contribution of this paper is twofold. Being one of the first studies to use a virtual environment for economic research, our paper proposes a novel methodological approach for economic experimentation with meaningful strengths for our topic in particular. Moreover, this paper adds to the empirical literature on the effectiveness of incentive schemes. Especially, we address the topic of the motivational effects of ex ante heterogeneity in tournaments, taking intermediate relative performance into account. The paper is organized as follows. After an overview of the literature on heterogeneity in tournaments and peer effects, we discuss the application of virtual environments (VE) for experimental economic research. In section 2 the experimental design is described. The results are given in section 3. The paper closes with a discussion of our findings in section 4.

2. Background

2.1 Heterogeneity in Tournaments

A whole bunch of theoretical as well as empirical and experimental studies on the effect of ex ante heterogeneity on tournament incentives already exists[1]. Starting with the seminal paper by Lazear and Rosen (1981), theoretical analyses predict that tournament incentives should be highest in symmetric tournaments, i.e. where the contestants are equally able. The intuition behind these findings is the so called "discouragement effect" (Dechenaux, Kovenock and Sheremeta 2015): In asymmetric tournaments, the weaker agent realizes that the likelihood of winning the tournament is too low and reduces effort, while the stronger agent anticipates this and withholds effort herself (e.g. O′Keeffe 1984, McLaughlin and Ehrenberg 1988, Kräkel and Sliwka 2004, Gürtler and Kräkel 2010, Gürtler and Gürtler 2015, Harbring and Lünser 2008). For example, Casas-Arce and Martínez-Jerez (2009) analyze the sales contests organized by a commodities company and find that winning participants decrease their effort as their lead extends, whereas the effort of trailing participants fades only when the gap between them and a winning position becomes very large.

[1] For an extensive overview see, e.g., Dechenaux, Kovenock and Sheremeta (2015).

For obvious reasons, tournaments in real organizations usually involve heterogeneous contestants. Such a heterogeneity of abilities has, for example, been analyzed empirically by implementing different cost of effort or output functions or discrimination procedures in an abstract laboratory environment (Schotter and Weigelt 1992, Harbring and Ruchala 2003, Orrison, Schotter and Weigelt 2004, Harbring et al. 2007, Fonseca 2009, Gürtler and Kräkel 2010, Fehr and Schmid 2011, Kimbrough, Sheremeta and Shields 2014). In these lab experiments, findings are similar. In line with the theoretic prediction, in lab experiments with chosen effort more heterogeneity leads to lower effort provision (Davis and Reilly 1998, Anderson and Stafford 2003, Anderson and Freeborn 2010, Fonseca 2009, Kimbrough, Sheremeta and Shields 2014).

Of course, implementing such abstract models of a real tournament in a laboratory environment boils down the strategic situation to its very essentials and it is unclear whether certain motivational effects do still occur[2]. Hence, in lab experiments using real effort instead of making decisions by choosing numbers, the results are ambiguous. Gill and Prowse (2012) find evidence for a discouragement effect in real-effort experiments while the findings of, e.g., (Berger and Pope, 2011) tell against it.

Kräkel (2008) argues, for example, that weak contestants may feel pride if they make it to achieve the winner prize in a tournament against a superior player. This pride leads to additional incentives by enlarging the subjectively perceived winner prize. More able agents may overexert effort (compared to the theoretic prediction) as they are determined not to lose.

While chosen effort experiments realize heterogeneity via cost functions, production functions, prize valuations, or head starts of one contestant, experiments in real-effort environments – where participants actually work on a task in the laboratory – naturally capture different abilities of subjects. However, the experimenter loses control in the sense that it is very unlikely to always find a pair with a specific ability (for an example of heterogeneity in a real-effort experiment, see van Dijk, Sonnemans and van Winden 2001, Hammond and Zheng 2013).

However, to be able to analyze the effect of a certain degree of heterogeneity, it is crucial to match participants of known abilities. One exception is the work of Eriksen, Kvaløy and Olsen (2011), who assign participants, after these have been asked to work on a simple cognitive real-effort task, to one of three ability groups and then vary the matching to tournaments. They, however, analyze behavior in tournaments where participants may endogenously decide on the wage spread and, interestingly, find that effort is higher in heterogeneous settings, as agents set higher wage spreads. Similarly, Cason, Masters and Sheremeta (2010) manipulate the relative skills of participants in adding up two-digit numbers exogenously by letting

[2] For a detailed comparison of chosen-effort and real-effort in contest experiments see Dechenaux, Kovenock and Sheremeta (2015).

the participants compete against the pre-recorded scores of earlier subjects (historical data). They find that heterogeneity seems to discourage weaker contestants while stronger contestants are not affected by heterogeneous abilities. They do not take intermediate information with regard to performance during the tournament into account, though.

Another empirical approach to testing predictions derived from theory is to use sports data. In sports you naturally have tournament structures, the documentation of sports is outstanding, and sports data are easily accessible. Field studies with sports data find evidence for the theoretical prediction that effort tends to be higher the more homogeneous the competing individuals or teams are (Sunde 2009, Bach, Gürtler and Prinz 2009, Nieken and Stegh 2010, Brown 2011, Kocher, Lenz and Sutter 2012, Deutscher et al. 2013, Berger and Nieken 2016).

2.2 Peer Effects in Tournaments

An additional crucial feature of many tournaments in reality is that they are typically dynamic, i.e., one occasionally or constantly observes the competitor and has an idea of her performance so far. As asymmetry in ability tends to undermine tournament incentives from a theoretic perspective, one might be tempted to recommend avoiding observability of peer performance. While observability can quite easily be avoided in lab experiments, this seems hardly possible in organizations in many contexts. Thus, it remains unclear to which extent the incentives of a tournament root in the pay scheme or in the pure peer effect. In our design, we are able to address this additional aspect of peer observability in tournaments. Thus, we address the question of to what extent behavior is driven by tournament prizes or solely by the peer’s presence.

Based on previous studies, the latter is assumed to have an effect on the performance of subjects in working situations. For more than a decade, scholars have tried to quantify this spill-over effect mainly with observational data from situations in which humans work on individual tasks but are able to observe each other (Katz, Kling and Liebman 2001, Sacerdote 2001, Topa 2001, Mas and Moretti 2009, Bandiera, Barankay and Rasul 2010, Waldinger 2012, Chan, Li and Pierce 2014, Chan 2016). Most of these studies find evidence for the existence of peer effects. In one of the first lab experiments on peer effects, Falk and Ichino (2006) confirm their predictions that the average output of a pair of peers should exceed that of a single subject. In a meta study of 34 laboratory and field studies, Herbst and Mas (2015), who provide an excellent overview of empirical evidence on peer effects, find a mean study-level estimate of a change in a worker’s productivity in response to a 1 % increase in a co-worker’s productivity of $\gamma = 0.12$ (SE = 0.03). Laboratory experiments and field studies do not significantly differ in γ.

Gürerk et al. (2016b) conduct an experiment based on a virtual setting similar to the one used in this paper. They embed a virtual human as co-worker of a human subject into an immersive virtual environment and observe that low productive human subjects increase their work performance more when they observe a low productive avatar than when they observe a high productive co-worker. Similarly, Mas and Moretti (2009) investigate peer effects in a group production setting in a field study with cashiers in a grocery store. They find that peer effects dominate free-riding incentives. However, only those workers who can be seen by a high productive peer increase their efforts and overcome free-riding incentives. Cashiers who can observe the high productive worker, but cannot be observed by her, do not change their efforts. In order to control for social influences which may confound peer effects in the field, van Veldhuizen, Oosterbeek and Sonnemans (2018) transfer the Mas and Moretti (2009) setting into the lab but cannot replicate their findings.

There have been attempts to investigate the effect of being behind or ahead of others in an abstract lab environment (Schotter and Weigelt 1992, Gürtler and Harbring 2010, Müller and Schotter 2010, Ludwig and Lünser 2012). However, usually only the effect of some intermediate information (typically modeled by ex ante heterogeneity, e.g. a head start or by competing against scores achieved by subjects in previous rounds), is analyzed and not a real dynamic setting as it is difficult to implement in the lab (e.g. Kuhnen and Tymula 2012, Gill and Prowse 2012, Fallucchi, Renner and Sefton 2013, Mago, Samak and Sheremeta 2016, Deutscher and Schneemann 2017). Mago, Samak and Sheremeta (2016) find in a laboratory experiment that the provision of information feedback about the effort provided by other contestants at the end of each period does not affect the aggregate effort but decreases the heterogeneity of effort in a repeated Tullock contest. Ludwig and Lünser (2012) realize being ahead and lagging behind by providing information about a contestant's first period score just before starting the second period of a two-player rank-order tournament with two stages and find that participants adjust their effort to the information. And Kuhnen and Tymula (2012) provide intermediate feedback about the contestant's relative rank at the end of each round of a 18 round real-effort experiment.

Thöni and Gächter (2015) report that subjects in a chosen-effort lab experiment with intermediate information follow a low-performing but not a high-performing peer. Highly productive peers adjust their efforts down when they are informed about the other worker's effort. There are also some attempts to investigate the role of ex ante heterogeneity and intermediate information in the field (e.g. Grund, Höcker and Zimmermann 2013 analyze changes in strategies of basketball teams in the NBA with regard to intermediate scores). Conducting these field studies, however, it is neither possible to perfectly control for important issues of the working environment and of the competitor nor to capture mutual interdependencies in behavior at each point of time.

2.3 The Reflection Problem

Manski (1993), who coined the notion “reflection problem”, defines it as the impossibility to infer "whether the average behavior in some group influences the behavior of the individuals that comprise the group" (Manski 1993: 532)[3]. As an intuition he uses the thought experiment of a person with no idea about optics and human behavior who observes the reflection of another person in a mirror. The first person would be unable to tell whether a movement of the reflection is caused by the simultaneous movement of the second person or vice versa. Manski states that this inference is impossible as long as no further information about the group’s characteristics exist. And it remains difficult to impossible if these variables are functionally dependent or are statistically independent. In order to avoid the reflection problem, Manski demands data from controlled experimental studies. However, a reflection problem cannot easily be ruled out here, either. Also in experiments, in the case of an interaction of two or more participants it cannot be concluded whether the behavior of subject *j* influences the behavior of *i* or vice versa. Only if it is ensured that one of the two definitely does not react to the other is a clear inference possible. However, this should hardly be feasible in experiments with human participants.

The relevance of the reflection problem has already been acknowledged in the field, and, thus, there are interesting previous studies who mitigate the reflection problem in the lab with regard to different research questions. Van Veldhuizen, Oosterbeek and Sonnemans (2018) for example conduct a computerized real-effort lab experiment where the participants, depending on their treatment, are informed that another participant will receive information about the number of exercises solved by her teammate and that they themselves will receive such information about other teammates. As the flow of information always goes unidirectional, the authors may diminish the reflection problem. Similarly, Sausgruber (2009) runs a standard linear public good game and varies the information the participants receive about each other’s contribution. According to the author, the lab situation itself where the participants sit in separate cubicles assures the non-existence of an “identification problem”. Thöni and Gächter (2015) circumvent the reflection in gift-exchange experiment by making “the effort of the other agent exogenous. To achieve this, both agents first choose their efforts simultaneously and then, after having learned the effort decision of their co-agent, are given the opportunity to revise their effort, *holding their co-agent’s effort constant*. Since the design removes any material and strategic incentives to revise effort, revision decisions […] tell us

[3] Manski places the reflection problem in a series with other endogenous effects that sociological research deals with: peer influence, conformism, neighborhood effects, etc.. For him, the reflection problem is one reason why numerous social effects are well understood, but an integrating evidence-based theory on the effect of group behavior on individual behavior is missing.

about the extent to which people change their effort *because* of the effort chosen by the co-agent" (Thöni and Gächter 2015: 73).

These examples like many others (e.g. Eisenkopf 2010, Beugnot et al. 2013, Georganas, Tonin and Vlassopoulos 2015) avoid the reflection problem by simply precluding both mutual observability (participants shielded from each other with partition walls and treated anonymously) and/or mutual reaction (unidirectional flow of information, simultaneous decision making). Though, Manski's definition of the reflection problem also refers explicitly to situations where individuals observe each other in person (face-to-face or face-to-back). Such a setting seems particularly relevant in a majority of employment situations when employees work together.

Some studies therefore work with actors as confederates of the experimenter (Brockner 1987, England & Buskist 1995, Ijzerman, van Dijk and Gallucci 2007, van Kleef et al. 2009). The confederates are then asked to show rehearsed behavior and not to be influenced by the participant. Whether even a good actor is always able to behave the same over several sessions is at least questionable. Alternatively, the actor is videotaped and the video sequence is played to the participants (Lakin and Chartrand 2003, van Schie, van Waterschoot and Bekkering 2008). Here, however, inference is paid for with the abandonment of realism.

In our opinion, an obviously appropriate way to mitigate the reflection problem in a context where a participant may continuously observe others in a highly dynamic setting is to use virtual reality in experiments. Here, it is possible to let the participants interact with a programmed avatar in an immersive environment. The avatar, because programmed, always behaves in a same way without tiring and certainly does not react to the participants' behavior (Gürerk et al. 2016a, Innocenti 2017). To sum up, while many studies have mitigated the reflection problem by avoiding observability, VR enables us to maintain observability in a highly dynamic context without falling for the reflection problem.

2.4 Experiments in a Virtual Environment

Augmented Reality (AR) and Virtual Reality (VR) are the emerging technologies at the workplace, in medicine, military, gaming, navigation, and private households. Some social or occupational activities possibly will be completely shifted into virtual environments (e.g. conferences around a virtual conference table with the participants represented by avatars) or will at least be assisted by augmented reality devices. Physicians use a combination of endoscopy and AR technology to create three-dimensional models of the structures to be diagnosed or operated (Székely and Satava 1999). Head-up displays in cars enrich the driver's visual field with warnings, highlighted obstacles or parking slots, and projections of braking points, or the optimal way to maneuver into a parking bay. Facebook managers are said to already use virtual meeting places for their conferences. The application of (fully or semi) immersive VR/AR systems is likely to become common in the future and so will the interaction of real human beings with virtual humans. Thus, studying the interaction of real and virtual humans is more than warranted.

Since the early 1990s, psychologist have tested the application of VR systems for situations which are too dangerous or impractical to be carried out in reality. Such applications are, for example, vehicle simulations, military training or psychotherapy, namely the exposure of patients suffering from phobias to their phobic stimuli (spiders, open places, narrow chambers, etc.) (Slater et al. 2009). Especially for the success of psychotherapeutic applications it is crucial that the virtual environment is as immersive as possible and is accepted as near real by the patients.

Numerous studies covering a broad range of topics, such as interpersonal distance keeping, public speaking anxiety, evacuation scenarios, or language interaction have already demonstrated the external validity of VE experiments in a variety of contexts (Pertaub, Slater and Barker 2002, Bailenson et al. 2003, Slater et al. 2006, Kothgassner et al. 2016, Iachini et al. 2016, Moussaïd et al. 2016, Heyselaar, Hagoort and Segaert 2017). The results of these experiments impressively demonstrate that, the reactions of human beings to virtual humans (avatars), objects, and events in VE is not very different from their reaction to the same stimuli in real situations. Following this line of research our experiment is based on the assumption that the interaction between real and virtual humans allows inference on the interaction between real human beings.

One interesting example for such a study is Slater et al. (2006) who replicated the famous experiment on obedience by Milgram (1963) using a virtual environment with the objective of understanding whether a virtual peer could simulate a real human being in experimental research. In the experiment the participants were asked to give electric shocks to the (virtual) “Learner” in case questions were not correctly answered. The authors conclude that subjects reacted similarly towards the virtual avatar compared to a human counterpart regarding different subjective (self-assessment via questionnaire), physiological (e.g. skin conductance, heart rate), and behavioral (withdrawal from experiment) measures, although all of them knew that the avatar was just a program. This finding is in line with other studies from psychology, all indicating that virtual avatars can serve as appropriate proxies of human counterparts for experimental research.

Another illustrative example is provided by Bailenson et al. (2003) who found that regarding interpersonal distance, human participants treated virtual humans in a manner similar to actual humans. They let human participants walk through a virtual room and meet a virtual avatar. The spatial distance between the participant and the avatar – being an established measure of intimacy – is the dependent measure. The results clearly indicate that participants keep spatial distance from virtual humans (avatars) in the same way that they do from real humans (greater distance when approaching the avatars’ fronts compared to their backs; giving more personal space to avatars engaging the subjects in mutual gaze; moving farthest from avatars when these invaded their personal space). Finally, Pertaub, Slater and Barker (2002) tested the public speaking anxiety of subjects in response to three different types of virtual audience. They found that the subject reacted massively to a (virtual) audience showing disrespect for the speech that the subjects were ask to hold. The negative audience clearly provoked an anxiety response, irrespective of the normal level of public speaking confidence of the subject.

Implementing a tournament in VR allows us to *ceteris paribus* control for the performance of a virtual co-worker (an agent-avatar, "avatar" in the following), and to vary and to adapt it systematically to the subject's beforehand measured ability. Thus, we get an understanding of the subject's reaction to the avatar in a dynamic setting by analyzing behavior changes during working at a task with respect to relative intermediate performances and disentangling these from pure peer effects.

3. Experimental Procedure

The experimental setting was implemented in the aixCAVE, a five-sided, room-mounted immersive environment with a size of 5.25m x 5.25m x 3.30m, hosted by the Virtual Reality Group of RWTH Aachen University (Cruz-Neira et al. 1993, Kuhlen 2014). The VR Group has programmed the aixCAVE for us in such a way that a human subject entering the room perceives it as a virtual production hall. The four walls as well as the floor display the projections. Subjects wear special 3D glasses that allow a stereoscopic projection and the scenario is shown as a quasi-holographic image similar to 3D cinemas. The projection is user-centered and depends on the subject's position in the room. Moreover, we tracked the movement of the subject's hand by attaching tracking markers, by means of a hand band, to the participant's dominant hand. Thus, the participant can use her own hand to grab a cube and put it into the bin (see figure 1).[4] The advantages of the VR environment in this study are the more realistic simulation of the real-effort task (see section 3.1), the near-perfect experimental control, the easy manipulation of the avatar's performance, the gaining of additional data (e.g. tracking of subjects' movements), the possibility to realize a dynamic competition, and the reduction of the reflection problem (Manski 1993).

[4] For an overview of economic experiments in virtual environments and a comprehensive discussion of their external validity see Gürerk et al. (2016a).

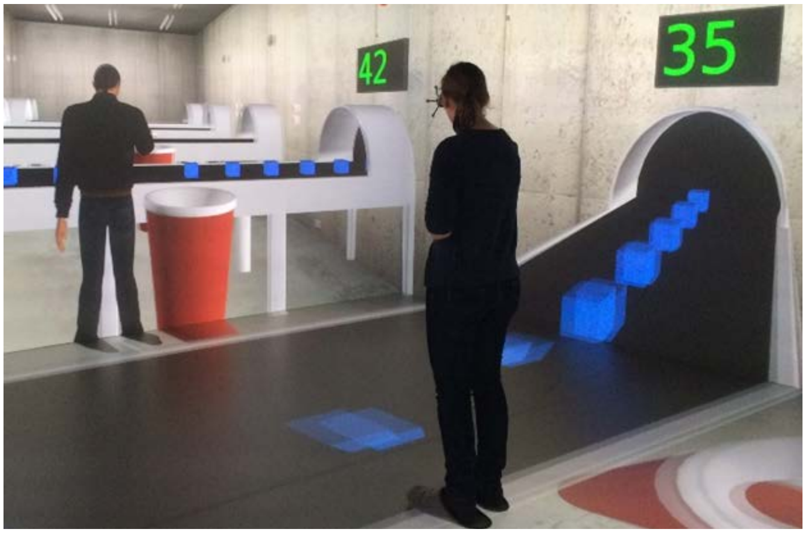

Figure 1: Subject in the foreground working on the conveyer belt task in the aixCAVE of RWTH Aachen University in the presence of a virtual co-worker. Note the score boards to the right of the conveyor belts. In this picture, the projection of the virtual production hall appears fuzzy because it is stereoscopic. For the participant in the foreground, who is wearing VR-glasses, the projection is distinct.

3.1 Real-effort Task and Sequence of Experiment

The human subject is asked to perform a task. In all (but one) treatments, the participant is able to observe an avatar working on the same task (see figure 1). Participants have to sort colored cubes, which pass by the subject on a virtual conveyor belt (36 cubes/minute). Regular cubes are blue on each side and are to be left on the conveyor belt. Defective cubes are red on one side (either on the right, back, or bottom side) and are to be sorted into a virtual bin. Subjects can easily take cubes from the belt and turn them in their hand to check for red sides. Each defective cube in the bin (hit) counts as one performance point; each regular cube in the bin (false alarm) counts as minus one performance point. We keep the order of defective and regular cubes constant for each subject. During each part of the experiment, 360 cubes, of which 180 cubes have a defect, pass by the subject. To keep the task's difficulty constant over time, we divide the stream of 360 cubes into 10 *sequences* of 36 cubes each. Within one sequence, always 18 cubes have a defect, of which always six cubes have the defect on the right/bottom/down side each. Thus, cube defect types are equally distributed within one sequence and the distribution of cube defects and defect types within the sequences follows a random order but is the same for each subject.

Figure 2 depicts the set-up of the experiment. After entering the hall where the aixCave is located, participants are provided with written instructions regarding the training phase as well as the ability check phase. During the training phase, subjects enter the aixCave and have the chance to get acquainted with the real-effort task as well as with the particular technology of grabbing the cubes. Participants are allowed to ask questions to the experimenter and they do not yet receive a payment. During the ability check phase (in the following also denoted as “part I”), they work on the task for 10 minutes and are told that they will receive a piece-rate of 4 Eurocents per performance point – one point per correctly sorted defective cube less one penalty point per incorrectly sorted cube without a defect. In sum, 360 cubes pass during the 10 minutes each 180 with and 180 without a defect. The maximum number of performance points is therefore 180 (= 7.20 €). Each participant’s accumulated performance points are displayed by a virtual counter at the wall at any time during the experiment.

After part I, participants receive written instructions for part II of the experiment, i.e., the treatment phase. Again, they work for 10 minutes on the same task – now with a different order of cubes. In all but one treatment an avatar is now working on a conveyor belt in front of the human subject. Subjects are informed that there is another worker on the other conveyor belt who is controlled by the computer. The payment procedure is outlined in the written instructions. In the piece-rate treatments participants receive a wage per performance point independent of others; in the tournament treatments participants receive a higher wage if they obtain more performance points than the other “worker” (in the following: avatar). The accumulated performance points are displayed by two virtual score boards at the wall, one for the subject and one for the avatar. Thus, the participant has the chance to observe the score of both workers at any time during part II.

Part I does not only serve as a measure of the participant’s basic ability to do the task in order for us to get a measure of the performance increase between the two parts due to the treatment intervention, but also enables us to use the subject’s ability for manipulating the performance level of the avatar. Dependent on the treatment that we implement, an avatar is able to achieve (i) the same performance score as the subject in part I, (ii) 10 percentage points more or (iii) 10 percentage points less. This procedure enables us to perfectly control for the avatar’s ability keeping everything else constant. Participants are not informed about the ability level of the avatar but can observe both players’ score throughout part II.

After part II, participants answer a questionnaire at a desk outside the aixCave. The questionnaire covers some demographic information, behavior, and decisions throughout the experiment as well as the subjects’ experience with VR-technology in general and the aixCave in particular, their experience with conveyor belt work, their perception of the avatar and its performance and of the immersive sensation, a measure of co-presence (Poeschl and Doering 2015), measures of competitiveness and risk attitude, and beliefs about their own ability, effort, and performance. In open questions, the participants can comment on technical problems, their experiences with VR, their perception of the avatar, and their opinion on the experiment in general. At the end, participants received their payment. On average, the experiment lasted for approximately 30 minutes and they receive a payment of 12.84 Euro (SD = 2.11). A number of *n* = 131 subjects participated.

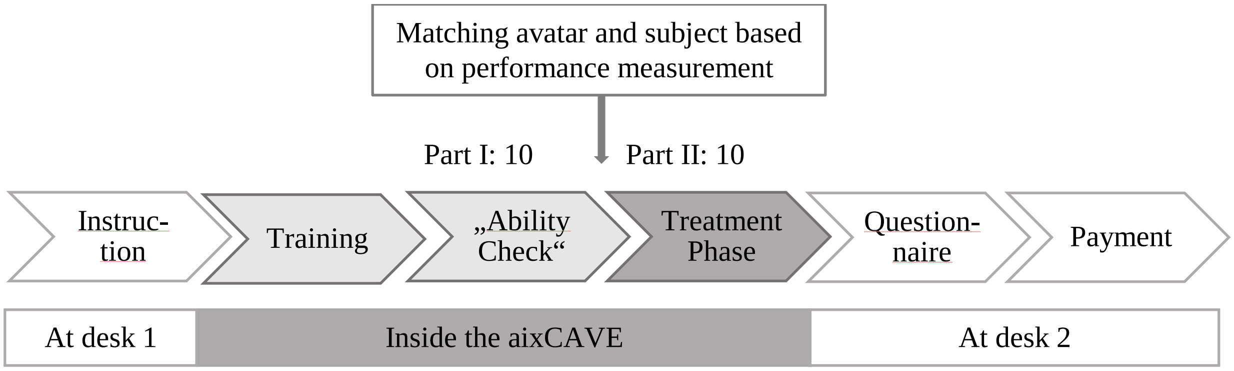

Figure 2: Set-up of Experiment

3.2 Treatments

As described above we vary the performance level at which the avatar is fulfilling his task, and we can thus, measure the influence of the avatar's performance on a differently able subject. The avatar's ability is adjusted to the participant's ability that has been measured in part I. This allows us to control for heterogeneity to a very large extent, which would hardly be feasible if we had two human competitors.

All treatments follow the same sequence outlined above. In part I, there is no avatar and all subjects work under a piece-rate pay scheme. In part II, the pay scheme varies by treatment and is as follows:

In the two treatments *PR* and *PR_equal* subjects are paid a piece rate incentive of 4 Eurocents per performance point (as in part I). Only in *PR_equal* is an avatar – of the same ability – also working on the task, so that we can compare behavior to participants who work alone in part II in *PR*. In the three tournament treatments, the participant receives a high payment of 8 € in the case where she achieves a higher performance score than the avatar and 2 € otherwise. Based on the results of a pre-test, the tournament prizes were chosen in the way that the expected payoff in the tournament treatments approximates the payoff in the piece-rate treatments when the participant is facing an equally able avatar. The ability of the avatar is manipulated based on the subject's performance score achieved in part I. (see figure 2).

Table 1: Overview of Treatments

	Payment Scheme	Avatar	Performance of Avatar relative to subject's performance in part I
PR	Piece rate	No	-
PR_equal	Piece rate	Yes	Equal
T_equal	Tournament	Yes	Equal
T_disad	Tournament	Yes	Avatar with higher performance (10 % points)
T_adv	Tournament	Yes	Avatar with lower performance (10 % points)

4. Results

We use the number of errors defined as the inverse of performance points mentioned above (number of errors = 180 – performance points) as our (inverse) performance indicator. Errors occur if subjects miss a defected cube (miss) or sort out cubes without defects (false alarm).

Figure 3 shows the average number of errors made per treatment and part. It becomes obvious that errors are much more likely to be made in part I (during the ability check) than in part II. For every treatment, the difference in the number of errors made between part I and part II is significant ($p<0.01$ each, Wilcoxon matched-pairs signed-ranks test, two-tailed). For our analysis we mainly focus on the comparison of behavior in part II, as we are interested in systematic effects due to our treatment manipulation. Based on the data of 131 subjects, the descriptive statistics reveal that the performance of subjects in *PR_equal* is higher than in *PR* and similar to those in *T_equal* and *T_disadv*, although only the difference between *PR* and *T_equal* is statistically significant ($p < 0.1$, Mann-Whitney-U, two-tailed).[5]

[5] A two-tailed Mann-Whitney-U-test reveals a significant difference in the number of errors in part I between *PR* and *PR_equal*. For our analyses we will mainly use within comparisons in order to identify treatment effects and will control for the ex ante ability in our estimations such that these differences in part I do not distort our major findings.

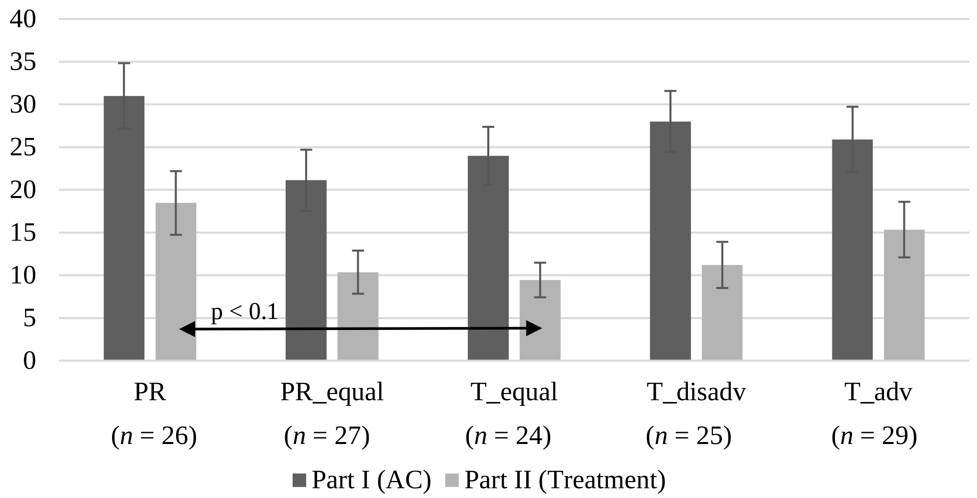

Figure 3: Average number of errors made per treatment and part (*n* = 131).

Comparing *PR* directly to *T_equal*, *T_disadv*, and *T_adv* is not unproblematic as it falls for multiple treatment variations. From *PR* to tournament treatments, two factors change: the incentive scheme and an avatar joining the game. Thus, what motivates the subject to provide more effort in part II and even more in the tournament treatments? Is it the effect of the tournament incentive or is it the motivational effect of a (virtual) peer? In the *PR_equal* treatment, subjects work under a piece-rate pay scheme while an equally able avatar is working on the same task opposite the subject. Thus, the subjects are able to observe the avatar's actions and her point score constantly but they know that their payoff is fully independent from the avatar's performance. The performance difference between *PR* and *PR_equal* indicates a peer effect, but is not statistically significant. This potential peer effect will be analyzed in detail using binary probit estimations in the following section.

Figure 4 shows the average number of errors made per sequence over all treatments. The performance increases between part I and part II could also be caused by learning over time. This may seem plausible, as the task of working on the VR-based conveyor belt is new and the subjects are most likely to be unfamiliar with it.[6] The handling of the virtual conveyor belt (VR-glasses, VR-markers on the working hand, grabbing and ungrabbing virtual cubes, etc.) is easy, but needs a moment to get used to. However, a trend in terms of learning over the sequences within the parts is hardly visible. Neither the means vary significantly over the sequences within one part nor do the standard deviations. This pattern is almost

[6] In the post-experimental questionnaire, the vast majority of subjects stated that they had no experience of the aixCave.

identical in all treatments.[7] The abrupt performance increase between sequence ten of part I and sequence one of part II is obvious in all treatments. At first glance, this finding is reminiscent a kind of "restart effect", which is relevant in cooperation problems and social dilemma situations, as previously highlighted by Selten and Stoecker (1986) and Andreoni (1988). Subjects in our setting, however, do not face any cooperation problem. We rather consider that subjects reflect on the task during the break after the first part of the experiment and conceive themselves effectively fulfilling the sorting of the cubes in the upcoming second part. We may therefore rather speak of a possible "recollecting-effect" as a very specific form of learning by recollecting past experiences and actively reflecting on them. Therefore, we have to hint that it appears in our experiment that our subjects are actually not that disadvantaged in our *T_disadv* treatment because this performance increase that starts from the beginning of part II in each treatment. We stick to our treatment denomination defined before the experiments took place, because the performance of the avatar is chosen based on past performance in part I. We reconsider the relevance of this phenomenon when discussing our results below.

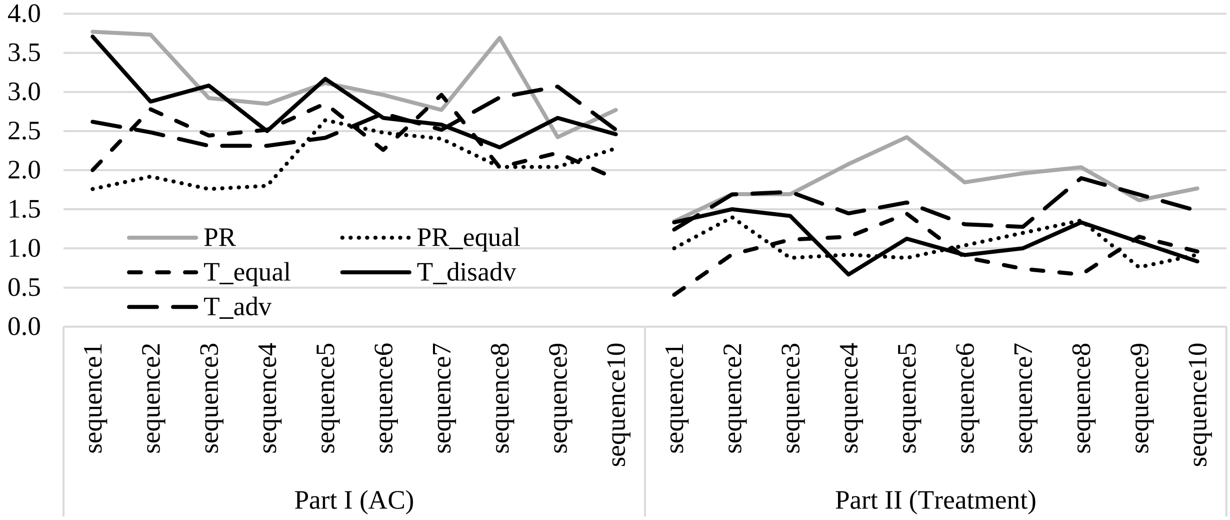

Figure 4: Average number of errors made per sequence over all treatments.
Notes: Sequences are groups of 36 cubes. In each sequence, the number of defect cubes and the distribution of the three cube defect types (right, back, down) are constant.

[7] So et al. (2017), who compare the impact of piece rate and tournament pay schemes on learning in a cognitively challenging task, find that it is only in the tournament pay scheme, and particularly in a more complex version of the task, that subjects show significant learning over time. Taking into account that our task is much less complex than the one used by So et al. (2017), the absence of learning over time in our experiment does not seem too surprising in the light of those findings.

Thanks to the virtual environment in which subject and avatar activities as well as environmental changes are tracked extremely precisely, we are able to uncover further interesting findings. Importantly, the different kinds of cube defects (red surface on back, right, or bottom side) are not equally difficult to detect. Subjects are significantly more likely to make errors when the red surface is on the cube's back side. In all treatments, the most errors are made with cubes which have their defect on that side. Thus, ability differences between subjects become especially obvious when it comes to sorting "difficult cubes". Moreover, these "difficult cubes" offer the possibility for highly motivated subjects to increase their performance compared to the avatar and compared to their own performance in the first part. Thus, the treatment effect should be stronger when looking at the "difficult cubes" only. The percentage of errors of the four defect types per treatment to be detected is given in figure 5.

Based on theoretical predictions and previous empirical findings, we expect the motivational effect of tournament incentives to be highest when the avatar and the subject are of equal ability. As hypothesized, we find that subjects in *T_equal* make significantly fewer errors on average with cubes with defects on their back side than subjects in the other treatments, indicating that participants may be more motivated in the symmetric tournament to work thoroughly and increase their output. The difference between *T_equal* and *PR* is statistically significantly different (*PR* vs. *T_equal*, Mann-Whitney-U-test, p = 0.000, two-tailed).

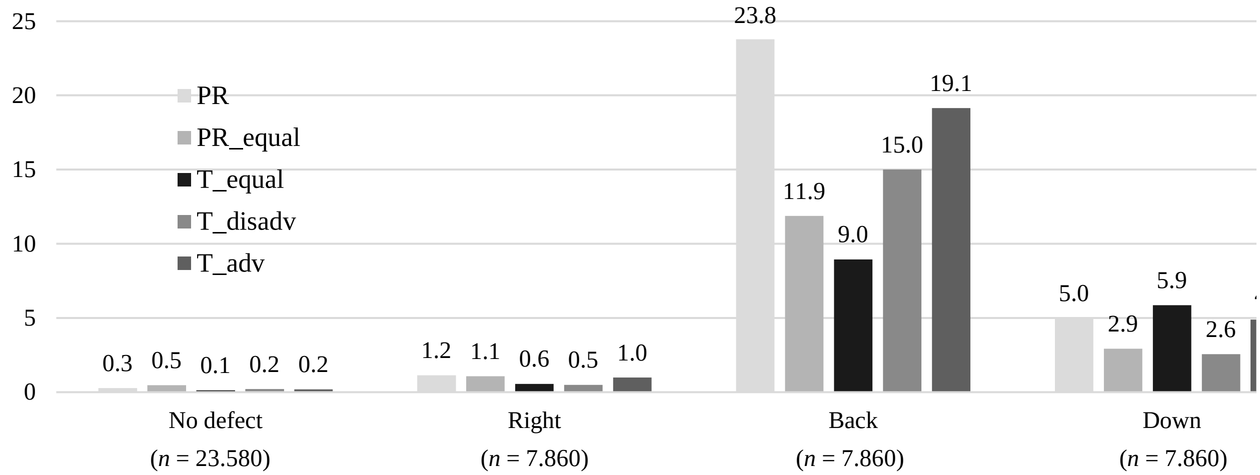

Figure 5: Percentage of errors per defect type and treatment. Regarding cubes with defects on their backside only, all differences between treatments are statistically significant (Mann-Whitney-U-test, two-tailed): *PR_equal* vs. *T_disadv* (*p* = 0.012), all other pairwise differences p<0.01.

However, the difference between *T_equal* and *PR* could also be due to the peer effect, as two parameters are different between the treatments, i.e., the incentive scheme as well as the presence of an equally able avatar. Indeed, the second significant difference that we find is between *PR* and *PR_equal*. The sheer presence of an equally able avatar seems to increase the subject's performance (p = 0.000, Mann-Whitney-U, two-tailed).

However, the error rate is even statistically lower ($p = 0.0003$) in *T_equal* (9.0%) than in *PR_equal* (11.9%), indicating tournament incentives may matter on top of pure peer effects. We have to hint that observations are not independent, as always 360 cubes are sorted by one subject. This finding indicates, as hypothesized, that tournament incentives are most effective when the contestants are homogeneous in ability, although the effect seems to driven by a peer effect to a major part.

4.1 Multivariate analysis

In the following we may further take other factors into consideration which may account for error making in our setup by estimating binary probit models on the base of a single cube. Moreover, we may control for each subject's initial performance level without our treatment intervention. Since errors hardly occur in cases of non-defective cubes, we only take defected ones into account. A not-sorted cube is defined as an error (miss), which acts as the dependent variable (1 = *yes*) with a mean error rate of 0.07 for these observations. We therefore use 180 observations per subject and account for individual differences by clustering robust standard errors at the individual level. Table 2 presents marginal effects (at the means of the variables) and corresponding standard errors of the estimations.

In model (1) we only add the cube defect type in addition to the treatment dummies. The presence of an avatar decreases the probability of an error in the amount of 3.3 percentage points on average between piece rate treatments (which corresponds to a percentage difference of 0.47 given the base error rate of 0.07). Interestingly, the same size effects occur for *T_equal*. With regard to cube defect type, we use "defect down" as the reference group and confirm much lower (higher) error probabilities, if the defect is on the right hand side (at the back) – both in the amount of 6 percentage points.

In model (2) we add two additional variables: First, we control for the number of the cube (180 cubes between 1 and 360) in order to capture possible learning or inertia effects. Confirming the illustration from figure 4 above, there is obviously no trend during the treatment phase of our experiment, indicating that neither learning nor inertia play a key role or indicating that they cancel each other out. Second, we use the number of errors in the first part of the experiment as a proxy for task-specific ability. Not surprisingly, this measure is positively related to the probability of making errors in the treatment part. More importantly, adding this control does not change the effects regarding the treatment dummies. There is now a weakly significant difference of *T_disadv* compared to the piece-rate setting, with actual p-values of *T_disadv* being almost the same: 0.093 compared to 0.101 in model (1).

Subjects need more time to sort a defective cube than to check a regular one and put it back onto the conveyor belt. Therefore, we additionally control for the information whether the precedent cube had a defect in model (3) (so that we cannot capture the very first cube of each subject here). Indeed, the probability of making an error is two percentage points higher if the preceding cube also had a defect, so that there is some spillover. The results of the other variables are not affected.

In model (4) we re-estimate model (3) for the subgroup of cubes with a defect on the back side. As revealed by the subsequent estimations, detecting defects is by far the most difficult for subjects for this defect type. The results are much more pronounced for this defect type. First, treatment effect sizes are much higher for both *T_equal* and *PR_equal*. Second, the just mentioned spillover effect from a preceding cube with a defect is much more pronounced. Perhaps, subjects act hectically in this situation and do not have enough time left to check the upcoming cube carefully. The consequence then is by far the most severe if the defect is only hardly detectable.

Further analyses (not reported; available from the authors upon request) show no meaningful interaction effects between treatments and either cube number or numbers of errors during the ability check part.

Table 2: Binary probit estimations , marginal effects, dependent variable: miss defective cube (1=yes)

	(1)		**(2)**		**(3)**		**(4)**	
							miss defected cube (back)	
Treatment (reference: PR)								
PR_equal	-0.033*	(0.017)	-0.032*	(0.017)	-0.032*	(0.017)	-0.109**	(0.053)
T_equal	-0.033**	(0.016)	-0.033**	(0.016)	-0.032**	(0.015)	-0.146***	(0.048)
T_disadv	-0.028	(0.017)	-0.028*	(0.017)	-0.028*	(0.016)	-0.077	(0.054)
T_adv	-0.010	(0.016)	-0.010	(0.016)	-0.009	(0.016)	-0.037	(0.054)
Cube defect type (reference: defect down)								
Defect right	-0.062***	(0.009)	-0.062***	(0.009)	-0.060***	(0.008)		
Defect back	0.065***	(0.009)	0.065***	(0.009)	0.064***	(0.009)		
Cube number			0.000006	(0.00002)	-0.000002	(0.00002)	-0.0001**	(0.00007)
# mistakes ability check			0.019**	(0.060)	0.022***	(0.006)	0.087***	(0.022)
Preceding cube with defect					0.019***	(0.004)	0.049***	(0.009)
Observations	23,580		23,580		23,449		7,729	
Pseudo R²	0.138		0.139		0.145		0.033	

Notes: Robust standard errors clustered at the individual level in parentheses. ** p<0.01, ** p<0.05, * p<0.1.

Table 3 extends our analysis by explicitly taking the intermediate (absolute) score difference between the subject and the avatar into account. Hence, we have to abstain from the observations of the piece-rate treatment without the avatar here. In model (1) we just re-estimate model (3) of Table 2 to check that the results with respect to hitherto variables hold. This is the case.

Model (2) then reveals that the difference in intermediate score is indeed (weakly) significantly related to the probability of making an error. An additional score difference of 11 points leads to an error increase of one percentage point (or 14 percent). Since incentives are explicitly related to relative performance in our tournament treatments, this relation may be particularly relevant in these cases. If the winner of the tournament is practically established, there are no incentives any more, so that an increasing score difference may lead to decreasing effort. We therefore interact treatment dummies with the intermediate score difference in model (3). Indeed, the corresponding coefficients of the interactions terms have a negative sign. They are not significant on conventional levels.[8]

We cannot clearly distinguish between two possible opposing effects here. The possible relevance of an effort-enhancing effect of close intermediate scores in tournament settings may be weakened by a countervailing "choking under pressure" effect (Ariely et al. 2009). Some subjects may start to act hectically when realizing that winning the tournament depends on sorting the next cubes successfully. We address this issue in more detail in the following subsection by making use of another advantage of our set-up and taking information about subjects' movements into account.

In model (4), we disentangle possible differences between leading and lagging behind by introducing a new variable *Leading* (1 = *yes*) and interact this variable with the absolute intermediate score difference to the avatar. The statistically significant interaction effect reveals that the error-enhancing effect of score difference from the avatar is particularly relevant for situations of lagging behind. This may indicate cases of choking under pressure (exerting effort without increasing output) or resigning (decreasing effort).

We tested whether the observed peer effect may be perceived by subjects as a kind of implicit tournament without monetary incentives. If that is the case, the peer effect should be stronger for highly competitive subjects than for low competitive subjects[9]. However, a binary probit estimation revealed no significant interaction of the treatments and the competitiveness dummy.

[8] However, it is close to being weakly significant, as we find p = 0.106 for *T_disadv * Difference to avatar*.

[9] We used the seven-point Likert item "I wanted to beat the avatar in part II." from the concluding questionnaire in order to distinguish highly (6 or 7 on the Likert scale) and low (lower than 6 on the Likert scale) competitive participants.

Table 3: Binary probit estimations, marginal effects, dependent variable: miss defective cube (1=yes).

	(1)		**(2)**		**(3)**		**(4)**	
Treatment (reference: PR_equal)								
T_equal	-0.0004	(0.015)	-0.003	(0.015)	0.007	(0.019)	0.008	(0.019)
T_disadv	0.004	(0.015)	0.006	(0.015)	0.022	(0.020)	0.029	(0.021)
T_adv	0.020	(0.015)	0.012	(0.016)	0.021	(0.021)	0.020	(0.021)
Cube defect type (reference: defect down)								
Defect right	-0.057***	(0.009)	-0.057***	(0.009)	-0.057***	(0.009)	-0.057***	(0.009)
Defect back	0.054***	(0.009)	0.054***	(0.009)	0.053***	(0.009)	0.053***	(0.009)
Cube number	-0.000008	(0.00002)	-0.00005*	(0.00003)	-0.00005*	(0.00003)	-0.000009	(0.00003)
# mistakes ability check	0.024***	(0.006)	0.018***	(0.006)	0.022***	(0.006)	0.023***	(0.006)
Preceding cube with defect	0.019***	(0.004)	0.019***	(0.004)	0.019***	(0.004)	0.019***	(0.004)
Difference to avatar[a]			0.001*	(0.0005)	0.002*	(0.001)	0.006***	(0.002)
Leading (1=yes)							0.024*	(0.012)
Interactions								
T_equal * Difference to avatar					-0.0014	(0.0013)	-0.0013	(0.0014)
T_disadv * Difference to avatar					-0.0029	(0.0017)	-0.0044*	(0.0020)
T_adv * Difference to avatar					-0.0012	(0.0013)	-0.0011	(0.0014)
Leading * Difference to avatar							-0.0044**	(0.0022)
Observations	18,795		18,795		18,795		17,869[b]	
Pseudo R^2	0.130		0.133		0.135		0.132	

Notes: Robust standard errors clustered at the individual level in parentheses. ** p<0.01, ** p<0.05, * p<0.1.

[a] *Difference to avatar* gives the absolute difference in intermediate score to avatar. *Difference to avatar* and *Leading* refer to the score difference at two cubes proceeding the cube in the estimation.

[b] In model (4), observations with *subject's score – avatar's score* = 0 are not included.

4.2 Movement

The conveyor belt task is a real-effort task which demands both mental and physical effort. Subjects need to be focused in order to identify normal and defective cubes, they need to move their working hand while inspecting cubes from different sides, and frequently they need to step sideways along the conveyor belt to keep pace with it. This is especially the case when a subject misses one or two cubes and tries to catch up with them by hurrying after them and then hurrying back to the central position.

The subjects wear markers on the VR-glasses and on their working hand, which are tracked by cameras suspended from the ceiling of the aixCave. The head and hand positions are constantly tracked and stored every five seconds (0.2 Hz). Thus, thanks to the VR-technology, we are able to get a measure of activity, the movement of head and hand. Compared to the point score or the number of errors, which are rather output measures, this is a genuine effort measure. However, it has to be and will be analyzed whether more movement leads to a higher performance or whether more movement is the reaction to perceived underperformance, or whether the subjects intend to increase their performance by increasing their movement effort (“being hectic”) and thereby, unfortunately, reducing their output.

Figure 7 shows the head positions of two subjects as seen from above[10]. It becomes obvious how they entered/left the aixCave through the door, which is located at the bottom left corner of the groundplan, and then started to work (and move) along the conveyor belt. The conveyor belt is positioned on the horizontal zero line. Subject (a) is one of the top five performers. She moves much less than subject (b), who is one of the worst five performers. Perhaps, this indicates that less movement leads to better results. We will specify this later.

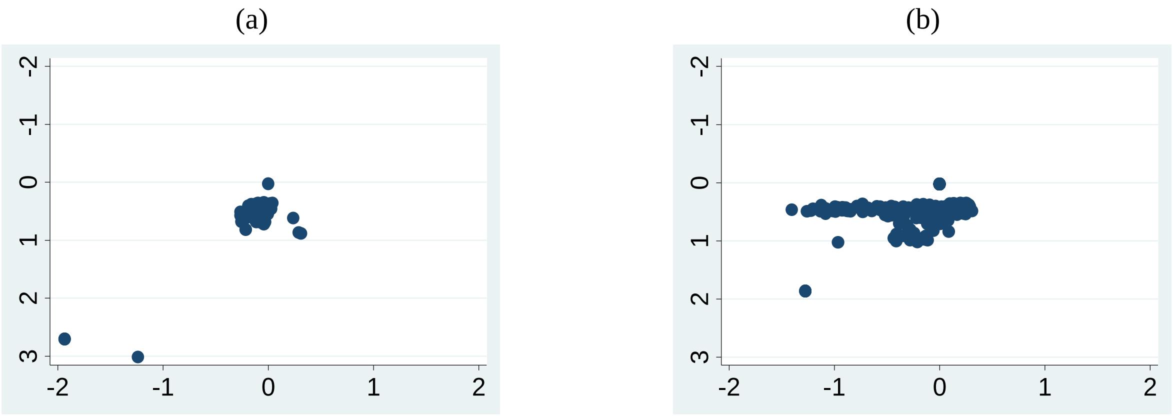

Figure 7: Head positions of two subjects. Subject (a) is one of the top five performers, subject (b) is one of the worst five performers.

[10] We have the possibility to track hand positions in the same way as head positions. In the analysis, we focus on head positions because they better represent the position of the whole body.

The movement data was tracked on a constant five-second base. Hence, we do not have the position information for each single cube but several cubes fall into a five-second interval. As we know which cubes were grabbed at the moment that the time stamp hit, we merged movement data and performance data on an aggregated level[11]. Thus, merged performance and movement data only exist on the aggregated level of a sequence.

Subjects' degrees of movement were calculated as Euclidean distances between the tracked head positions. Their movement patterns were quite different. During part II (the treatment part), the "most active" subject covered a total distance of 55.9 meters while the "most static" subject covered a distance of only 2.8 meters. The average total covered distance in part II per subject is 12.6 meters (SD = 8.4).

The total distance is not significantly different between the five treatments (Figure 8). However, distance correlates with the number of errors with $r = 0.25$ ($p < 0.001$) among all observations, indicating that more movement tends to reduce output.

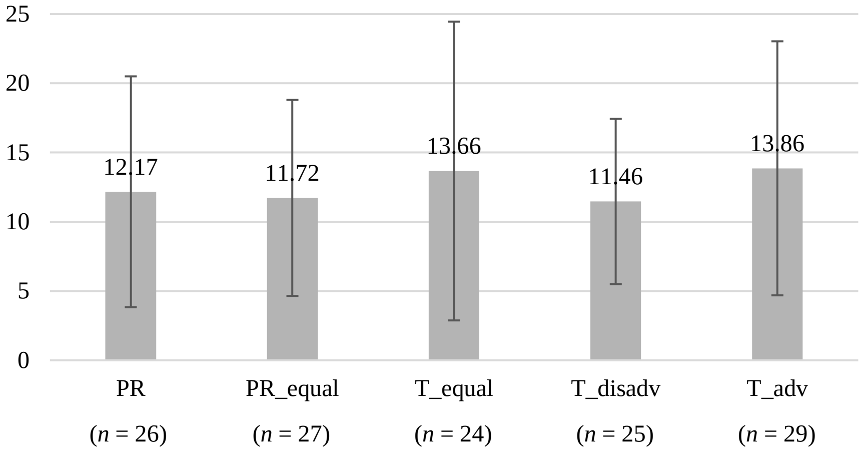

Figure 8: Average total distance in meters covered by subjects in part II per treatment. Differences not statistically significant (Mann-Whitney-U).

[11] One experimental round takes ten minutes and contains 360 cubes, while the subject's head and hand positions are tracked by 120 time stamps per round. Thus, one round can be divided into ten one-minute sequences containing 36 cubes and 12 time stamps. Starting with the last grabbed cube which was captured by the time stamp, we counted the preceding 120 time stamps and can thereby quite accurately identify the time stamp at which the first sequence starts.

Table 4: Binary Tobit estimations , β coefficients, dependent variable: miss defective cube (1=yes)

	(1)		**(2)**		**(3)**		**(4)**		**(5)**		**(6)**	
					PR_equal		***T_equal***		***T_adv***		***T_disadv***	
Treatment (reference: T_equal)												
PR equal	0.536	(0.634)	0.522	(0.639)								
T disad	0.834	(0.691)	0.823	(0.689)								
T adv	0.841	(0.703)	0.827	(0.705)								
Sequence	-0.081	(0.056)	-0.082	(0.558)	-0.091	(0.148)	-0.074	(0.073)	-0.151	(0.159)	-0.103	(0.101)
Difference from avatar	0.053**	(0.025)	0.041	(0.03)	0.125	(0.078)	0.032	(0.065)	0.025	(0.056)	-0.198	(0.137)
Movement	0.823***	(0.18)	0.727***	(0.249)	1.625***	(0.575)	0.881	(0.541)	-0.070	(0.463)	-0.320	(0.701)
Difference from avatar * Movement			0.009	(0.013)	-0.051	(0.049)	0.021	(0.045)	0.030*	(0.016)	0.211**	(0.082)
cons	-1.932***	(0.571)	-1.787	(0.63)	-2.229**	(0.999)	-1.790**	(0.898)	0.235	(1.076)	0.342	(1.224)
Pseudo R^2	0.028		0.029		0.040		0.077		0.017		0.012	
n	1050		1050		250		270		290		240	
left-censored n	579		579		143		159		140		137	
uncensored n	471		471		107		111		150		103	

Notes: Robust standard errors clustered at the individual level in parentheses. ** p<0.01, ** p<0.05, * p<0.1.

Again, we complement our analyses with corresponding multivariate estimation. We now have to rely on the sequence level because of our movement measure (10 observations per subject). Since there is a relevant fraction of sequences without errors, we apply Tobit estimations (see Table 4). The results confirm that movements increase the number of errors (model 1). There is no interaction effect between the effect of movements and absolute point difference from the avatar among all treatments, though (model 2).[12]

However, there may be treatment differences with regard to the latter interaction effect, since the point difference from the avatar does matter for payoffs in tournaments, but not in the piece-rate treatment. Therefore, we present treatment-wise estimations in models (3) to (6). Indeed, there are some differences. The interaction coefficient for *PR_equal* is negative and not significant. In contrast, the coefficient of the interaction effect has a positive sign for all three tournament treatments and is even statistically significant for *T_adv* and *T_disadv*. Apparently, increased moving leads to more errors for an increasing point difference in particular, when we exogenously match the subject with a better avatar (*T_disadv*). This leads to a higher number of cases in which the subject is lagging behind the avatar's score and may act more hectically, showing some kind of choking under pressure behavior. On the other hand, in *T_adv* where the subject is far ahead, the avatar in point score and the advantage is growing over time almost linearly, the subject may tend to slack off as she is very unlikely to lose the tournament. Thus, errors become more likely.

On average, increased movement leads to more errors. However, when looking at the individual patterns of movement over time[13] we found that the subjects show very different behavioral patterns. Thus, different behavioral patterns (consciously or unconsciously) exist. In order to identify subgroups of subjects showing similar behavioral patterns, we take a look at the association between movement and performance on an individual level and whether these patterns systematically vary with our treatment manipulation. In figure 9, the total number of errors per subject are plotted against the total distance covered by the subjects in part II. Each dot corresponds to one subject. The median number of errors (8) as well as the median distance (10.5) are highlighted. The faint positive correlation becomes obvious in all treatments, but the distribution of subjects over the four quadrants generated by the orthogonal median lines varies.

[12] As a robustness check we split the sample along the median of the individual performance during the ability check in low performing and high performing subjects (subjects with median performance were included in the high performing subsample). We re-estimated model (1) of table 4 separately for high performers and low performers and find a significant positive effect of *PR_equal* for the high performers. However, tournament incentives do not seem to affect subjects with heterogeneous ability differently, while high performing and low performing subjects react differently to differences in the point score. We find a highly significant negative effect of point difference on the number of missed cubes as well as a highly significant interaction of point difference and movement. The larger the (absolute) point difference, the fewer errors are made by the low performers. Low performers, thus, seem to fall for choking under pressure more easily than high performers by (unproductively) increasing their movement.

[13] Detailed graphs are available on request

Subjects in the bottom right quadrant move a lot and perform above the median. We could call them "active high performers". Subjects in the bottom left quadrant are relatively economic with their activity and achieve very high scores (very few errors). We could name these subjects "efficient high performers", while subjects in the upper left quadrant are "reflective low performers". They economize on their movements and achieve above median error rates. Subjects in the upper right quadrant are active but fail to turn activity into achievement. We call these subjects "hectic low performers".

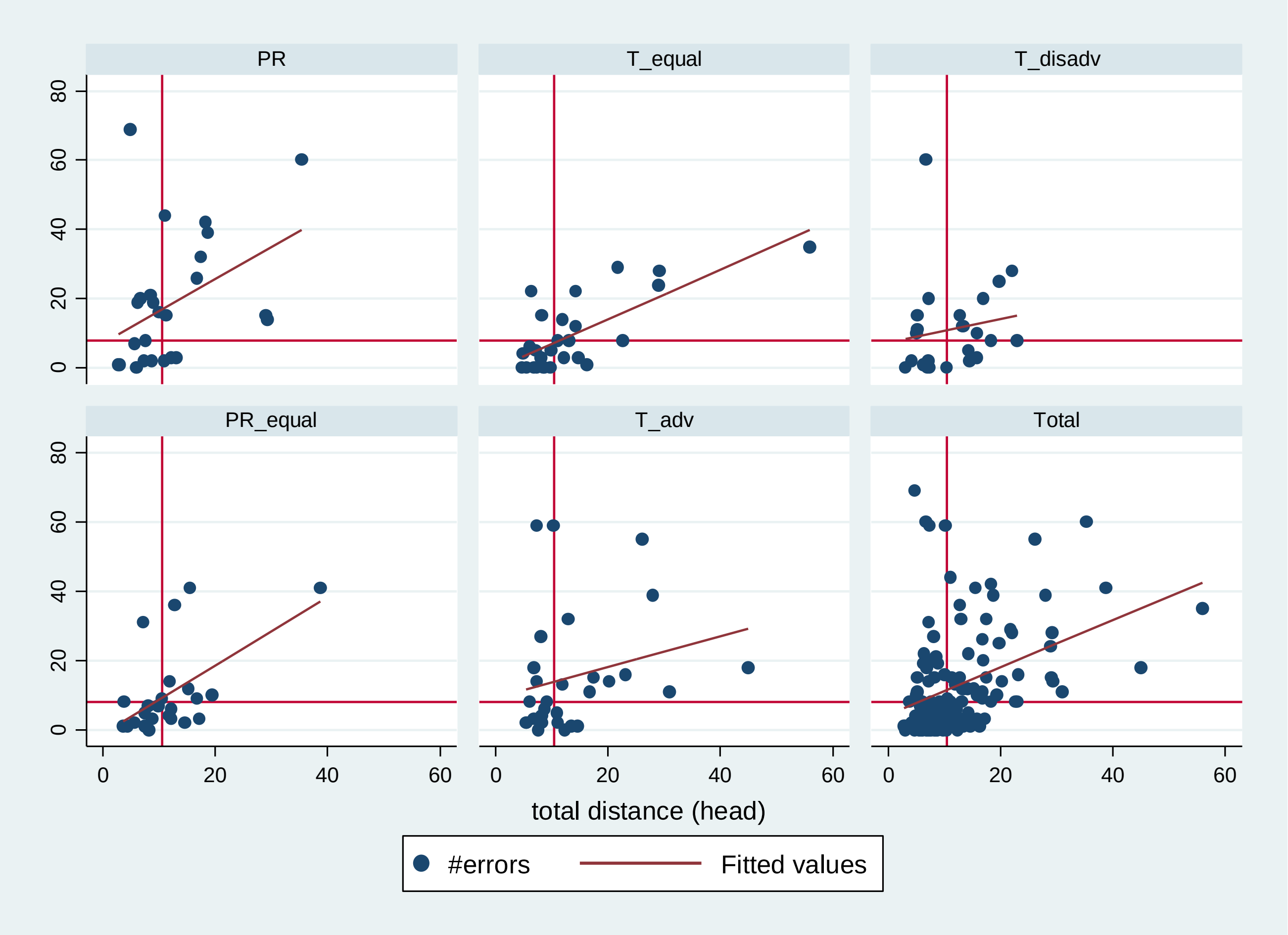

Figure 9: Scatterplot and linear fitted values of the total number of errors per subject and the total distance covered by the subjects in part II. Each dot corresponds to one subject. Medians highlighted.

In the scatterplots above it becomes obvious that the dispersion of dots (subjects) across the four quadrants (behavioral patterns) differ to some extent across treatments. For example, in *PR* – the treatment with the highest overall error rate (see figure 3) – the point cloud is relatively widely spread over the left half of the coordinate system. While in *T_equal*, the treatment with the lowest error rate, the point cloud is much more compressed in the bottom left quadrant and at the bottom left edges of the other quadrants (except for one outlier). Apparently, the good performance of the subjects in *T_equal* is at least partially rooted in their ability not to become too hectic.

5. Discussion

In this study we experimentally analyze the effect of heterogeneity in tournaments, which has already been captured by the seminal model of Lazear and Rosen (1981), predicting that the effort provided by subjects in a tournament is highest when the subjects are homogeneous in ability. We tested this prediction in an innovative experimental design in Virtual Reality (VR). Doing this we overcome the reflection problem although we focus on a highly dynamic context where participants may continuously observe each other.

Compared to classical lab experiments, the virtual environment provides a more realistic and completely controllable setting. Here, we used the Gürerk et al. (2016b) virtual production hall with a virtual conveyor belt, and a virtual co-worker (the avatar). VR allows us to control the avatar's ability precisely, and yet it provides a really dynamic setting in which the subjects are permanently able to observe their avatar's performance on a score board. As only the human subject reacts towards the avatar, our technology enables us to analyze individual behavior in a dynamic but highly controlled setting. Thereby, we are able *i*) to disentangle to which extent tournament incentives trace back to peer effects by combining tournaments with contestants of systematically heterogeneous ability and piece-rate pay schemes with and without a virtual co-worker and *ii*) mitigate the reflection problem (Manski 1993, section 2.3). To the best of our knowledge, ours is the first study to address the interplay of tournament incentives and peer effects.

At first sight, our results confirm the theoretical prediction that particularly tournaments among equally able contestants induce higher efforts than piece-rate pay schemes at least when interpreting the number of errors made as an inverse measure of effort. However, when taking into account the basic ability to do the real-effort task in the multiple regressions, we find that also the piece-rate scheme in which an equally able avatar can be observed while working on the same task induces significantly fewer errors compared to a piece-rate scheme without the presence of an avatar. Thus, our result seems to be mainly driven by a peer effect. This result is also confirmed by taking into account that the most difficult defects are most frequently detected in the symmetric tournament as well as in the piece-rate setting with an equally able avatar, thus, indicating that participants work most thoroughly in settings in which they may observe an equally able subject. Comparing the three tournament treatments, we find that *T_equal* and *T_disadv* are the most effective treatments. Given the performance increase from part I to part II, the actual degree of heterogeneity in *T_disadv* is considerably lower than in *T_adv* so that *T_disdav* can be considered as an almost symmetric setting. Also in reality one would always be forced to base the matching of contestants on performance measures of the past as we did in our setting. We have to reconsider, though, that there is no reason to doubt that all persons would show performance increases so that previous performance differences would persist. In sum, our results hint for the advantageousness of tournaments with rather homogeneous contestants.

The use of VR, moreover, enables us to implement a real *effort* measure – the degree and pattern of movement – besides the reached score or the number of errors, while the latter ones are rather *output* measures. We find interindividual differences in the subjects' movements. The distance that they cover during the experiment varies dramatically between subjects as well as the individual movement patterns. Some subjects stay static over the whole experiment, others are constantly hectic, others seem to learn over time that small movements are more effective than fast movements, and yet others seem to try to catch up lags in score by moving faster towards the end of the experiment. We find a negative correlation between the amount of movements and performance. However, we find no direct association of the total movement measures with either the treatments or with the score difference between subject and avatar. Thus, it seems to be rather a matter of the subject's characteristics whether she increases movement in order to (falsely) enhance performance or whether she learns that small and controlled movements are more effective and reacts to score differences with less hectic work.

Our approach of using this innovative technology to conduct an economic experiment enables us to analyze individual behavior in a highly dynamic but controlled setting. For future studies the virtual environment can easily be used to study other relevant economic questions that are difficult to simulate in traditional laboratory experiments. For example, our real-effort setting with competing agents could be extended by helping and sabotage activities which can be integrated into the task as a more natural option compared to previous lab experiments. In those former experiments on sabotage participants had to be explicitly informed that they may harm others (see e.g. Carpenter, Matthews and Schirm 2010, Harbring and Irlenbusch 2011, Balafoutas, Lindner and Sutter 2012, Charness, Masclet and Villeval 2013).

Interestingly, we may show in our study that – independent of the monetary incentive scheme – the effect of observing an equally able worker drives up performance compared to working alone under a simple piece-rate scheme. The peer effect that we observe occurs in a highly transparent setting in which participants may continuously observe the contestant's performance. Comparing the tournament settings one may conclude that matching a subject to a similarly well or better performing individual seems superior to matching someone to a weaker contestant. Our results add to the rich empirical literature on the effectiveness of incentive schemes. Moreover, our approach – being one of the first studies to use a virtual environment for economic research – is believed to reveal some new methodological avenues for economic experimentation in the (near) future

6. References

Anderson, L. R., B. A. Freeborn. 2010. Varying the intensity of competition in a multiple prize rent seeking experiment. *Public Choice* **143**(1/2) 237–254.

Anderson, L. R., S. L. Stafford. 2003. An experimental analysis of rent seeking under varying competitive conditions. *Public Choice* **115** 199–216.

Andreoni, J. 1988. Why free ride?: Strategies and learning in public goods experiments. *Journal of Public Economics*.

Archer, B. 2017. Roger Federer reveals how Rafael Nadal was so important for his career Grand Slam record, https://www.express.co.uk/sport/tennis/862138/Roger-Federer-Rafale-Nadal-motivation-Grand-Slam-record.

Ariely, D., U. Gneezy, G. Loewenstein, N. Mazar. 2009. Large Stakes and Big Mistakes. *Review of Economic Studies* **76**(2) 451–469.

Bach, N., O. Gürtler, J. Prinz. 2009. Incentive effects in tournaments with heterogeneous competitors: An analysis of the Olympic Rowing Regatta in Sydney 2000. *Management Revue* **20**(3) 239–253.

Bailenson, J. N., J. Blascovich, A. C. Beall, J. M. Loomis. 2003. Interpersonal distance in immersive virtual environments. *Personality & Social Psychology Bulletin* **29**(7) 819–833.

Balafoutas, L., F. Lindner, M. Sutter. 2012. Sabotage in Tournaments: Evidence from a Natural Experiment. *Kyklos* **65**(4) 425–441.

Bandiera, O., I. Barankay, I. Rasul. 2010. Social Incentives in the Workplace. *Review of Economic Studies* **77**(2) 417–458.

Berger, J., P. Nieken. 2016. Heterogeneous Contestants and the Intensity of Tournaments: An Empirical Investigation. *Journal of Sports Economics* **17**(7) 631–660.

Beugnot, J., B. Fortin, G. Lacroix, M. C. Villeval. 2013. Social Networks and Peer Effects at Work. *IZA Discussion Papers*(7521).

Brockner, J. 1987. Survivors' reactions to layoffs: We get by with a little help for our friends. *Administrative Science Quarterly*.

Brown, J. 2011. Quitters never win: The (adverse) incentive effects of competing with superstars. *Journal of Political Economy* **119**(5) 982–1013.

Bull, C., A. Schotter, K. Weigelt. 1987. Tournaments and piece rates: An experimental study. *Journal of Political Economy* **95**(1) 1–33.

Carpenter, J., P. H. Matthews, J. Schirm. 2010. Tournaments and Office Politics: Evidence from a Real Effort Experiment. *American Economic Review* **100**(1) 504–517.

Casas-Arce, P., F. A. Martínez-Jerez. 2009. Relative performance compensation, contests, and dynamic incentives. *Management Science* **55**(8) 1306–1320.

Cason, T. N., W. A. Masters, R. M. Sheremeta. 2010. Entry into winner-take-all and proportional-prize contests: An experimental study. *Journal of Public Economics* **94**(9-10) 604–611.

Chan, D. C. 2016. Teamwork and Moral Hazard: Evidence from the emergency department. *Journal of Political Economy* **124**(3) 734–770.

Chan, T., J. Li, L. Pierce. 2014. Compensation and peer effects in competing sales teams. *Management Science* **60**(8) 1965–1984.

Charness, G., D. Masclet, M. C. Villeval. 2013. The Dark Side of Competition for Status. *Management Science* **60**(1) 38–55.

Cruz-Neira, C., D. J. Sandin, DeFanti, T. A. 1993. Surround-screen projection-based virtual reality: the design and implementation of the CAVE: The design and implementation of the CAVE. Mary C. Whitton, ed. *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*. Association for Computing Machinery, 135–142.

Davis, D. D., R. J. Reilly. 1998. Do too many cooks always spoil the stew?: An experimental analysis of rent-seeking and the role of a strategic buyer. *Public Choice* **95** 89–115.

Dechenaux, E., D. Kovenock, R. M. Sheremeta. 2015. A survey of experimental research on contests, all-pay auctions and tournaments. *Exp Econ* **18**(4) 609–669.

Deutscher, C., B. Frick, O. Gürtler, J. Prinz. 2013. Sabotage in Tournaments with Heterogeneous Contestants: Empirical Evidence from the Soccer Pitch. *Scandinavian Journal of Economics* **115**(4) 1138–1157.

Deutscher, C., S. Schneemann. 2017. The Impact of Intermediate Information on Sabotage in Tournaments with Heterogeneous Contestants. *Managerial and Decision Economics* **38**(2) 222–237.

Eisenkopf, G. 2010. Peer effects, motivation, and learning. *Economics of Education Review* **29**(3) 364–374.

England, D. E., Buskist, W. 1995. The Effects of Instructions on Subjects' Disclosure of Information about Operant Tasks. *The Psychological Record* **45**(3) 451–461.

Eriksen, K. W., O. Kvaløy, T. E. Olsen. 2011. Tournaments with Prize-setting Agents. *Scandinavian Journal of Economics* **113**(3) 729-753.

Falk, A., A. Ichino. 2006. Clean evidence on peer effects. *Journal of Labor Economics* **24**(1) 39–57.

Fallucchi, F., E. Renner, M. Sefton. 2013. Information feedback and contest structure in rent-seeking games. *European Economic Review* **64** 223–240.

Fehr, D., J. Schmid. 2011. *Exclusion in the all-pay auction*: *An experimental investigation.* Discussion paper / Social Science Center Berlin (WZB), Research Area Markets and Choice, Research Unit Market Behavior, Berlin.

Fina. 2018. Database on athletes and records, https://www.fina.org. Retrieved January 18, 2018.

Fonseca, M. A. 2009. An Experimental Investigation of Asymmetric Contests. *International Journal of Industrial Organization* **27**(5) 582–591.

Georganas, S., M. Tonin, M. Vlassopoulos. 2015. Peer pressure and productivity: The role of observing and being observed. *Journal of Economic Behavior and Organization* **117** 223–232.

Gill, D., V. Prowse. 2012. A Structural Analysis of Disappointment Aversion in a Real Effort Competition. *American Economic Review* **102**(1) 469–503.

Grund, C., J. Höcker, S. Zimmermann. 2013. Incidence and consequences of risk-taking behavior in tournaments: Evidence from the NBA. *Economic Inquiry* **51**(2) 1489–1501.

Gürerk, Ö., A. Bönsch, L. Braun, C. Grund, C. Harbring, T. Kittsteiner, A. Staffeldt. 2016a. *Experimental Economics in Virtual Reality*. MPRA 71409.

Gürerk, Ö., T. Kittsteiner, A. Bönsch, A. Staffeldt. 2016b. *Avatars as peers at work*: *An experimental study in virtual reality*.

Gürtler, M., O. Gürtler. 2015. The Optimality of Heterogeneous Tournaments. *Journal of Labor Economics* **33**(4) 1007–1042.

Gürtler, O., C. Harbring. 2010. Feedback in Tournaments under Commitment Problems: Experimental Evidence. *Journal of Economics and Management Strategy* **19**(3) 771–810.

Gürtler, O., M. Kräkel. 2010. Optimal Tournament Contracts for Heterogeneous Workers. *Journal of Economic Behavior and Organization* **75**(2) 180–191.

Hammond, R. G., X. Zheng. 2013. Heterogeneity in Tournaments with Incomplete Information: An Experimental Analysis. *International Journal of Industrial Organization* **31**(3) 248–260.

Harbring, C., B. Irlenbusch. 2011. Sabotage in Tournaments: Evidence from a Laboratory Experiment. *Management Science* **57**(4) 611–627.

Harbring, C., B. Irlenbusch, M. Kräkel, R. Selten. 2007. Sabotage in Corporate Contests: An Experimental Analysis. *International Journal of the Economics of Business* **14**(3) 367–392.

Harbring, C., G. K. Lünser. 2008. On the Competition of Asymmetric Agents. *GER* **9**(3) 373–395.

Harbring, C., G. K. Ruchala. 2003. *Separating the Wheat from the Chaff*: *An Experimental Study on the Sorting Function of Tournaments*. GEABA Discussion Paper 03-13, German Economic Association of Business Administration e. V.

Herbst, D., A. Mas. 2015. Peer effects on worker output in the laboratory generalize to the field. *Science* **350**(6260) 545–549.

Heyselaar, E., P. Hagoort, K. Segaert. 2017. In dialogue with an avatar, language behavior is identical to dialogue with a human partner. *Behavior Research Methods* **49**(1) 46–60.

Iachini, T., Y. Coello, F. Frassinetti, V. P. Senese, F. Galante, G. Ruggiero. 2016. Peripersonal and interpersonal space in virtual and real environments: Effects of gender and age. *Journal of Environmental Psychology* **45** 154–164.

Ijzerman, H., W. W. van Dijk, M. Gallucci. 2007. A bumpy train ride: a field experiment on insult, honor, and emotional reactions. *Emotion* **7**(4) 869–875.

Innocenti, A. 2017. Virtual Reality Experiments in Economics. *Journal of Behavioral and Experimental Economics* **69** 71–77.

Katz, L. F., J. R. Kling, J. B. Liebman. 2001. Moving to Opportunity in Boston: Early results of a randomized mobility experiment. *The Quarterly Journal of Economics* **116**(2) 607–654.

Kimbrough, E. O., R. M. Sheremeta, T. W. Shields. 2014. When parity promotes peace: Resolving conflict between asymmetric agents. *Journal of Economic Behavior and Organization* **99** 96–108.

Kocher, M., M. V. Lenz, M. Sutter. 2012. Psychological Pressure in Competitive Environments: New evidence from randomized natural experiments. *Management Science* **58**(8) 1585–1591.

Kothgassner, O. D., A. Felnhofer, H. Hlavacs, L. Beutl, R. Palme, I. Kryspin-Exner, L. M. Glenk. 2016. Salivary cortisol and cardiovascular reactivity to a public speaking task in a virtual and real-life environment. *Computers in Human Behavior* **62** 124–135.

Kräkel, M. 2008. Emotions and the Optimality of Uneven Tournaments. *Review of Managerial Science* **2**(1) 61–79.

Kräkel, M., D. Sliwka. 2004. Risk taking in asymmetric tournaments. *GER* **5**(1) 103–116.

Kuhlen, T. 2014. Virtuelle Realität als Gegenstand und Werkzeug der Wissenschaft. S. Jeschke, L. Kobbelt, A. Dröge, eds. *Exploring Virtuality*: *Virtualität im interdisziplinären Diskurs*. Springer Spektrum, Wiesbaden, 133–147.

Kuhnen, C. M., A. Tymula. 2012. Feedback, Self-Esteem, and Performance in Organizations. *Management Science* **58**(1) 94–113.

Lakin, J. L., T. L. Chartrand. 2003. Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Science* **14**(4) 334–339.

Lazear, E. P., S. Rosen. 1981. Rank-Order Tournaments as Optimum Labor Contracts. *Journal of Political Economy* **89**(5) 841–864.

Ludwig, S., G. K. Lünser. 2012. Observing your competitor – The role of effort information in two-stage tournaments. *Journal of Economic Psychology* **33**(1) 166–182.

Mago, S. D., A. C. Samak, R. M. Sheremeta. 2016. Facing Your Opponents. *Journal of Conflict Resolution* **60**(3) 459–481.

Manski, C. F. 1993. Identification of endogenous social effects: The Reflection Problem. *Review of Economic Studies*(204, S. 531-542).

Mas, A., E. Moretti. 2009. Peers at work. *American Economic Review* **99**(1) 112–145.

McLaughlin, K. J., R. G. Ehrenberg. 1988. Aspects of Tournament Models: A Survey. *Research in Labor Economics* **9**(1) 225–256.

Milgram, S. 1963. Behavioral Study of Obedience. *The Journal of Abnormal and Social Psychology* **67**(4) 371–378.

Moussaïd, M., M. Kapadia, T. Thrash, R. W. Sumner, M. Gross, D. Helbing, C. Hölscher. 2016. Crowd behaviour during high-stress evacuations in an immersive virtual environment. *Journal of the Royal Society Interface* **13** 20160414.

Müller, W., A. Schotter. 2010. Workaholics and Dropouts in Organizations. *Journal of the European Economic Association* **8**(4) 717–743.

Niederle, M., L. Vesterlund. 2007. Do Women Shy Away From Competition? Do Men Compete Too Much? *Quarterly Journal of Economics* **122**(3) 1067–1101.

Nieken, P., M. Stegh. 2010. Incentive effects in asymmetric tournaments: Empirical evidence from the german hockey league. *SFB/TR 15 Discussion Paper*(305).

O'Keeffe, M. 1984. Economic contests: Comparative reward schemes. *Journal of Labor Economics* **2**(1) 27–56.

Orrison, A., A. Schotter, K. Weigelt. 2004. Multiperson Tournaments: An Experimental Examination. *Management Science* **50**(2) 268–279.

Pertaub, D.-P., M. Slater, C. Barker. 2002. An Experiment on Public Speaking Anxiety in Response to Three Different Types of Virtual Audience. *Presence: Teleoperators and Virtual Environments* **11**(1) 68–78.

Poeschl, S., N. Doering. 2015. Measuring Co-Presence and Social Presence in Virtual Environments - Psychometric Construction of a German Scale for a Fear of Public Speaking Scenario. *Studies in Health Technology and Informatics* **219** 58–63.

Sacerdote, B. 2001. Peer Effects with Random Assignment: Results for Dartmouth Roommates. *Quarterly Journal of Economics* **116**(2) 681–704.

Sausgruber, R. 2009. A note on peer effects between teams. *Exp Econ* **12**(2) 193–201.

Schotter, A., K. Weigelt. 1992. Asymmetric tournaments, equal opportunity laws, and affirmative action: Some experimental results. *The Quarterly Journal of Economics*(429, S. 511-539).

Selten, R., R. Stoecker. 1986. End behavior in sequences of finite Prisoner's Dilemma supergames A learning theory approach. *Journal of Economic Behavior and Organization* **7**(1) 47–70.

Slater, M., A. Antley, A. Davison, D. Swapp, C. Guger, C. Barker, N. Pistrang, M. V. Sanchez-Vives. 2006. A virtual reprise of the Stanley Milgram obedience experiments. *PloS one* **1** e39.

Slater, M., P. Khanna, J. Mortensen, I. Yu. 2009. Visual realism enhances realistic response in an immersive virtual environment. *IEEE computer graphics and applications* **29**(3) 76–84.

So, T., P. Brown, A. Chaudhuri, D. Ryvkin, L. Cameron. 2017. Piece-rates and tournaments: Implications for learning in a cognitively challenging task. *Journal of Economic Behavior and Organization* **142** 11–23.

Sunde, U. 2009. Heterogeneity and performance in tournaments: a test for incentive effects using professional tennis data. *Applied Economics* **41**(25) 3199–3208.

Székely, G., R. M. Satava. 1999. Virtual Reality in Medicine. *British Medical Journal* **319**(7220) 1305.

Thöni, C., S. Gächter. 2015. Peer effects and social preferences in voluntary cooperation: A theoretical and experimental analysis. *Journal of Economic Psychology* **48** 72–88.

Topa, G. 2001. Social interactions, local spillovers and unemployment. *Review of Economic Studies* **68**(2) 261–295.

van Dijk, F., J. Sonnemans, F. A. A. M. van Winden. 2001. Incentive Systems in a Real Effort Experiment. *European Economic Review* **45**(2) 187–214.

van Kleef, G. A., A. C. Homan, B. Beersma, D. van Knippenberg, B. van Knippenberg. 2009. Searing sentiment or cold calculation?: The effects of leader emotional displays on team performance depend on follower epistemic motivation. *AMJ* **52**(3) 562–580.

van Schie, H. T., B. M. van Waterschoot, H. Bekkering. 2008. Understanding action beyond imitation: reversed compatibility effects of action observation in imitation and joint action. *Journal of Experimental Psychology: Human Perception and Performance* **34**(6) 1493–1500.

van Veldhuizen, R., H. Oosterbeek, J. Sonnemans. 2018. Peers at work: Evidence from the lab. *PloS one* **13**(2) e0192038.

Waldinger, F. 2012. Peer Effects in Science: Evidence from the Dismissal of Scientists in Nazi Germany. *Quarterly Journal of Economics* **79**(2) 838–861.