Avatars as Peers at Work: An Experimental Study in Virtual Reality

Özgür Gürerk* Thomas Kittsteiner† Andrea Bönsch‡ Andreas Staffeldt†

September 22, 2016

Abstract

Identification of peer effects is often complicated by the *reflection problem*: Does agent i influence agent j, or vice versa? To be able to identify a clear causality, we embed a virtual human (avatar) as co-worker of a human subject into an *immersive virtual environment*. We observe that low productive human subjects increase their work performance more when they observe a low productive avatar – compared to a high productive avatar. This result is in line with the predictions of the social comparison theory, in as much as we observe stronger peer effects when the perceived similarity in abilities between the peers is high.

Keywords: peer effects, real effort, virtual reality, virtual human, reflection problem

1 Introduction

Understanding peer effects is of utmost importance for the organization of workers in the workplace and the composition of teams. Does a "bad apple" have a negative influence on her peers' performances? Does a high performer induce fellow workers to high effort levels? The answers to these questions have many practical implications for the organization of labor and the economic consequences of team composition, such as whether shifts should be composed of workers with homogenous or heterogeneous abilities.

In this context, the contribution of this paper is twofold: First, by means of a controlled lab experiment, we investigate whether and how a worker's performance is influenced by the presence of another worker (in the same room), who performs an identical but otherwise unrelated task. Second, we introduce and apply a new experimental methodology for testing

*RWTH Aachen University, School of Business and Economics, corresponding author, e-mail: ozgur.gurerk@rwth-aachen.de

†RWTH Aachen University, School of Business and Economics

‡RWTH Aachen University, Visual Computing Institute

peer effects. We create a virtual work environment, where subjects work together with other workers who are (programed) virtual humans (avatars).[1] Subjects work in a realistic environment conducting real effort tasks while at the same time the experimenter has control over the actions of (virtual) peers. This allows identification in a lifelike environemnt. Furthermore, our methodology allows us to collect tracking data in the virtual environment, in particular the position and orientation of a subject's head and hand(s). This data allows us to refine our findings as it can be used to distinguish subjects who are working more carefully, from others whose productivity is relatively more affected by chance rather than by skill when working at the sorting task (see Section 6.4).

More specifically, we investigate how – if at all – a human subject adjusts her own performance in a simple individual (real effort) sorting task to the performance of the observable virtual human (her peer), who simultaneously performs the same task in the same virtual production hall. Tasks are independent, i.e., output is independent of the peer's performance and its output, and a worker's payoff is output independent. In addition, we investigate whether human subjects perceive their respective virtual peer as manlike.

To address these questions, and to control for heterogeneity in ability, we design an experiment consisting of two stages. In the first phase of the experiment, to which we refer as the ability-check phase, subjects perform the sorting task in the absence of a peer. Then, in the second phase, to which we refer as the treatment phase, they perform the task in the presence of an avatar (a peer). In two treatments, we model the avatar either as a low or a high productivity worker. We refer to these treatments as the SLOW and the FAST treatment, respectively.

We find that human subjects with low ability, i.e., subjects with low productivity in the ability-check phase, increase their productivity more in the treatment phase (relative to their productivity in the ability-check phase) when they are working in the SLOW treatment, in the presence of a low productive avatar – as compared to when they are working in the FAST treatment, with a highly productive avatar (see Section 6.2 and Section 6.3). This finding is consistent with social comparison theory and the assumption that subjects hold perceptions of avatars as others to whom they compare themselves. This is because the presence of an avatar with a low productivity (rather than high productivity) makes it easier for a low-productivity subject to outperform the avatar by working harder. This finding cannot be explained by standard incentive theory, since the monetary incentive is the same between treatments. Also implicit learning cannot explain this finding, as the productivity increase is not larger in the FAST treatment with the more productive avatars.

Based on a post-experimental survey, we find that most subjects considered the display

[1] While a virtual human is a computer-generated and programed model looking and acting like a human, an avatar is a representation of a human in a virtual environment. We ignore this subtle difference in this paper and use the terms virtual human and avatar interchangeably.

of the avatar and the environment as natural (see Section 6.1). Subjects evaluated the highly productive virtual human indeed as the more productive one – compared to the low productive avatar. This means, subjects perceived the difference in the performances of both avatars exactly as intended by us. Furthermore, a significant share of subjects claimed that they wanted to beat the avatar, which we interpret as further evidence that subjects acknowledge and react to the presence of avatars.

1.1 The Reflection Problem and How we Handle It

A major challenge in the empirical as well as the experimental research on peer effects is the reflection problem (Manski, 1993). Does agent i influence agent j, or vice versa? Similar to a person in a mirror chamber who cannot determine the causality of reflections she is seeing, simultaneous observation of peers in a group does not reveal who is influencing whom. This reflection problem severely hampers identification, if data is collected from real life or close to real life working environments, where, in principle, information is generated and received by subjects simultaneously. Identification is further complicated by the problem that (at least some) information between subjects flows freely and cannot be controlled. This is the case in (natural) field studies but also causes problems in lab experiments, such as in the study by Falk and Ichino (2006), where human subjects are physically located in the same room and can see each other, and each others characteristics.

So far, the reflection problem has been addressed in lab experiments by controlling the information each subject receives on others' productivity. In Van Veldhuizen, Oosterbeek, and Sonnemans (2014), a monitoring subject i is provided with productivity information on the monitored subject j, but not vice versa. Thöni and Gächter (2015) provide subjects with relevant information concerning their peers' performance only after they have provided effort themselves, and then – after having received the information on peers – allowing them to reconsider their own effort choices. Such limitations on information flow and decision making, however, add to the artificiality of the work situation and effort measurement that are typical for lab experiments.[2]

In this paper, we address these problems differently. First, we design an economic real effort experiment to be conducted in an *immersive virtual environment*. An immersive virtual environment (abbrev. as IVE) is defined as a virtual environment that perceptionally surrounds the subject (Blascovich et al., 2002). In such environments, subjects have a strong feeling of "being there" in the synthetically created situation.[3] Being physically inside this virtual environment, the human subject performs a real-effort sorting task at a virtual

[2] Virtually all lab experiments either use simple choices as a proxy for effort or "real effort" tasks that are merely cognitive. In addition, in the lab, subjects only obtain numerical information about their peers' effort choices.

[3] In Section 3, we provide more information on IVEs.

conveyor belt, interacting with virtual objects. Second, and maybe more importantly, a subject's peer is an avatar (a virtual human). The avatar, though being a (fully controlled) computer program, resembles in its movements and appearance a human, and performs the same task as the human subject. Since the avatar is not influenced by the subject, we are able to observe non-confounded causal peer effects of the virtual human's working behavior on the human subject's performance. Obviously, the potential "cost" of this approach is that it is, per se, unclear whether subjects react similarly to humans and to avatars – though here, we build and rely on a literature in social sciences and computer science that is mainly affirmative to this question (see Section 3.2).

2 Related Literature

For the sake of brevity, this literature section only covers closely related studies. It does not serve as a comprehensive review of both topics of this paper: experiments on peer effects, and experiments in immersive virtual environments. A recent survey by Herbst and Mas (2015) covers 11 peer effect studies from the lab and 23 based on field data. This study shows that there is a great deal of congruence of the observed peer effects in the lab and the field.

2.1 Peer Effects Experiments

If peer effects exist, then agents working within the visual range of each other should have more similar performances than individuals working alone. In other words, the standard deviation within peers' performances should be less than the standard deviation between individual workers' performances.[4] This is exactly what Falk and Ichino (2006) observe in an experiment. In their setup, two subjects placed in the same room who can observe each other, individually work at the same task, the tasks and payoffs being independent, as in our setting. The data shows that peers' performances are similar and moreover, peers on average exert higher individual efforts than agents working alone. Contrary to conventional wisdom, among peers, "bad apples" do not have a negative effect on the performance of "good apples". It seems rather that the low-productivity workers exert higher efforts in the peer treatment compared to the single treatment. By design, however, there is no possibility to identify which types of workers (low and high productive) worked together as peers. Thus, Falk and Ichino (2006) cannot unambiguously answer the question which team composition (of low and high productive workers) exactly lead to higher average effort levels. In our experiment, we are able to offer an answer to this question. As mentioned

[4] A possible behavioral explanation is conformism (Kandel & Lazear, 1992).

before, our design allows us to identify low and high productive subjects according to their performance in the ability-check phase.

In an empirical study, Mas and Moretti (2009) investigate peer effects in a group production setting. They analyze cashiers' scanner data in a grocery store. This situation involves an individual incentive for each agent to free-ride on peers' efforts. The slower one works, the more others' workload increases. The data shows, however, peer effects dominate free-riding incentives. If there is a highly productive peer present, on average less productive workers' performances increase. The authors quantify the effect as follows: for a 10 percent permanently more productive peer in the given shift, other workers in that shift increase their efforts by 1.7 percent points. Interestingly, monitoring possibilities are crucial for this effect. Only workers who can be seen by the high productive peer increase their efforts. Cashiers who can see the high productive worker, but cannot be seen by him or her, do not change their efforts.[5]

Van Veldhuizen et al. (2014) try to replicate the study by Mas and Moretti (2009) in the lab. With lab settings, one can exclude many channels for social influences that may be present in the field. These unobserved relations may also influence behavior and confound peer effects that are assumed to result from productivity differentials. The main finding of Mas and Moretti (2009) is that a low productivity cashier increases her productivity when she is observed by a highly productive peer. Van Veldhuizen et al. (2014) argue that this effect could be due to a high productive worker being more likely to become a superior in the future, and this explains why a low productive worker strategically increases her performance. In the lab, such strategic considerations can be ruled out. Van Veldhuizen et al. (2014) are not able to replicate the results of Mas and Moretti (2009). They do not find any peer effects whatsoever.

Thöni and Gächter (2015) measure productivity peer effects in a lab experiment. In their setup, two agents independently choose their respective effort level after the principal sets the wage. The agents' productivity and their payoffs do not depend on each other. After exerting effort, both agents learn the output of the other worker and are given the possibility to adjust their effort level up or down. Thöni and Gächter (2015) report that agents follow a low-performing but not a high-performing peer. Highly productive peers adjust down their efforts when they are informed about the other worker's effort.

In a recent study, Gravert and Gerhards (2016) find that subjects show more perseverance in an anagram task (skipping less often difficult anagrams), if they know that an observing subject is informed about the own avoidance behavior, compared when they work alone without being observed.

[5] Bandiera, Barankay, and Rasul (2005) report from a field study with fruit pickers. In their setting, an agent's individual production has negative externalities on co-workers. The authors observe that highly productive agents decrease their individual performances if they can be monitored by their peers.

2.2 Experiments in Immersive Virtual Environments

Blascovich et al. (2002) define an immersive virtual environment as a virtual environment that perceptionally surrounds the subject. In such environments, subjects have a strong feeling of "being there", i.e., being in the synthetically created situation. In this section, we refer to experiments where a subject is "immersed" into virtual reality either entering a virtual reality chamber, a CAVE (Cruz-Neira, Sandin, & DeFanti, 1993); or wearing a virtual reality headset, available such as Oculus' *Rift*, or HTC's *Vive*. The technological setup of our experiment will be explained in Section 4. We do not refer to experiments conducted in *virtual worlds*, which differ from immersive virtual environments, as in virtual worlds subjects usually sit in front of a standard monitor and navigate by using the keyboard or the mouse. For the research potential of virtual worlds, see Bainbridge (2007). Harrison and Rutström (2011) provide an thoughtful methodological remark on the potential use of virtual reality in behavioral research, and the differences between experiments in virtual worlds and experiments in virtual reality.

2.2.1 Social Psychology Experiments in Immersive Virtual Environments

Researchers started to use immersive virtual environments (IVEs) in the 1990s, in domains such visual perception, spatial cognition, psychotherapy, education and learning as nicely reviewed by Blascovich et al. (2002). The use of IVEs in social psychology has been proposed by e.g. Biocca and Levy (1995), and one of the first experiments was conducted by Loomis, Blascovich, and Beall (1999). Blascovich et al. (2002) consider IVEs as a tool that can possibly overcome some of the long-term methodological problems of experimental social psychology research. In particular, IVEs can increase mundane realism without losing experimental control, i.e., in IVEs one can create experimental environments that more closely resemble reality while keeping the advantages of a controlled lab setting. Second, for studies involving human social interaction, IVEs can also help to overcome the difficulties of replication of experimental results in social sciences. In IVEs, one can create virtual humans, who act exactly in the same way in each observation and who are immune to be influenced by the participant. For a comprehensive overview of experimental research in IVEs, we refer to the book by Blascovich and Bailenson (2011) which nicely covers the research in the 2000s and before. For a recent review of the social interaction research in IVEs and a detailed rationale on the unique advantages of such research, see Bombari, Schmid Mast, Canadas, and Bachmann (2015). The most important results concerning the validity of the virtual environment and the acceptance of virtual humans by participants are reported in Section 3.

2.2.2 Economics Experiments in Immersive Virtual Environments

In economics, the idea of conducting experiments in IVEs has been proposed by Fiore, Harrison, Hughes, and Rutström (2009) who also develop the concept of "virtual experiments". They formulate the objective of virtual experiments as to "bridge the gap between the artefactual controls of laboratory experiments and the naturalistic domain of field experiments or direct field studies". This paper nicely presents many arguments in favor of the use of IVEs in (experimental) economic research. The authors propose that subjects can evaluate risks associated with wildfires more accurately when they can watch realistic and physically accurate models of wildfires presented on 3D screens, compared to when only still pictures are provided. Indeed, subjects' beliefs about risks are closer to actual risks when they are exposed to the 3D screen compared to still pictures of the 3D animation.

Our setup is very different from that in Fiore et al. (2009). First, our virtual environment offers a higher degree of immersion since subjects are not looking on small flat 3D monitors, but they enter a virtual reality chamber, and are surrounded by large 3D projection walls. Second, our subjects are not only visually stimulated as in Fiore et al. (2009) but they can move in the virtual reality lab, and they can also interact with virtual objects in a natural way, by using their hands (which facilitates realistic real effort tasks). Third, our subjects are confronted with virtual humans.

The potential benefits of IVE experiments in economic research was further explored in Gürerk et al. (2014). The proof of concept of a real-effort experiment in an IVE was brought forward by Gürerk, Grund, Harbring, Kittsteiner, and Staffeldt (2015). A variant of the design of that study is used to investigate a question in the context of behavioral operations management (DeHoratius, Gürerk, Honhon, & Hyndman, 2015). The present paper shares the basic technological setup with these studies, but it differs from these substantially in that it introduces a co-working virtual human.

3 The Validity of Experiments in Immersive Virtual Environments

Our study is based on the assumption that subjects' actions in virtual environments and interactions with virtual humans are informative about how subjects behave in reality. Whereas in Section 6.1, we investigate whether and to what extent participants in our experiment consider the specific virtual environment as "natural", we do not validate the method as such. Here, we rely on an already significant body of research showing that IVEs generally are perceived as realistic and that subjects consider and react to avatars in a similar way as to humans.

3.1 The Validity of the Virtual Environment

There are two important concepts defining the validity of a virtual environment: *immersion* and *presence*. According to Slater and Wilbur (1997), the difference is as follows: While variety and the quality of the sensory information determine the quality of the immersion, presence refers to the participants' subjective feeling of "being there" in the immersive virtual environment. Immersion relates to objectively measurable, technical aspects of the virtual evironment, for example, the resolution of the displays, whereas presence describes the intensity of the subjective experience. Given the same immersive virtual environment, i.e., for a given degree of immersion, subjects may experience differently strong presence. A recent review by Cummings and Bailenson (2015) covering 83 VR studies reports that some aspects of immersion such as the accuracy of participant-tracking, or the field of view of the visual displays have a greater impact on presence than other aspects, such as the quality of the visual and auditory content.

3.2 The Perception of Virtual Humans

The term *personification* describes the graphical/visual quality of a virtual human, and how natural it appears. The quality of personification influences the acceptance of avatars by their human counterparts and how humans interact with them (Kasap & Magnenat-Thalmann, 2007). To be considered manlike, avatars should not only look like real humans, but also act/move manlike. Therefore, plausible facial expressions, correct eye movements and natural movements of the body are important for the quality of personification. The generally accepted *uncanny valley* theory [6] states that as the quality of personification increases, a human's emotional response to a virtual human gets increasingly positive and emphatic, up to a point beyond which the response quickly turns into strong revulsion and then becomes positive again. Thus, the uncanny valley theory postulates that there is a non-trivial relationship between quality of personification and the acceptance of avatars. For our experiments, facial expressions do not matter as the virtual human is visible only from behind whereas natural body movements is central to achieve a sufficiently high quality of personification.

A number of studies show that participants accept virtual humans as real and behave as if they were human. Bailenson, Blascovich, Beall, and Loomis (2003), for example, measure the interpersonal distance between the human participant and a virtual human in a virtual room, when the participant should walk around the virtual human. As is the case with human-to-human studies, participants move closer to the virtual human when the virtual

[6] The uncanny valley theory originated from the field of robot-human interaction (Mori, MacDorman, & Kageki, 2012) and has been transferred to avatar-human research (Brenton, Gillies, Ballin, & Chatting, 2005)

human is not having any eye contact with them – as compared to a situation where the virtual human is looking at them. In general, when participants encounter virtual humans, one can observe an increase in arousal, which is the physiological and psychological state of being awake, and an indicator of human emotions. Slater et al. (2006), for example, measure the increase in arousal through physiological responses such as skin conductance and heart rate variability. For a comprehensive review of the evidence on humans' perception of virtual humans in social interactions, see Bombari et al. (2015).

4 Formulation of Hypotheses and Experimental Design

We investigate whether and how a worker's (referred to as she) performance is influenced by the physical presence of another worker (referred to as co-worker or he) working on an identical but otherwise unrelated task. In particular, we ask whether the observable productivity of a co-worker affects an agent's own work performance in a simple economic setup without any interdependencies between co-worker and agent. In particular, co-worker and agent do not work on the same task and an agent's monetary payoff is independent of her own and the co-worker's performance; the agent receives a fixed amount (unrelated to her performance). Furthermore, whereas the agent can observe the co-worker performing his task, the co-worker does not observe the agent (who faces the co-worker's back). The only difference between our two experimental treatments is the productivity of the co-worker, which is exogenous.

We are interested in whether subjects adjust their work speed to that of the avatar. To be able to observe an adjustment, we have to know how subjects' "normal" performance is and control for interpersonal differences in subjects' abilities. Thus, our experimental design consists of two phases. In phase 1, the *ability-check phase*, subjects work on their own, knowing that there will be a second phase, the *treatment phase*, but subjects do not know that they will work in presence of an avatar in the second phase.

4.1 Standard Economic Prediction

Standard economic theory (assuming a payoff maximizing worker) predicts no peer effects as payments and work outcome of agent and co-worker are independent. In our setting, there is no possibility to free-ride on others' efforts and there are also no incentives to do so. Because of the fixed wage, there is no incentive to exert any effort anyway. Ignoring dynamic aspects such as the possibility of learning and the possibility that subjects can get bored and fatigued in the course of the experiment, standard theory leads to the following hypothesis:

Hypothesis 1 *There are no differences in agents' average performance whether the agent is working in the presence of a low productive co-worker (in the SLOW treatment) or in the presence of a highly productive co-worker (in the FAST treatment).*

If we observe peer effects, these effects cannot be driven by (selfish) financial motivations. In the following, we elaborate on two alternative possible sources of motivation.

4.2 Social Learning through Imitation (Implicit Learning)

Social learning can occur through imitation (Bandura & McClelland, 1977). In our setting, the agent could (implicitly) learn how to sort cubes by observing and imitating the co-worker. In our experimental design, however, co-workers in both treatments do not display different sorting techniques. The only difference between the SLOW and the FAST treatments is the that the co-worker in the SLOW treatment is taking short times of rest whereas the co-worker in the FAST treatment works continuously. We suspect that subjects in our experiment cannot learn a more efficient sorting technique in the FAST treatment than in the SLOW treatment. As subjects in the FAST treatment can observe more (but not better) sorting incidents, we cannot exclude the possibility that there is a learning effect. It should be noted, however, that learning is hampered by the fact that subjects were not able to fully observe the co-worker's actions as these were partly covered by the co-worker's body. Even though we suspect social learning not to occur in our setup, there might be other/additional reasons for imitation which would imply that subjects in the SLOW treatment would possibly inspect less cubes because the avatar does so. In any case, if social learning mattered, it would give rise to the following hypothesis:

Hypothesis 2 *The agents' average productivity in the SLOW treatment is lower than in the FAST treatment.*

4.3 Competition through Social Comparison

Social comparison theory (Festinger, 1954) states that humans are inclined to evaluate their own opinions and abilities (cf. Hypothesis 1 in (Festinger, 1954)). In the absence of objective (non-social) feedback sources as in our setting, people compare themselves to similar others (peers) for evaluation purposes. Festinger (1954) hypothesizes that the tendency to compare oneself with someone else decreases, the more different (from oneself) the other person's perceived opinions and abilities are (cf. Hypothesis 3 in (Festinger, 1954)). For our experimental design this implies that a low ability person is more likely to compare herself to a low ability peer, and a high ability person to a high ability peer.

Festinger (1954) also states that, in order to be able to make comparisons and evaluate themselves accurately, people try to reduce the discrepancies between themselves and others. This means, that someone who is a bit slower than the person whom she is comparing herself to, tries to catch up with this person, and a competitive behavior emerges. Hence, a low-ability agent works harder in the presence of a low-productive co-worker (as compared to the presence of a high-productive co-worker); as for a low-ability agent, the perceived chances to beat a low-productive peer are higher than to beat a high-productive co-worker. If, however, the discrepancy between the low-ability agent and the high-ability co-worker is too large, for the low-ability agent, the "upward-drive" does not work. The reason is, for the low-ability agent, while the chances to beat the co-worker is small, the costs of effort to catch up with the highly productive co-worker are too high. In case of a low-ability agent, this means, that she does not try to catch-up with the other person who is far better than she, simply it is too costly to do so.

The arguments depicted above imply that a low-productive worker will increase her productivity more in the SLOW treatment than in the FAST treatment. A high-ability subject should also be more productive in presence of a similar productive avatar, though for a high-ability agent, the potential to improve productivity is probably smaller.

Hypothesis 3 *Low-ability agents are more productive in the SLOW treatment compared to the FAST treatment. High-ability agents are more productive in the FAST treatment than in the SLOW treatment.*

5 The Lab and the Procedures

The experimental setup including the avatar is an extension of the setup we already pre-tested in the virtual reality lab *aixCAVE* of the RWTH Aachen University (Gürerk et al., 2015). In our experimental setting, when a human subject enters the aixCAVE, she perceives the virtual environment as a production hall. Being inside this immersive virtual environment, the human subject performs a real-effort sorting task at a virtual conveyor belt while seeing a virtual human that independently performs the same task. Below, we explain the features of the virtual reality lab, the task, and the avatar.

5.1 The Virtual Reality Lab: aixCAVE

The aixCAVE of the RWTH Aachen University (Kuhlen, 2014) is a five-sided immersive virtual environment with a size of $5.25m \times 5.25m \times 3.30m$ (width $\times$ length $\times$ height). The five projection displays, i.e., the four walls and the floor, enclose the user giving her a 360 degree field of view in the horizontal plane in the computer-generated environment.[7]

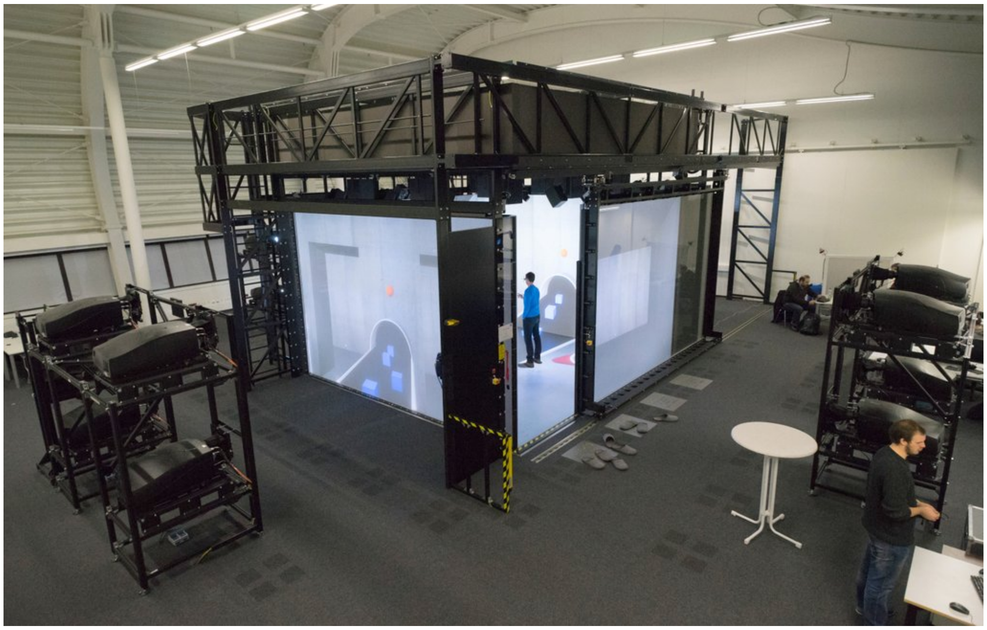


Figure 1: The virtual reality lab aixCAVE of the RWTH Aachen University

By employing a stereoscopic projection approach in conjunction with special 3D glasses, the virtual scenario is presented in a three-dimensional, quasi-holographic way as known from 3D cinemas. Due to the high resolution[8] a very detailed and crisp image of the virtual environment can be presented to the user, thereby immersing her into it. Images are generated by means of a user-centered projection, i.e., depending on the user's position and viewing direction. To this end, tracking markers are attached to the glasses, thereby giving the system knowledge about the user's head pose[9]. As a direct consequence, a user can intuitively see the scene from any desired perspective. To facilitate a natural way of interaction with the virtual environment, and an intuitive sorting of virtual cubes on a conveyor belt, the subject's hand has to be tracked. For this, tracking markers are attached to the user's dominant hand by means of a hand band to which the markers were attached. Using the pose information of the user's hand, the human subject is able to pick up the virtual cubes and inspect them, and throw them into a bin, just by moving her own hand.

[7]CAVE is indeed an acronym for Cave Automatic Virtual Environment (see Cruz-Neira et al. (1993) for the basic working principles of a CAVE). One can imagine a CAVE also like a "holodeck" from the sci-fi series "Star Trek, The Next Generation" which the spaceship crew uses mainly for recreation by visualizing experiential settings.

[8]In total 24 projectors are used with a resolution of 1920 $\times$ 1200 pixels each, four per wall and eight for the floor.

[9]This setup causes to the aixCAVE to be a single user system.

5.2 The Task

We use a simple real effort task. A constant stream of objects (abstract cubes) approach the subject from the right.[10] There are two types of cubes. The first type of cubes has six blue-shaded sides, they represent objects without a defect. The other type of cubes has five blue-shaded and one red-shaded side, they represent objects with a defect. The red side is not visible to the subject when the cube appears on the belt. The defect side is located either on the bottom, the back or the right-hand side of the cube. The task is to sort out the objects with a defect (having a red side). To inspect a cube, the subject must grasp a cube, rotate it, and, if defect, put it into a bin which is located to the right or, if not defect, put it back onto the belt. In a series of pilot sessions, we adjusted the speed of the belt such that it is challenging to inspect each and every cube in the given time but it should be doable for most subjects.

5.3 The Main Interest Variable

We measure a subject's *productivity* as the sum of the number of sorted cubes with a defect put into the bin (correctly sorted), minus the number of cubes without a defect into the bin (falsely sorted). We denote the productivity in the ability-check phase with $ACProd$, and the productivity in the treatment phase with $Prod$. Note that sometimes we refer to a subject's *ability*, when we talk about the subject's productivity in the ability-check phase.

The main performance variable we are interested in is the relative increase in productivity between the ability-check and the treatment phase. We denote this variable $ProdInc$ and define it as the difference between the productivity in the treatment phase and (two times[11]) the productivity in the ability-check phase divided by (two times) the productivity in the ability-check phase (ability). Hence, the relative increase in productivity for subject i is equal to: $ProdInc_i = \frac{Prod_i - 2 * ACProd_i}{2 * ACProd_i}$

There are no penalties whatsoever for the defect cubes that were left unsorted on the belt. Note, we do not provide any financial incentives to work. As we designed our experiment, we deliberately decided against piece rate to observe non-confounded "pure" peer effects. During the instructions, we simply told subjects that they should "reach a work output as good as possible".

In the ability-check phase, 34 out of 168 cubes are defect, and in the treatment phase 68 out of 336 cubes. The order of cubes was randomly drawn before the start of the first session. All subjects are confronted with the exact same cube order. Between the first and second phase, there is a short break, where the experimenter enters the VR lab and

[10] We asked subjects for their handedness before starting the experiment and reversed the whole virtual environment for the left-handed. In particular, cubes approached left-handed subjects from the left.

[11] Note that the ability-check phase only had half the number of cubes to be sorted.

informs the subject about the second phase handing out a short note (see Appendix). In particular, the subject learns that in the second phase there will be a "co-worker" who performs the same task as the subject but does so independently. The subject also learns that her co-worker is computer-controlled and represented by a virtual human.

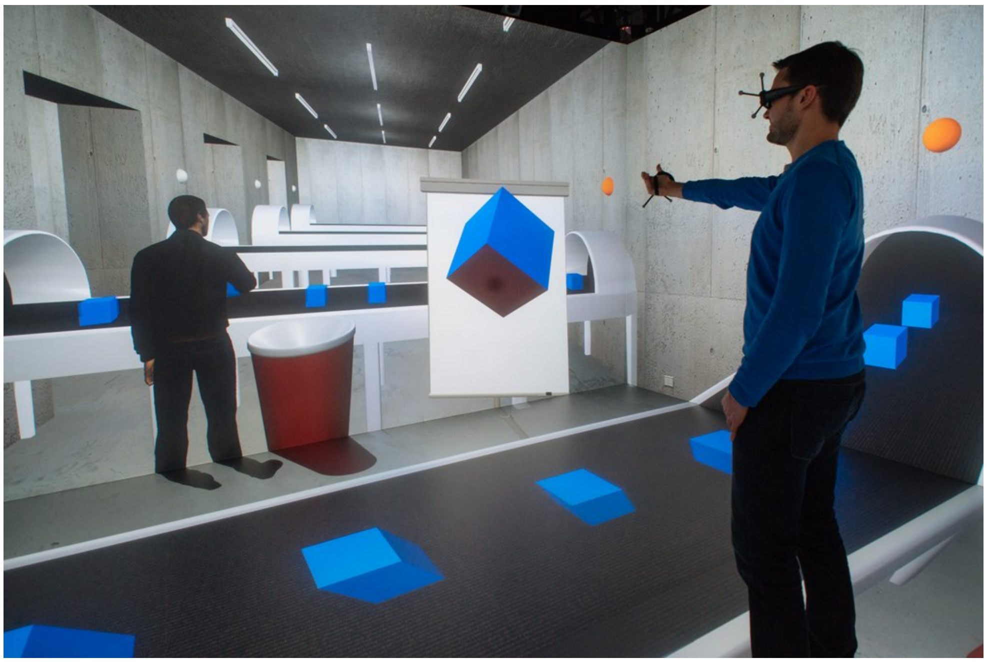


Figure 2: The experimental setup in the virtual reality lab

5.4 The Virtual Human as a Peer

In order to make the virtual human move as naturally as possible, we captured a student assistant's movements via a Kinect 360 (Windows, 2011) while the student performed the task in the aixCAVE. Afterwards, we transferred the data using 3ds max (Autodesk, 2014) into the model of a male virtual human, provided by SmartBody (USC Institute for Creative Technologies, 2014). Using this procedure, we modeled two otherwise identical virtual humans with different behaviors: one worker sorting virtually all cubes, and one worker only sorting a subset of cubes while having short rests during the task.[12]

[12] See the video clips on https://www.youtube.com/watch?v=RuX688VliB4 for the avatar we modeled for the SLOW treatment, and https://www.youtube.com/watch?v=xZgBdlD1ypQ, for the FAST treatment, respectively.

5.5 The Experimental Procedures

Subjects were recruited with ORSEE (Greiner, 2004). The participants were paid a fix amount of 12 Euros. The whole experimental procedure including instructions and the questionnaire took approximately 45 minutes. The experimental software was developed by the Virtual Reality and Immersive Visualization Group of RWTH Aachen University. Programming of the questionnaire was done using z-Tree (Fischbacher, 2007). The data was collected in 108 sessions, conducted one ofter another, since only one participant can experience the immersive environment with correct visual perspective in the VR lab at the same time. We have 54 observations in the SLOW and 54 observations in the FAST treatment.[13] The sessions were conducted on eight days. To avoid possible biased behavior emerging from different days and different times of the day, we constantly alternated the sessions between the SLOW and FAST treatment.

6 Results

We derive and discuss two different types of results. First, in Section 6.1, we are concerned with subjects' perceptions of the virtual environment, and in particular, with their perceptions of the virtual co-worker. These results are based on the survey that subjects completed after the experiment. They support the internal validity of the method used. Second, in Section 6.2, we are concerned with peer effects. There we test the hypotheses derived in Section 4 based on the outcome of the experiment.

6.1 Subjects' Perception of the Virtual Environment and the Avatar

To evaluate the subjects' perception of the virtual environment including the virtual human and the task, we use a variant of the presence questionnaire developed in (Slater, Usoh, & Steed, 1994). Table 1 summarizes the results concerning subjects' general experience in the virtual environment in the aixCAVE for the two different treatments. Subjects evaluated different aspects of the virtual environment on a 7-point Likert scale with high numbers indicating higher realism. In general, subjects evaluate their experience in the aixCAVE as being realistic, and for the four of six categories the ratings in both treatments are very similar. Only in two categories, subjects in SLOW consider their "sense of being" and the realism of the virtual environment (weakly) significantly higher than those in FAST ($p = 0.078$ and $p = 0.012$, Mann-Whitney rank-sum test, two-sided)[14], see the first two

[13] We had to cancel another 11 observations due to software crashes.

[14] Unless otherwise stated, we use the non-parametric Mann-Whitney rank-sum test for between treatment comparisons, and report the p-values of two-sided tests.

Table 1: Subjects' Experience in the aixCAVE

Treatment	Sense of Being	Virt. Env. Was Reality	Saw vs Visited	Where Were You	Memory Structure	Physically in Virt. Env.	SUS score
SLOW	5.57	4.81	4.04	4.91	4.39	4.44	2.22
	(0.92)	(1.26)	(1.94)	(1.55)	(1.55)	(1.64)	(1.40)
FAST	5.13	4.07	4.31	4.87	4.50	4.30	2.07
	(1.32)	(1.58)	(1.85)	(1.61)	(1.75)	(1.89)	(1.32)

Note: Standard deviations in parentheses. All questions were measured on a 7-point Likert scale with higher numbers indicating a greater sense of realism in the virtual environment.

questions of Table 1.[15]. The scores we observe are similar to those obtained in other CAVE studies (Usoh, Catena, Arman, & Slater, 2000).

Another measure of subjects' perception of IVEs is the (binary) Slater-Usoh-Steed (SUS) score (Slater et al., 1994). The SUS score sums for a subject, the number of high evaluations (6 or 7) this subject gave for the six questions. We do not find a significant difference in the SUS score between our treatments (2.22 and 2.07).[16]

For the experimenter, another way to get insights about subjects' perception of the virtual environment is to observe subjects' behavioral responses. How do subjects move in the virtual environment, and how do they react to the avatar? Anecdotal observations of participants' movements in the aixCAVE suggest a realistic interaction with the virtual environment. Indeed, it happened that one participant of our experiment wanted to lean on the virtual conveyor belt and she fell on the ground.[17]

6.1.1 Subjects' Perception of the Avatar

In the post-experimental questionnaire, subjects were also asked about their perception of the avatar (see Table 2). In particular, subjects were asked to rate the virtual human's perceived performance on a Likert-scale ranging from 1 to 7, where 7 stands for a very good performance. Note that participants were not able to exactly observe the avatar's productivity in terms of correctly or false sorted cubes, simply because they were standing behind the virtual human with no clear vision on all sides of a cube in the avatar's hand. Subjects in the FAST treatment rated the avatar's performance as being considerably higer than subjects in the SLOW treatment (SLOW: 4.09 and FAST: 5.15, $p < 0.001$).

[15] The numbers in Table 1 are similar to those in a related study conducted in the aixCAVE that did not involve avatars (DeHoratius et al., 2015). For example, "sense of being" was rated on average 5.53 and realism, the second question, was rated 4.26. The remaining questions were rated with 4.28, 4.94, 4.01, and 4.42, respectively

[16] The SUS score we obtain is also similar to the SUS score of 2.00 reported in (DeHoratius et al., 2015).

[17] Luckily, she did not hurt herself and could go on with the experiment.

Table 2: Subjects' Perceptions of the Avatar and the Task

Treatment	Avatar is productive	Avatar moves natural	Enjoyed CAVE	Enjoyed Task	Exerted Effort
SLOW	4.09	4.78	6.44	5.56	6.04
	(1.61)	(1.45)	(0.84)	1.34)	(0.95)
FAST	5.15	4.76	6.39	5.41	5.78
	(1.50)	(1.33)	(0.88)	(1.45)	(1.24)

Note: Standard deviations in parentheses. All questions were measured on a 7-point Likert scale.

Result 1 *Subjects in the FAST treatment evaluate the avatar's work performance significantly higher than subjects in the SLOW treatment.*

This result shows that our intended treatment variation, that the virtual human in the FAST treatment is more productive compared to the avatar in the SLOW treatment, was experienced by subjects in the intended direction. To see whether the observed difference between the ratings can be attributed to possible differences in the perception of the avatar's movements, we also asked subjects about how they considered the realism of the computer-programed virtual human's movements. In both treatments, participants judge the avatar's movements similarly (SLOW: 4.78 and FAST: 4.76). This numbers underline, that the perceived differences in the avatars' behavior cannot be attributed to the perceived differences in avatars' movements.

6.1.2 Do Low and Highly Able Subjects Perceive the Avatar Differently?

To further investigate possible treatment differences in the avatar's perceived productivity, we split all subjects in two equally sized groups based on their respective performances in the ability-check phase (recall that subjects do not see an avatar in or before this phase of the experiment). Subjects who had a productivity of more than 20 units we classify as the high-ability subjects (Hi-Ab), and those with a productivity of 20 units or less, we denote the low-ability subjects (Lo-Ab).

Averaged over both treatments, Hi-Ab subjects rate the avatar's productivity significantly lower than Lo-Ab subjects, (4.19 and 5.05, $p < 0.01$). This suggests that subjects rate the avatar's behavior relative to their own – subjectively judged – performance. Hi-Ab subjects perceive the avatar as less productive than Lo-Ab subjects. Both Lo-Ab and Hi-Ab subjects rate the avatar's performance in the SLOW treatment as significantly less productive as the avatar in the FAST treatment. Lo-Ab subjects rate the avatar's performance in SLOW with 4.6, and in FAST with 5.4 ($p = 0.034$) while Hi-Ab subjects rate the avatar's performance in SLOW with 3.8 and in FAST with 4.8, respectively ($p = 0.057$). These facts nicely add to the observation that subjects are not indifferent with respect to the virtual human, and that they perhaps perceive it as a peer.

The questionnaire also reveals (see Table 2) that subjects highly enjoyed the experience of being in the aixCAVE (SLOW: 6.44 and FAST: 6.39, no significant differences between treatments). The task was also enjoyed, but less than the experience of the aixCAVE (SLOW: 5.56 and FAST: 5.41, again no differences between treatments). This indicates that performance differences between treatments cannot be attributed to differences in how exciting the CAVE experience was or to the degree of task enjoyment. The last item in Table 2 nicely supports this congruence of perception, as it shows that subjects' self-judgment of their exerted effort level is rather high and not significantly different between treatments.

6.2 Peer Effects

We first test whether subjects' performances are affected by the working behavior of the avatar. To accommodate for differences in ability in performing the sorting task, we compare relative increases in productivity between the ability-check phase and the treatment phase. Recall that the ability-check phase was identical for all subjects, in particular there was no virtual human. Throughout this subsection, we measure performance as the relative increase in productivity as defined in Section 5.3.

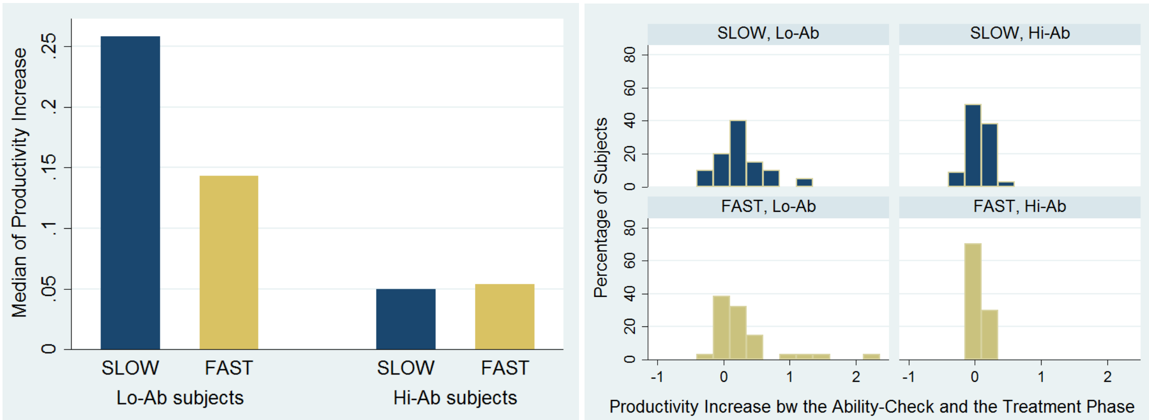
Result 2 *Over all subjects, there is no significant difference in average performance between the SLOW and the FAST treatment.*

Over all subjects, we do not observe a statistically significant difference between the relative increases in productivity between the two treatments (SLOW: 0.12, and FAST: 0.09, $p = 0.851$). This result is in line with Hypothesis 1.

If we restrict attention to high-ability (Hi-Ab) subjects as they were defined in Section 6.1.2, we find no significant difference in performance between the two treatments (SLOW: 0.04, and FAST: 0.05, $p > 0.1$), which supports Hypothesis 1. An alternative explanation for this observation could be that a Hi-Ab subject has less potential to increase her productivity in the treatment phase if she performed already close to her limit in the ability-check phase.[18] If we now look at the Lo-Ab subjects, we find that in the SLOW treatment, they increase their productivity considerably stronger than in the FAST treatment (median SLOW: 0.26, and FAST: 0.14, $p = 0.502$), though the increase is not statistically significant.

[18] Note, however, no subject achieved the best possible work productivity (sorting out all defect cubes) in the ability-check phase.

Figure 3: Relative Increase in Productivity between the Ability-Check and the Treatment phase



(a) Medians over treatments and types

(b) Histograms over treatments and types

6.2.1 Do More Competitive Subjects Show Higher Productivity Increases?

As discussed in subsection 6.1, subjects acknowledged the presence of a virtual co-worker and can judge its performance. To understand whether and how subjects react to the avatar, in the post-experimental questionnaire, we asked subjects to state whether they cared about being more successful than the avatar. More than a quarter of subjects (27.8 percent) were affirmative to this question. We refer to these participants as the competitive subjects.

Interestingly, significantly more subjects in the SLOW treatment can be classified as competitive as compared to the FAST treatment (SLOW: 19 of 54 subjects, and FAST: 11 of 54, one-sided probability test, $p = 0.043$). Conditional on being competitive, those in the FAST treatment had significantly higher productivity in the ability-check phase than subjects in the SLOW treatment (SLOW: 17.7 and FAST: 23.3, $p = 0.006$). This suggests that in both treatments, (competitive) subjects judge the avatar's productivity to different extents: compared to the FAST treatment, in SLOW, rather subjects with lower productivity (ability) state that they can beat the avatar.

Averaged over both treatments, competitive subjects have a much higher (relative) productivity increase, as compared to non-competitive subjects. The difference, however, is not statistically significant (median of the competitive: 14.9 percent and non-competitive: 8.3 percent, $p = 0.111$). Interestingly, if we only look at subjects in the FAST treatment, competitive subjects have a significantly higher productivity increase than the non-competitive subjects (medians: 20.5 percent and 7.4 percent, $p = 0.029$). In the SLOW treatment, however, we do not observe such a difference (medians: 10 and 12 percent).

Result 3 *(a) The percentage of competitive subjects is significantly higher in the SLOW treatment. (b) The competitive subjects in the FAST treatment, however have higher abilities, and (c) compared to non-competitive subjects – they show higher productivity increases than the competitive subjects in the SLOW treatment.*

6.3 Regression Analyses

To better understand the relationship between a subject's performance and that of an avatar, we conducted a series of regressions. The dependent variable in all our models is productivity $(\text{Prod})_i$ in the treatment phase (as defined in 5.3). As independent variables, we include productivity in the ability-check phase $(\text{ACProd})_i$, an indicator variable $(\text{SLOW})_i$ for the treatment ($(\text{SLOW})_i = 1$ if subject i is in the treatment with the low productive avatar), and an interaction term between SLOW and ACProd (see Equation 1). In some models, we add further control variables, e.g., for gender, age, whether subjects were left-handed,[19] whether subjects perceived the avatar's movements as natural, and other variables reported in Table 1 and Table 2, and whether subjects were studying for an engineering degree. None of these controls turned out to be significant. In other variants of the models, we also included as independent variables the items from the immersion questionnaire, and the items on the subjects' perception of the avatar. None of these variables were significant either.

$$(\text{Prod})_i = \alpha + \beta \cdot (\text{ACProd}_i) + \gamma \cdot (\text{SLOW}_i) + \delta \cdot (\text{ACProd}_i \cdot \text{SLOW}_i) + \epsilon_i, \qquad (1)$$

As depicted in Table 3, the coefficients of the respective variables of model (I) reveal, first, that the productivity in the treatment phase depends positively and significantly on the productivity exerted in the ability-check phase. Second, there is no significant treatment difference regarding the level of productivity, as both the coefficient for the treatment dummy SLOW as well as the interaction term SLOW $\times$ ACProd are not significant.

Model (II) takes into account a number of control variables of which none is significant. In regressions we do not include here (to save space), we also ran models including the control variables from Table 1 and Table 2. All controls remain insignificant. We exclude these variables in the models (III and IV). Note, in fact, we did run additional regressions to be sure that these variables remain insignificant when we add them to models (III) and (IV).

[19] Recall that the VR-environment was mirrored for left-handed subjects and thus should not discriminate against these subjects

Table 3: Determinants of the productivity

	(I, all)	(II, all)	(III, Hi-Ab)	(IV, Lo-Ab)
ACProd	1.012	1.039	1.537	-0.219
	(0.231)***	(0.236)***	(0.563)***	(0.580)
SLOW	-2.614	-0.714	14.846	-10.417
	(7.342)	(7.623)	(17.500)	(18.201)
SLOW $\times$ ACProd	0.151	0.068	-0.589	0.633
	(0.342)	(0.353)	(0.695)	(1.064)
AGE		0.063		
		(0.148)		
MALE		-2.469		
		(1.847)		
ENGINEER		0.984		
		(1.776)		
LEFT-HANDED		5.986		
		(4.331)		
_cons	26.000	24.668	14.364	45.633
	(4.730)***	(6.360)***	(14.253)	(9.822)***
Adjusted R-squared	0.29	0.32	0.21	0.01
N	108	108	54	54

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

6.3.1 Do Data Support the Social Comparison Hypothesis?

The social comparison hypothesis predicts higher performance when subjects are working in the presence of similarly productive co-workers. In particular, the Lo-Ab subjects should demonstrate a higher productivity in the SLOW treatment, when they are working in the presence of a comparable virtual co-worker, than in the FAST treatment. Similarly, the Hi-Ab subjects should show a higher productivity in the FAST treatment. To check these predictions, model (III) includes only the Hi-Ab subjects, and model (IV) only the Lo-Ab subjects.

Let us first look at model (IV) with only Lo-Ab subjects: As the coefficient of the AC-Prod shows, in the FAST treatment, the productivity of the Lo-Ab subjects is negatively correlated with ability. The interaction term SLOW $\times$ AC-Prod indicates, however, for Lo-Ab subjects, there is a positive correlation between ability and productivity in the SLOW treatment. The coefficient, is, however not significant. Hence, these numbers provide directional (but not significant) support for the social comparison hypothesis.

Now, let us look at the model (III) with only Hi-Ab subjects. In the FAST treatment, we observe a significant positive correlation between ability and productivity (see the coefficient AC-Prod) while this correlation is less pronounced for Hi-Ab in the SLOW treatment as the interaction term SLOW $\times$ AC-Prod shows. Hence, for the Lo-Ab subjects, the social comparison hypothesis is supported.

Result 4 *Data partly support the social comparison hypothesis. For Hi-Ab subjects, the data supports the social comparison hypothesis, for Lo-Ab the direction of the effect is OK but the effect is not significant.*

6.4 An Alternative Performance Measure

A possible explanation for the insignificance of the treatment variable in the previous regressions is that productivity is very noisy because it might be affected by subjects' luck to select the "easy" cubes. To better understand this, consider the following hypothetical situation: a subject only inspects every other cube, checks it from below and only sorts it out if she can see a defect on one of the inspected sides. Then, depending on whether the subject started its inspection with the first or with the second cube, she might get very different productivity values, depending on how defects are distributed over cubes, and the sides of a cube.

As in an immersive virtual environment we can track subjects' movements, we are able to detect how carefully a cube was inspected. More precisely, as we track the position and orientation of a subject's hand and head, we are able to detect which sides of a cube a subject actually inspected. In particular, we can infer from the tracking data, whether a cube that was put into the bin was inspected thoroughly by looking at each side of the cube, or whether it was simply taken from the belt and thrown into the bin. In the following, we use this information to construct a novel performance measure.

6.4.1 Tracking Data Allows to Determine the Quality of Inspections

To quantify the quality of inspections, we count the number of carefully inspected cubes. A cube is defined as inspected "carefully", if

(i) it is a non-defect cube (i.e., a cube without a red side), and was inspected on all three hidden sides and put back on the belt, or

(ii) if it is a defect cube (i.e., a cube with a red side), and the subject correctly placed the cube into the bin.

In particular, a cube that is grabbed, but put on the belt without looking at all of its hidden sides, is classified as not inspected carefully.

The number of carefully inspected cubes is significantly ($p < 0.001$) and positively correlated with productivity ($R^2 = 0.748$).[20]

We use the increase in the number of carefully inspected cubes to identify subjects, whose productivity was not (primarily) driven by the result of careful inspections. The idea is as following: for each subject, we measure the relative increase of her careful inspections between the ability-check and the treatment phase, and compare it to the relative increase of her productivity. Note, to account for differences in abilities, we consider the relative increases. If the performance of a subject in terms of the relative increase of her careful inspections and her relative increase in productivity are similar, then we interpret this work as high quality performance. If, on the other hand, a subject exhibits a larger difference in these two measures, we interpret this as evidence of low-quality work, in as much as the sorting of cubes was apparently less the result of a high quality inspection but stems from (good or bad) chance. We define a threshold to seperate subjects with "quality-driven" performances from the subjects whose performances were "chance-driven". If the difference between both variables is not too large, i.e., below the threshold, then we label these subjects as quality-driven, otherwise subjects' performances are identified as chance-driven.

In what follows, we just consider the 54 subjects with quality-driven performances,

[20]We do not find any significant treatment differences for the number of careful inspections, neither in the ability-check phase (median SLOW: 101, median FAST: 94, $p = 0.174$), nor in the treatment phase (median SLOW: 239, median FAST: 234.5, $p = 0.332$). There is also no significant difference between treatments for the increase in the number of careful inspections between the ability check phase and the treatment phase, (median SLOW: 0.092, median FAST: 0.127, $p = 0.559$). There is also no difference if we look at the performance increase between both phases for both subject types separately. For the Lo-Ab subjects, we obtain: median SLOW: 0.153, median FAST: 0.186, $p = 0.591$, and for the Hi-Ab subjects, we observe: median SLOW: 0.087, median FAST: 0.086, $p = 0.986$.

Table 4: Determinants of the productivity for quality-driven performers

	(I, all)	(II, all)	(III, Hi-Ab)	(IV, Lo-Ab)
AC-Prod	2.159	2.129	2.478	3.059
	(0.323)***	(0.334)***	(1.240)*	(0.734)***
SLOW	39.961	39.972	49.553	65.506
	(9.640)***	(10.101)***	(35.908)	(17.689)***
SLOW × AC-Prod	-1.619	-1.625	-1.925	-3.039
	(0.414)***	(0.434)***	(1.373)	(0.907)***
AGE		-0.034		
		(0.170)		
MALE		-0.161		
		(1.978)		
ENGINEER		0.948		
		(1.854)		
LEFT-HANDED		5.504		
		(3.809)		
_cons	-0.012	0.843	-9.383	-16.526
	(7.363)	(9.143)	(32.407)	(14.285)
Adjusted R-squared	0.52	0.55	0.16	0.55
N	54	54	30	24

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

and exclude all subjects that were chance-driven from our analysis.[21] For this sub-sample consisting of 54 subjects, we re-estimate models (I-IV) from Section 6.2. The main findings are summarized in Table 4. The treatment dummy SLOW, ability (ACProd) and the interaction term of these two variables are all highly significant in models (I), (II) and (IV). As the interaction term is negative, the difference in productivity between FAST and SLOW is increasing in ability.

An ex-post estimation of the marginal effect (dy/dx) of the treatment dummy SLOW on productivity for different levels of ability is given in Table 5. From this, we can infer that subjects with low ability ($ACProd \leq 22$) in SLOW are more productive than in FAST, and this is significant (at the 1 percent level for $ACProd < 22$ and at the 5 percent level for $ACProd = 22$). For subjects with higher ability, the treatment effect is either not significant ($22 < ACProd < 27$) or it is negative, i.e., subjects in FAST are more productive than in SLOW.[22]

If we further split the sample in Hi-Ab and Lo-Ab subjects according to their ability-

[21] Whenever the difference between increase in productivity and the increase of carefully inspected cubes is larger than 12 percentage points, a subject is considered chance-driven and excluded.

[22] The significance levels are 10 percent for $ACProd = 28$, 5 percent for $ACProd = 29 and ACProd = 30$ and 1 percent for $ACProd = 31 and ACProd = 32$

Table 5: The Marginal Effects in Model (I) in Table 4

ACProd	13	14	15	16	17	18	19	20	21
dy/dx	18.91	17.29	15.68	14.06	12.44	10.82	9.20	7.58	5.96
Std. Err.	4.45***	4.07***	3.70***	3.34***	2.99***	2.67***	2.37***	2.11***	1.90***

ACProd	22	23	24	25	26	27	28	29	30	31	32
dy/dx	4.34	2.72	1.11	-0.51	-2.13	-3.75	-5.37	-6.99	-8.61	-10.23	-11.85
Std. Err.	1.77**	1.73	1.79	1.94	2.16	2.43	2.74*	3.07**	3.42**	3.78***	4.15***

Note: dy/dx for factor levels is the discrete change from the base level.
Standard Errors bootstrapped with Delta-method.
* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

check productivity (a median split of quality-driven subjects results in 24 Lo-Ab and 30 Hi-Ab subjects) and run the regressions on the Lo-Ab subset, we can confirm that the lower the ability, the higher the performance in the presence of a slow avatar as compared to the fast avatar (see Table 4). This effect is not significant for Hi-Ab subjects, for which we only find that ability is positively correlated with productivity (see model III in Table 4). These findings are consistent with the social comparison hypothesis (Hypothesis 3) but not with Hypotheses 1 and 2:

Result 5 *The data stemming from quality-driven observations support the social comparison hypothesis. Low ability subjects show a higher performance in SLOW than in FAST whereas subjects with a higher ability show a higher performance in FAST than in SLOW whereas for subjects with intermediate levels of ability we do not obtain significant result. In general, the difference in performance between FAST and SLOW is decreasing in ability.*

7 Conclusions

We present a novel methodology to investigate peer effects in the workplace. By incorporating a virtual human in an immersive virtual environment as the peer of human subjects, we avoid the reflection problem: while the avatar is immune to be influenced by the human subject, it may influence the human subject. We find that human subjects rate the virtual environment and the virtual peer as quite natural. We observe stronger peer effects when the perceived similarity in abilities between the human and virtual peers is high. This result is in line with the predictions of the social comparison theory.

The applied methodology allows for interesting new investigations based on our study. Firstly, one can easily extend the focus of the study and, for example, ask to what extent the findings would be robust when we have a female avatar. One important advantage of the virtual reality experiments is that we could do same experiment with the exact setup including the same model of a virtual peer that would do everything the same way except looking like a female. The virtual human model we used can also easily be adapted to display some other physical attributes, such as height, weight, or ethnicity, with all other things being equal. Hence, immersive virtual environments are ideal to investigate how a variety of different characteristics of a person influences her peers.

Secondly, further analysis of subjects' movements using tracking data could be used to address additional research questions. For example, tracking data can be used to measure interpersonal distances, in order to evaluate (unconscious) behavioral patterns. A study by Dotsch and Wigboldus (2008) shows that people keep greater distances to avatars with different ethnic looks. Another study by McCall and Singer (2015) uses interpersonal distance measures to predict financial behavior.[23]

We believe that experiments in immersive virtual environments is a very promising research methodology that could nicely complement classic lab and field experiments. Using the VR technology, one can create lab experiments with more natural contextual cues, making lab experiments a bit more like field experiment; however, without losing any experimental control. Tracking data can used to be for novel analyses, that may help discover new insights. Finally, with experiments in virtual reality, there are virtually no limits for creating counter-factual settings. One can conduct experiments that would be expensive, dangerous or simply impossible otherwise.

Acknowledgments

We thank Jonathan Wendt for his assistance in conducting the experiment, and Lucas Braun for preparing the raw data for the analysis. We also thank Almut Balleer, Robert Böhm, Bernd Irlenbusch, the participants of the seminars in Bamberg, Erfurt, Göttingen, Karlsruhe; the conference participants at the 2015 Economic Science Association Meetings in Sydney, Heidelberg, and Dallas; and the participants of the 15th TIBER Symposium in Tilburg, for their valuable comments. We acknowledge the generous financial support from the "Excellence Initiative (ZUK II)" of the German Science Foundation (DFG).

[23] In the experiment, first, subjects played several rounds of a trust game with two confederate players, one being "fair" and the other "unfair". After the trust game, entering a virtual room, human subjects could approach the avatars of the fair and the unfair player. After the virtual encounter, subjects could punish the other players, by investing a token for a deduction of three tokens on the punished player's account. McCall and Singer (2015) find that (i) when being in the virtual room, subjects kept the fair player closer, and (ii) subjects who invested in financial punishment of the unfair player were more likely to stand directly in front of those players.

References

Autodesk. (2014). *3ds Max.* "http://www.autodesk.de/products/3ds-max/overview". ([Online; accessed 30-May-2015])

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

Bainbridge, W. S. (2007, 27 July). The scientific research potential of virtual worlds. *Science*, *317*(5837), 472–476.

Bandiera, O., Barankay, I., & Rasul, I. (2005, August). Social preferences and the response to incentives: Evidence from personnel data. *The Quarterly Journal of Economics*, *120*(3), 917–962.

Bandura, A., & McClelland, D. C. (1977). Social learning theory.

Biocca, F., & Levy, M. R. (1995). Virtual reality as a communication system. *Communication in the age of virtual reality*, 15–31.

Blascovich, J., & Bailenson, J. (2011). Infinite reality: Avatars, eternal life, new worlds, and the dawn of the virtual revolution.

Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002). TARGET ARTICLE: Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry*, *13*(2), 103–124.

Bombari, D., Schmid Mast, M., Canadas, E., & Bachmann, M. (2015, 12 June). Studying social interactions through immersive virtual environment technology: virtues, pitfalls, and future challenges. *Frontiers in Psychology*, *6*.

Brenton, H., Gillies, M., Ballin, D., & Chatting, D. (2005). The uncanny valley: does it exist. In *Proceedings of conference of human computer interaction, workshop on human animated character interaction.* Citeseer.

Cruz-Neira, C., Sandin, D. J., & DeFanti, T. A. (1993). Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the CAVE. In *SIGGRAPH '93: Proceedings of the 20th annual conference on Computer graphics and interactive techniques* (pp. 135–142). New York, NY, USA: ACM. doi: http://doi.acm.org/10.1145/166117.166134

Cummings, J. J., & Bailenson, J. N. (2015). How immersive is enough? a meta-analysis of the effect of immersive technology on user presence. *Media Psychology*, *0*(0), 1–38.

DeHoratius, N., Gürerk, Ö., Honhon, D., & Hyndman, K. B. (2015, 20 October). Understanding the behavioral drivers of execution failures in retail supply chains: An experimental study using virtual reality. *Chicago Booth Research Paper No. 15-47*.

Dotsch, R., & Wigboldus, D. H. (2008, July). Virtual prejudice. *Journal of Experimental*

Social psychology, *44*(4), 1194–1198.

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

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*(2), 117–140.

Fiore, S. M., Harrison, G. W., Hughes, C. E., & Rutström, E. E. (2009, January). Virtual experiments and environmental policy - frontiers of environmental and resource economics. *Journal of Environmental Economics and Management*, *57*(1), 65–86.

Fischbacher, U. (2007, 7 February). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178.

Gravert, C., & Gerhards, L. (2016). Because of you I did not give Up-How peers affect perseverance. *Available at SSRN*.

Greiner, B. (2004). An online recruitment system for economic experiments. , *63*, 79–93.

Gürerk, Ö., Bönsch, A., Braun, L., Grund, C., Harbring, C., Kittsteiner, T., & Staffeldt, A. (2014, December). Experimental economics in virtual reality.

Gürerk, Ö., Grund, C., Harbring, C., Kittsteiner, T., & Staffeldt, A. (2015). Real effort on the holodeck: Experimental economics in virtual reality. *http://ssrn.com/abstract=2544739*.

Harrison, H. E., GW, & Rutström, E. (2011). Remarks on virtual world and virtual reality experiments. *Southern Economic Journal*.

Herbst, D., & Mas, A. (2015, 30 October). Peer effects on worker output in the laboratory generalize to the field. *Science*, *350*(6260), 545–549.

Kandel, E., & Lazear, E. P. (1992, August). Peer pressure and partnerships. *Journal of Political Economy*, *100*(4), 801–817.

Kasap, Z., & Magnenat-Thalmann, N. (2007). Intelligent virtual humans with autonomy and personality: State-of-the-art. *Intelligent Decision Technologies*, *1*(1), 3–15.

Kuhlen, T. W. (2014). Exploring Virtuality: Virtualität im Interdisziplinaren Diskurs. In S. Jeschke, L. Kobbelt, & A. Dröge (Eds.), (p. 133-147). Springer Fachmedien Wiesbaden.

Loomis, J. M., Blascovich, J. J., & Beall, A. C. (1999). Immersive virtual environment technology as a basic research tool in psychology. *Behavior Research Methods, Instruments, & Computers*, *31*(4), 557–564.

Manski, C. F. (1993, 1 July). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies*, *60*(3), 531–542.

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

McCall, C., & Singer, T. (2015, 12 February). Facing off with unfair others: introducing proxemic imaging as an implicit measure of approach and avoidance during social interaction. *PloS one*, *10*(2), e0117532.

Mori, M., MacDorman, K. F., & Kageki, N. (2012, June). The uncanny valley [from the field]. *IEEE robotics & automation magazine / IEEE Robotics & Automation Society*, *19*(2), 98–100.

Slater, M., Antley, A., Davison, A., Swapp, D., Guger, C., Barker, C., ... Sanchez-Vives, M. V. (2006, 20 December). A virtual reprise of the stanley milgram obedience experiments. *PloS one*, *1*, e39.

Slater, M., Usoh, M., & Steed, A. (1994). Depth of presence in virtual environments. *Presence*, *3*(2), 130–144.

Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence Teleoperators and Virtual Environments*, *6*(6), 603–616.

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

USC Institute for Creative Technologies, U. o. S. C. (2014). *Smartbody.* ”`http://smartbody.ict.usc.edu/`”. ([Online; accessed 30-May-2015])

Usoh, M., Catena, E., Arman, S., & Slater, M. (2000). Using presence questionnaires in reality. *Presence: Teleoperators and Virtual Environments*, *9*(5), 497–503.

Van Veldhuizen, R., Oosterbeek, H., & Sonnemans, J. (2014). Peers at work: From the field to the lab.

Windows, M. (2011). *Kinect For Windows.* ”`http://www.microsoft.com/en-us/kinectforwindows/`”. ([Online; accessed 30-May-2015])

A Instructions (Translated From German)

Welcome to today's experiment in the aixCAVE. Please read the following instructions carefully. If you have a question, please ask the instructor. Please turn off your mobile phone!

Today's experiment consists of an **introduction**, a **testing part**, **main part**, and a concluding **questionnaire**. The introduction takes about 5 minutes and will familiarize you with how to execute today's **work task** in the aixCAVE. In the subsequent **testing part** you work exactly **5 minutes** without assistance. The introduction as well as the testing part are **not relevant for your payoff**. The **main part** starts immediately after the **introduction** and takes exactly **10 minutes**. You will work on the same tasks as in the testing part. Different from the testing part, you will receive a payment in the main part, as explained in detail below.

Your work task

Today, you will work at a conveyor belt. Your working area is the area in front of the conveyor belt. Your task is **to find as many defect objects as possible on the conveyor belt and to remove them**. In order to do so, you grasp the object and rotate it with your hand carefully. The objects are represented as blue cubes. The object is **OK (non-defect)** if all **six sides** of the cube are **blue**. If **one** side of the cube is colored **red**, than the object is **defect**. **You have to take away a defect object from the conveyer belt and to put it into the bin that is designated for this object**. Non-defect objects must remain on the conveyor belt. The **work output** measures your performance in this task.

Your work output

The following apply:

- Each **defect** object you put into the bin **increases** your work output by 1 unit.
- Each **non-defect** object in the bin **decreases** your work output by 1 unit.

Introduction

- You work several minutes on the task described above.
- You should get used to the (technical) implementation of the task.
- You can ask the present instructor questions.
- You receive for this part of the experiment no payoff.

Testing Part

- You work exactly 5 minutes on the task described above.
- You should achieve a working result as high as possible.
- You receive for this part of the experiment no payoff.

Main Part of the experiment Relevant for payoff

- You work exactly 10 minutes on the task described above.
- You should achieve a working result as high as possible.
- You receive for this part of the experiment a payoff.
- You will be informed about your payoff at the beginning of the main part.

The Questionnaire

- A short questionnaire follows.
- You will receive your payoff from the experiment in cash, after completing the questionnaire.

Please follow the points below while you are inside the aixCAVE:

- Please enter the aixCAVE with the slippers provided here.
- Please avoid quick or hectic movements in the aixCAVE.
- Please do not touch the walls or the floor of the aixCAVE.

We wish you success!

Information: The main part of the experiment begins! (Given to the subjects before the treatment phase starts.)

- You work exactly 10 minutes on the task.
- You should achieve a work output as high as possible.
- You will be payed for this part of the experiment.
- Your payoff is 12 Euro, independent of the level of your work output.

Information: In the main part, you have a co-worker who works at the same task as you do! (Given to the subjects before the treatment phase starts.)

- Your co-worker is computer programed.
- You both work independent of each other. This means in particular, that your work output does not depend on your co-worker's work output, and vice versa.

B Survey Questions About Subjects' Experience in the CAVE

The following contains detailed information on the questions reported in Table 1 about subjects' experience in the virtual environment. In all cases, the measure is a 7-point Likert scale with higher numbers indicating a greater sense of realism in the virtual environment. The questions were adapted from slater1994depth.

1. **Sense of Being:** Please rate your sense of being in the virtual environment, on a scale of 1 to 7, where 7 represents your normal experience of being in a place. I had a sense of "being there" in the virtual environment *not at all* (1) – *very much* (7).
2. **Virtual Environment Was Reality:** To what extent were there times during the experience when the virtual environment was the reality for you? There were times during the experience when the virtual environment was the reality for me *at no time* (1) – *almost all the time* (7).
3. **Saw vs. Visited:** When you think back to the experience, do you think of the virtual environment more as images that you saw or more as somewhere that you visited? The virtual environment seems to me to be more like *images that I saw* (1) – *somewhere that I visited* (7).
4. **Where Were You/Sense of Location:** During the time of the experience, which was the strongest on the whole, your sense of being in the virtual environment or of being elsewhere? I had a stronger sense of *being elsewhere* (1) – *somewhere that I visited* (7).
5. **Memory Structure:** Consider your memory of being in the virtual environment. How similar in terms of the structure of the memory is this to the structure of the memory of other places you have been today? By "structure of the memory" consider things like the extent to which you have a visual memory of the virtual environment,

whether that memory is in color, the extent to which the memory seems vivid or realistic, its size, location in your imagination, and other such structural elements. I think of the virtual environment as a place in a way similar to other places that I've been today *not at all* (1) – *very much so* (7).

6. **Physically in Virtual Environment:** During the time of your experience, did you often think to yourself that you were actually in the virtual environment? During the experience, I often thought that I was really standing in the virtual environment *not very often* (1) – *very much so* (7).

The following contains detailed information about the questions reported in Table 2 about subjects' perceptions with the treatments. In all cases, the measure is a 7-point Likert scale with higher numbers indicating greater agreement with the question.

1. **Enjoyed Cave:** I enjoyed the experience in the cave. ((1) Do not agree at all; (7) agree fully)

2. **Enjoyed Task:** I enjoyed the execution of the task. ((1) Do not agree at all; (7) agree fully)

3. **Exerted Effort:** I exerted much effort in the execution of the task. ((1) Do not agree at all; (7) agree fully)

C Survey Questions About Subjects' Perception of the Avatar

The following contains detailed information about the questions reported in Table 2 about subjects' perceptions of the avatar.

1. **Avatar's productivity:** What do think about the avatar's work output. ((1) It was not good at all; (7) it was pretty good)

2. **Avatar moves natural:** Please give a rating, to which extent the avatar moves naturally. ((1) Not naturally at all; (7) pretty much naturally)

D Survey Questions About Subjects' Willingness to Compete with the Avatar

1. **Success:** Was it important to you to be more successful than the avatar. ((0) No; (1) Yes)