VR vs. Lab: A Study on the Opportunities of Experimental Economics in Virtual Reality

Completed Research Paper

Felix Brandstetter*, Chiara Krisam*, Anton Rewitzer*, Sebastian Riedinger*, Janika Vödisch*

Introduction

Whether shopping, learning, or working–more and more areas of daily life are taking place on the Internet. One term that is currently gaining massively in importance is the concept of the *Metaverse*. Metaverse is a collective term for a digital, three-dimensional world of experience where people come together to play, shop, meet with colleagues, attend concerts, etc. (Höfler & Kroller, 2022). Mark Zuckerberg even describes the Metaverse as "the next chapter for the Internet" and predicts that "within the next decade, the Metaverse will reach a billion people, host hundreds of billions of dollars of digital commerce, and support jobs for millions of creators and developers" (Zuckerberg, 2022). One of the ways to access the Metaverse is through the technology of *Virtual Reality* (VR). VR enables the deployment of virtual worlds in which people can interact using VR headsets and avatars. Thus, VR helps to create a completely new level of digital life.

In summary, Metaverse has numerous potential application scenarios. In the context of Experimental Economic research, the question arises whether one could prospectively conduct economic experiments in the Metaverse via VR. For example, one could perform experiments such as the Ultimatum Game, the Trust Game, or the Dictator Game by enabling interactions between people in the Metaverse. Innocenti (2017) investigated this question and examined the use of VR as a laboratory tool in Economics. According to Innocenti (2017), VR offers the opportunity to specifically incorporate contextual cues into experiment design. This overcomes an important limitation of the standard laboratory approach in Economics as well allows for a greater degree of control compared with field experiments (Fiedler & Haruvy, 2009). In search of the ideal experimental environment, many studies are concerned with the distinction between laboratory experiments and field experiments. The most frequently discussed aspects are usually control and realism (Fiedler & Haruvy, 2009). While laboratory experiments allow a high degree of control, field experiments feature more realism (Fiedler & Haruvy, 2009). Thus, one of the characteristics of laboratory experiments is high *internal validity*, while field experiments benefit from high *external validity* (Kornmeier, 2007). Therefore, one can argue that experiments in VR may combine the advantages of laboratory and field experiments by overcoming the typical trade-off between control and realism (Fiedler & Haruvy, 2009) respectively the trade-off between internal and external validity. Meißner et. al (2019) support this idea and additionally show that the use of VR can increase *ecological validity*. Specifically, this means that the use of VR creates an opportunity to generalize the study results to other similar situations (Peukert, Pfeiffer, Meißner, Pfeiffer, & Weinhardt, 2019). Furthermore, VR offers the possibility to change the experiment environment easily and flexibly during experiments. For example, experimenters could adjust wall colors or room sizes more easily in VR than in the laboratory or the field.

The use of VR as an alternative experimental environment thus offers many potential advantages for experimental economics research. However, the question arises if one can transfer the results of economic experiments in VR to the results in the classical physical laboratory. Only if the economic behavior in VR and the laboratory does not differ, it is possible to draw conclusions from an experiment in VR to the results of the same experiment in the laboratory. Therefore, it must be shown that experiment participants make the same economic decisions in both environments. Previous studies have already shown that while VR creates an artificial world, it still triggers real-world feelings (Karabasz, 2020). However, it has not yet been shown that people make the same economic decisions in VR and the laboratory. The purpose of the present work is thus to investigate in an exploratory way whether the results of a classical economic laboratory

*Email: Firstname.Surname@student.kit.edu

experiment can be replicated when conducting an experiment in VR. Further, it is explored whether a VR environment and thus the advantages of VR can be used for experimental economic research. In summary, this study thus examines the following **research question**:

1. When conducting experiments in VR, to what extent can the results from classical physical economic laboratory experiments be replicated, while leading to more perceived enjoyment and higher attention?

To answer this research question, this study conducts an experiment with two groups, both playing the Dictator Game. The first group conducts the experiment as a classical laboratory experiment and plays the Dictator Game on the desktop. The second group conducts the experiment in VR, meaning each of the participants wears VR glasses and plays the Dictator Game in the VR environment. To additionally investigate and compare the factors of *perceived enjoyment, attention,* and *intention to participate again* between VR experiments and laboratory experiments, we ask the participants to fill out questionnaires after the Dictator Game. In a subsequent analysis, we compare both the results in these factors and the behavior in the Dictator Game of both groups. In our experiment, the participants' decisions were equal between a group in a VR environment and a group in a classical laboratory. In the VR treatment, the physical discomfort is significantly higher than in the control group, yet the perceived enjoyment is also significantly higher. Thus, the VR is more attractive for the participants, making it more likely that they will attend again.

The remainder of the paper has the following structure: Section 2 provides definitions for VR as well as existing findings from other VR studies. Section 3 presents the initial hypotheses regarding the results of this study. The following Section 4 describes the methodology, focusing on the experimental procedure and the statistical measurement for evaluating the experimental results. After section 5 summarizes the results of the study, section 6 discusses these results and lists the limitations.

Theoretical Background and Related Work

Theoretical Background. From a technological point of view, VR is a computer-generated environment in which people act in a real-time simulated environment generating artificial places through an interface stimulating one or more senses (Innocenti, 2017). The created environment appears synchronously with the user's actions and tracks the user's movements. Thus, VR enables the simulation of situations and tasks under the precise control of the state experienced by the user. VR technologies can be applied to two types of environments, which differ in the degree of perceived immersion. *Low-immersive virtual environments* (LIVE) are computer screen-based renderings of real-world environments or virtual worlds (De Witt, 2018). In these, users interact via digital models called avatars that embody their virtual selves using computer features like a keyboard, a screen, a touchpad, or a mouse. In LIVE experiments factors like anonymity and virtual identity present potential disadvantages, because users may participate multiple times, may falsify their identity, and state or act in groups by changing avatars, e-mail, or IP address. On the other side, virtual worlds enable more diverse subject pools than standard lab populations. Examples of LIVE include Second Life and The Sims (De Witt, 2018). *Highly immersive virtual environments* (HIVE) use, among other things, head-mounted displays (VR glasses) to enter the VR. By using these VR glasses instead of a computer screen, it is possible to replace the user's real environment so convincingly that they can suspend their disbelief and fully engage with the created environment. This results in a higher perceived immersion compared to LIVE.

Another dimension of presence strongly connected with VRs is the so-called perceived telepresence. Telepresence describes the state of "(…) the perception of being present in an unreal environment." (Peukert, Pfeiffer, Meißner, Pfeiffer, & Weinhardt, 2019). Furthermore, technical equipment assures a full control on of the experience of the subjects and therefore over the stimulus presentation. Regarding the high degree of immersion in HIVE experiments, often the question arises whether subjects isolated from the real world by VR glasses helmets or caves are responsive to the design created by the experimenter and therefore are able tocan focus solely on the experimental stimuli (Innocenti, 2017). However, studies such as those by Fox & Bailenson (2019) have shown that although immersion dominates subjects' perception during an investigation, it does not prevent them from being fully active and cognitively engaged (Fox & Bailenson, 2019; Innocenti, 2017). Nevertheless, the use of VR (especially HIVE) technology can lead to so-called cybersickness, defined as a "(…) side effect experienced by users of immersive interfaces commonly used for Virtual Reality." (Nesbitt & Nalivaiko, 2018, p. 1). It is associated with symptoms such as nausea, postural instability, disorientation, headaches, eyestrain, and tiredness (Nesbitt & Nalivaiko, 2018).

When conducting experiments in VR, one can use diverse types and technologies of VR glasses. Amin et al. (2016) conducted a study to examine the difference in the level of immersion between a *Cardboard head-mounted display (HMD)* and a *traditional HMD*. A Cardboard HMD simply uses the smartphone as a display. In the simplest case, these VR glasses are do-it-yourself systems with thin cardboard. More advanced models made of plastic consequently offer greater stability. The lenses built into the front of the frame serve as a magnifying glass and enlarge the image so that it fills almost the entire viewing. To convey a spatial impression, the inserted smartphone shows a slightly offset image on the display for both eyes. Traditional VR glasses integrate the screens directly into the glasses. Thus, traditional HDMs offer significantly better resolution and a larger field of view than Cardboard HMDs and take a fundamentally different technological approach (Amin, Gromala, Tong, & Shaw, 2016). Nevertheless, Amin et al. (2016) were able to show that "(…) the Cardboard VR, despite its simplicity and small screen size, is capable of providing an acceptable level of immersion compared to (…) the larger screen size" of a traditional HMD (Amin, Gromala, Tong, & Shaw, 2016, p. 269). Mehrfard et al. (2019) also examined VR glasses and compared different VR technologies in terms of text readability, comfort, display quality, contrast perception, and compatibility with wearing glasses. The study shows that the technologies examined performed differently in the criteria (Mehrfard, et al., 2019). This should always remain a consideration when conducting studies with VR.

Related Work. Regarding the focused research question, a couple of studies already dealt with the question of comparability between laboratory and VR experiments by transferring different economical games into a VR environment. These studies were partially conducted in LIVE or HIVE. Fiedler & Haruvy (2009) investigated the impact of personal interaction through a virtual world on the Trust Game in a LIVE. They conducted a study of Trust Games in Second Life and contrasted the results with laboratory studies. In addition to the specific environment, personal interaction and the subject pool served as treatment variables. The authors showed that personal interaction through a virtual world interface significantly increases the amount sent relative to laboratory results. Furthermore, subjects recruited in the virtual world significantly give and return less than the control group of the laboratory. In a second study, Fiedler et al. (2011) investigated the impact of social distance on economic choices, also using the Trust Game in a LIVE (Second Life). Fiedler et al. (2011) point out that the experiment participants act more socially in a LIVE than in the classic laboratory environment. Both studies indicate that participants behave differently during an experiment in a LIVE than in a laboratory experiment. Based on Fiedler & Haruvy's (2009) and Fiedler et al. (2011) studies, it would not be possible to transfer the results of a VR experiment to the laboratory. Chesney et al. (2009) come to different conclusions in their study. They use five standard economic experiments, such as the Dictator or Ultimatum Game, to investigate whether subjects' virtual behavior in LIVE (Second Life) is consistent with previous laboratory findings of other studies. In addition to the experimental platform, the distinctive subject pools of previous laboratory studies also served as treatment variables. Controlling for age, Chesney et al. (2009) were unable to detect significant and systematic overall differences between the behavior in Second Life as well as in traditional laboratory settings. In HIVE, there are few studies so far that have conducted economic experiments in a virtual world using VR glasses. Gürerk & Kasulke (2018) investigated the impact of a HIVE on empathy and donation. As a result of their study, they found that a VR environment increased empathy and, in some cases, may also lead to higher average donations.

These studies do not provide a clear answer to the question of how the results of a VR experiment differ from the results of a classical laboratory experiment. One possible reason for the different results is that in the studies presented, other treatment variables besides the experimental environments "laboratory" and "VR" influence the results. In addition, it is important to note that only a few studies considered a HIVE via VR glasses, which will serve as the VR environment for this work. This may be due to the fact that HIVE is still in its early days. Therefore, this study conducts exploratory research with a HIVE to investigate the extent to which the results of the VR experiment differ from those of the laboratory experiment. The experimental results in both environments may further converge due to the higher immersion in a HIVE. The user's real environment may be replaced so convincingly that they can fully engage with the created environment and therefore behave as if they were in the laboratory.

Hypothesis

To use VR as a substitute for a laboratory, the perceived characteristics of virtual environments would have to be the same as those of the laboratory. Early on, researchers addressed environments such as VR give "an illusion that a mediated experience is not mediated" (Lombard & Ditton, 1997, p. 1). Thus, we no longer perceive VR as an intermediary medium. Studies with VR already demonstrate high perceived telepresence, which measures this phenomenon (Schnack, Wright, & Holdershaw, 2019). This would result in participants no longer being aware that they are in a virtual environment. Thus, if one conducts a classical economic laboratory experiment, in which one part of the participants sits in a virtual but identical-looking laboratory, the results of the two groups should be equal. Furthermore, (Chesney, Chuah, & Hoffmann, 2009) already conducted an experiment in which standard economic experiments were implemented in VR, as already described in the previous section. The difference was that the experiments did not take place in HIVE but in LIVE. In concrete terms, the experiments took place in the virtual environment "Second Life". It was shown, however, that standard economic experiments can be replicated in the environment without affecting the results. This leads to the first hypothesis.

> **H1:** The sent amounts of a Dictator Game performed in a virtual laboratory can be considered equal to those of a Dictator Game performed in a classical laboratory to a given level of significance.

As described before, one of the great advantages of the classical laboratory is that the experiments have a high validity because they offer good control. Other studies point out similar advantages of VR. A meta-study on studies in the educational context mentions amongst others the advantages of motivation, presence, and engagement (Pirker & Dengel, 2021). According to a study from 2017, VR also prevents distractions in children with attention deficit hyperactivity disorder (ADHD) (Bashiri, Ghazisaeedi, & Shahmoradi, 2017). Furthermore, a study by Cho et al. (2002) addressed the question of whether cognitive training in VR can be used for attention enhancement. The training was conducted by one treatment group in VR and the other group without VR. The researchers were able to show that the VR group increased their attention more than the other group. In the course of economic experiments, however, there is no study yet that refers to the effects of VR on attention. The question of whether VR represents an advantage to a conventional laboratory, which has a lot of control, is therefore still unanswered. VR could help control the environment even more strongly and easily than in the laboratory. This leads to the second hypothesis.

> **H2:** The attention of participants in the virtual laboratory is higher than that of participants in the traditional laboratory.

Methodology

Dictator Game. The experiment is based on the dictator game. This is played in different environments where the results are compared. "The Dictator Game is an experimental paradigm in which one participant (the dictator) receives an endowment and then decides to what extent she/he wants to split this endowment with another, anonymous participant (the recipient)." (Leder & Schütz, 2018, p. 1). The recipient cannot refuse the dictator's offer and thus assumes a purely passive role. The literature lists different variants of the Dictator Game and shows that the context can influence the behavior of the dictator. Other studies show that the dictator acts more socially in a more personal context, for example, when the dictator knows the recipient's deservingness. In this case, the dictator wants "(...) to draw upon the more personal context, be less homo economicus and more human." (Thunström, Cherry, McEvoy, & Shogren, 2016, p. 117). This shows that fairness and justice play a crucial role in the utility function of the dictator. Other studies show that the age of the dictator also plays a role. For example, older people tend to give away more of their money, while children divided fairly (Engel, 2010).

Treatments. The study investigates two different treatments: a laboratory group and a VR group. The between-subject design was chosen because the participants should not play the dictator game for too many rounds. If the dictator game is played multiple times, the previously played rounds may have an influence on the decisions of further rounds. Therefore, it was decided not to use a within-subject design. The experiment consists of two parts. First, the participants play the Dictator Dame, either in VR or on the computer, afterwards they fill out a questionnaire. We recruited participants for both treatments from a student pool. For each treatment, twenty participants take part in the experiment. We divided the total of forty participants into groups of four persons each, with one group containing only persons from the same treatment. In the laboratory group, participants sit in a soundproof and air-conditioned computer cabin equipped with a computer with screen, mouse, and keyboard during the experiment. The Dictator Game and the consecutive survey take place on this computer. The experiment for the VR group takes place in a highly immersive virtual environment (HIVE). For this treatment, we displayed the cabin as an 360° image in VR. Instead of the monitor the participants see a virtual screen in front of them (cf. Figure 1). The interactions for the game take place with the help of a cursor in the center of the field of view. By focusing the cursor on an option, they can select elements on the virtual screen. The participants in the VR environment wear a Cardboard HMD (VR Shark X6 with a Google Pixel 4a). Every fourth participants in the VR group conducted the experiment with a traditional VR HMD (HTC Vive Pro Eye). The Dictator Game was implemented for the laboratory group on the desktop using oTree (Chen, Schonger, & Wickens, 2016) and in the VR group using Unity, with similar interfaces in each case (cf. Figure 1, Figure 2).

Procedure. We invited both groups to the laboratory on site (cf. Figure 3). After arriving at the laboratory, the participants from both groups went to the laboratory cabins. There, they received an explanation of the process, rules, and payoff of the experiment on the computer. Both groups received an introduction to the Dictator Game and their role and task in it. Afterwards, the participants from the VR group put on the VR glasses. The experimenters assisted them to ensure that the glasses fit properly. Before the game started, participants in both groups received an explanation of how to interact with the corresponding environment (desktop or VR).

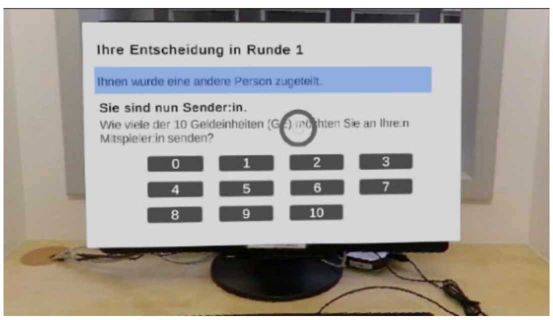

Figure 1 Unity interface of the VR group

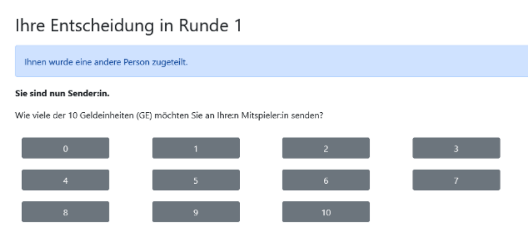

Figure 2 oTree interface of the laboratory group

Additionally, the first instructions page contained an attention check, according to Oppenheimer, Meyvis, & Davidenko (2009). Hereby we wanted to measure in which group the participants are more attentive. The Participants read that, to evaluate whether they were reading the instructions carefully, they should not click on the button labeled "Start the game." Instead, they were to click on the heading of the page to move to the next page. The attention check had no effect on the rest of the experiment nor on the payoff. The participants played the Dictator Game with a perfect stranger matching within their group. Prior to each round, they received the information that they were now playing with a new person. In each round, participants made the decision as to how much money to send to their teammate. The respective teammate in the round made this decision as well. Thus, in each round, the participants took both the role of the dictator and the role of the receiver. Participants receive ten monetary units each round, which they can divide up. One monetary unit is equal to ten cents in real life. To keep the input in the VR environment as simple as possible, they could only trade integer money units in both treatment groups. After each decision, participants see their own decision again before the next round begins. They do not see the other person's decision until the completed all rounds. This should ensure that each other's decisions do not influence the participants for the upcoming rounds and thus we can consider the rounds as separate decisions.

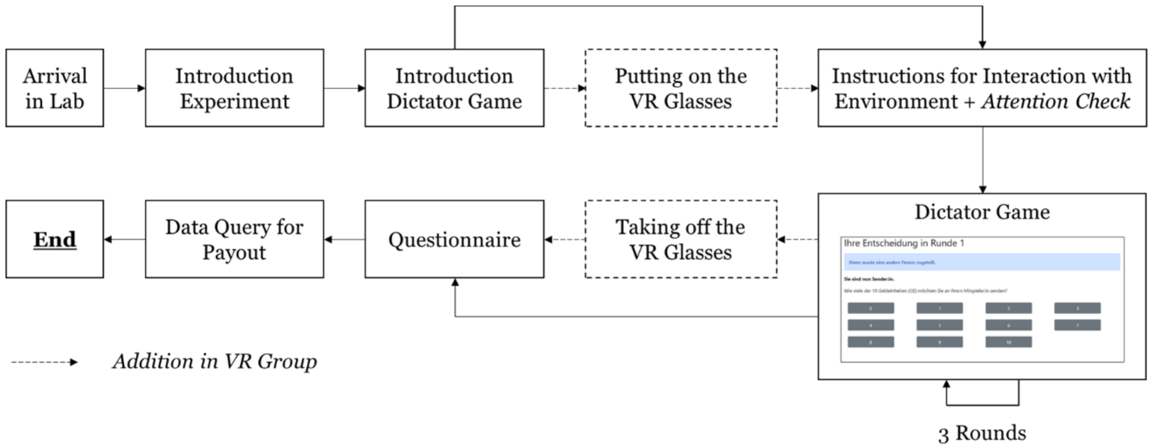

Figure 3 Design of the experiment for both treatments (VR and Laboratory)

After the Dictator Game, the participants in both groups filled out a questionnaire on a computer to measure their perceptions during the experiment. Here we examined, in addition to demographic variables, further factors of the environments, as already mentioned in the motivation. In addition, the participants had the opportunity to share further impressions from the experiment in a free text field. This allowed us to include factors noticed by the participants that did not appear in the questionnaire.

Payout. All participants received a fixed amount of 6.50 €. In addition, there was a performance-based amount of up to 3 € resulting from the dictator game. Therefore, in each round, one player from each pair was randomly selected. Their decision was the basis for the payout in that round. This means that they could win between 0 € and 1 € in one round. The total average payout was therefore 8 € per participant.

Measures. Our analytical procedure consists of the analysis of the data collected in the Dictator Game, the data collected via the applied questionnaires, and the data received due to the attention checks. In the case of the Dictator Game, the data are the allocations of monetary units determined by the dictators. Since we want to show that one can replicate laboratory experiments with VR (HIVE) experiments, we need to use a statistical test that can show the potential equality respectively equivalence of two independent samples. Mean value comparisons tests such as the t-test are, without further modification, only able to test for an inequality of the samples (Bortz & Schuster, 2005). A solution is the so-called *Two One-Sided Test Procedure* (TOST) (Schuirmann, 1987). As the name implies, this is a combination of two one-sided t-tests. One of the t-tests tests the null hypothesis (H_{01}) of whether the difference in group means is less than or equal to a lower interval limit. In a second t-test the null hypothesis (H_{02}) is tested whether the difference is greater than or equal to an upper interval limit.

$$H_{01}: \mu_{Lab} - \mu_{VR} \leq \theta_1 \qquad H_{02}: \mu_{Lab} - \mu_{VR} \geq \theta_2 \qquad \text{and} \qquad H_{11}: \mu_{Lab} - \mu_{VR} > \theta_1 \qquad H_{12}: \mu_{Lab} - \mu_{VR} < \theta_2$$

Thus, in the case of rejection of both null hypotheses (H_{01} and H_{02}), it can be shown that the difference is within a tested interval and the group means, depending on the established significance level, can be considered as equivalent ($\theta_1 < \mu_{Lab} - \mu_{VR} < \theta_2$) (Schuirmann, 1987). In order to perform statistical tests within a group, for example to investigate differences due to the effect of repeated measurements, depending on the distribution of the data, parametric or non-parametric mean comparison tests for dependent samples can be used (Bortz & Schuster, 2005). Regarding the necessary effect size for the evaluation of the Dictator Game, we followed the contributions of the meta-study by (Engel, 2010). In the analysis of several studies, it became visible that the realized effect sizes, which were sufficient to accept a rejection of the null hypothesis, very strongly. Regarding the sample size that we realized in this study, we followed the study by Brañas-Garza (2007) and set a Cohens d of 0.71 as targeted and sufficient benchmark.

We measured subjective perception during the experiment by standard questionnaires widely used in research. Since the experiment took place at a German university, we used German versions of the following questionnaires. Table 1 provides an overview of the standard questionnaires used. The exact questions are also included in the appendix. Using a seven-point Likert scale and four items we measured perceived telepresence (Nah, 2011). In addition to the individual items, we calculate a mean value of the four items to compare the two groups. The NASA Task Load Index (TLX) serves as a measure for mental demand, time demand, and frustration using a 20-point Likert scale, referencing to a total demand range between zero and 100 (Human Performance Research Group, 1988). The subjective evaluation of the technologies used (desktop or VR glasses) is done with the ten-item and five-level System Usability Scale (SUS), where the rated items are combined to an overall score between zero (low) and 100 (high) (Brooke, 1996). Within the SUS, we integrated an attention check item. For this purpose, we followed the proposals of Oppenheimer, Meyvis & Davidenko (2009). This also means that we excluded questionnaires from subjects who did not complete the attention checks correctly from the analysis. Perceived Enjoyment defined as "(...) the extent to which the activity of using a specific system is perceived to be enjoyable in its own right" (Venkatesh, 2000, p. 351), is measured using a seven-point Likert scale and the four items interest, enjoyment, excitement, and fun (Ghani, 1991). In addition to the individual items, we calculated a mean value of the items here to make the general factor enjoyment comparable. To record any physical discomfort during the experiment, we used the simulator sickness questionnaire (SSQ) (Kennedy, Lane, Berbaum, & Lilienthal, 1993). However, in this experiment not all items on the four-point Likert scale are queried, so that only the items general discomfort, fatigue, headache, eyestrain, nausea, blurred vision, and dizziness are relevant to this experiment (Kennedy, Lane, Berbaum, & Lilienthal, 1993).

In case of normal distributions and variance homogeneity, the parametric student's t-test for independent samples can be used (Bortz & Schuster, 2005). If no normal distribution but variance homogeneity can be verified, a non-parametric alternative, the Mann-Whitney U test, will be applied. If we cannot verify both, the normal distribution as well as the variance homogeneity, we will use Welch's t test. Normal distribution will be checked by performing the Shapiro-Wilk test as well as graphically by analyzing the Q-Q diagrams of the samples. To accept or reject the assumption of variance homogeneity Levene's test will be applied.

Table 1 Used standard questionnaires

Phenomenon	Measures	Source
Perceived telepresence	7-point Likert scale (-3,...,+3), 4 items	(Nah, 2011)
Mental demand, temporal demand, frustration	NASA Task Load Index (TLX) 20-point Likert scale, between 0 and 100	(Human Performance Research Group, 1988)
Subjective evaluation of the technologies	System Usability Scale (SUS)* 10 items, 5 level, between 0 and 100	(Brooke, 1996)
Perceived enjoyment	7-point Likert scale (+3,...,-3), 4 items	(Ghani, 1991)
Physical discomfort	Simulator Sickness Questionnaire (SSQ) 4-point Likert scale (1=none,..., 4=severe), 7 items	(Kennedy, Lane, Berbaum, & Lilienthal, 1993)

Results

Pretest. We assessed both treatments in a pre-test with a group of four participants each. At this point, the instructions for the experiment and the instructions for the game were still part of the VR environment. This means that the VR group started the experiment directly by putting on the VR glasses. It was not until the questionnaire that the participants removed the glasses and switched to the desktop. During the pretest, we noticed that the text was not well readable with the Cardboard HMD, resulting in mild headaches for some of the participants. We further observed that due to the actions in the VR environment, the VR image shifted increasingly to the right during the execution of the experiment. The cause was the inaccuracy of the accelerometer of the smartphone used. This shifting orientation may have caused nausea and dizziness for some participants. Therefore, we decided to take the instructions out of the VR environment and perform them on the desktop instead. Thus, the experiment for the VR group also started on the desktop and only switched to VR for the actual Dictator Game. In addition, we increased the font size in the VR environment to make it more legible. We suspected that the quality of the Cardboard HMD was one of the reasons for the poor readability. Therefore, we also decided that some participants should perform the experiment on the HTV Vive Pro Eye to have a comparison between two types of VR glasses. This traditional VR HMD has better resolution, a higher framerate, and a better acceleration sensor than the Cardboard HMD with the Google Pixel.

Results. The laboratory group consisted of thirteen men and seven women. The VR group, on the other hand, was composed of eleven men and nine women. A Chi² independence test showed that there was no significant correlation between the two variables group and gender (p = 0.519). For this reason, we did not perform a gender-related analysis in the following analyses. In the laboratory group, seven subjects were between the ages of 18 and 22, twelve were between the ages of 23 and 27, and one subject was over the age of 50. In the VR group, eleven subjects were between the ages of 18 and 22 and nine were between the ages of 23 and 27. Fifteen subjects in the laboratory group and eleven subjects in the VR group already had previous experience with VR. Furthermore, 19 laboratory subjects and 18 of the 20 VR subjects had had previously participated in an experiment at the Karlsruhe Decision and Design Laboratory (KD2Lab), a computer-based experimental laboratory used for this experiment.

Both the comparison of the results of the Dictator Game and the results of the different questionnaires were tested at a significance level of five percent ($\alpha = 0.05$). Since each subject made three decisions in a row within the Dictator Game, a sample size of N = 60 could be realized per group regarding the equivalence test for testing the first research hypothesis. To ensure a sufficient effect size in the analysis of the dictator game, we oriented on the realized effect sizes in a meta-study (Engel, 2010). Due to a power analysis conducted in advance, we could ensure a sufficient effect size of 0.61 (Cohen's d). Since we could not confirm the assumption of normal distribution and variance homogeneity, we used a Welch's t-test to perform the equivalence test. Although the average amounts sent were slightly higher for the VR group (3.05) than for the laboratory group (2.87), the null hypothesis could still be rejected at a five percent significance level (cf. Figure 3). Thus, regarding the first alternative hypothesis formulated in this study, we can reject the associated null hypothesis. To strengthen the assumption that each subject's three decisions are independent of each other and that the participants did not develop any strategies, we compared the average amount sent between rounds (within-subject design) for both the laboratory and VR group. Since we could not confirm the assumption of normal distribution and we compared measurement replications, we performed a Wilcoxon test at a five percent significance level in each case. We did not find any significant differences between the rounds for both groups, so that one can consider the independence assumption to be sufficient.

	t	df	p
t-test	−0.405	106	0.686
TOST Upper	−3.75	106	< .001
TOST Lower	2.94	106	0.002

	Low	High	90% Confidence interval Lower	Upper
Cohen's d	−0.610	0.610		
Raw	−1.51	1.51	−0.934	0.567

	N	Mean	Median	SD	SE
Lab	60	2.87	2.50	2.87	0.370
VR	60	3.05	3.00	2.01	0.260

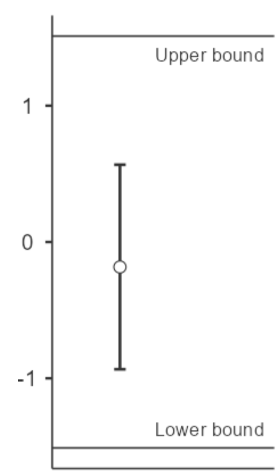

Figure 3: Descriptives and Results of the Equivalence test (TOST)

Eleven of the total twenty laboratory subjects successfully completed the attention check integrated into the experimental instructions of the Dictator Game. On the part of the VR group, only four of the twenty subjects were able to successfully complete this attention check. With respect to the second formulated alternative hypothesis, although we cannot perform statistical tests, the descriptive data clearly support not rejecting the associated null hypothesis. In contrast, all subjects in both groups answered the attention check integrated into the System Usability Scale for the survey correctly, so that we did not have to remove data records for the analysis of the questionnaires.

Table 2 presents the results of the analysis of the applied scales and questionnaires. Regarding the questionnaire perceived telepresence, the four included items (see Table A.1) showed a Cronbach's α value of 0.662, which initially does not indicate a high reliability. Since we could not increase the reliability by removing one or more of the four items and one tends to underestimate the reliability by using a Cronbach's alpha value to represent the internal consistency, one can consider this value to be acceptable. Perceived telepresence in the VR group (mean = -0.0375) was higher than in the laboratory group (mean = -0.713). Performing the one-sided student's t-test showed significant differences at a 5 percent level.

With reference to the three used items of the NASA-TLX, mental demand showed slightly higher results in the laboratory group (mean = 21.3) than in the VR group (13.8). In contrast, temporal demand (mean = 14.5) and frustration (mean = 17.0) was higher in the VR group than in the laboratory group (means = 6.5 and 15.3, respectively). Nevertheless, we did not find any significant differences (α = 5%) between the groups for any of the items.

The mean systems usability score (SUS score) was marginally higher for the laboratory group (mean = 85.6) than for the VR group (85.5). Using the two-sided respectively one-sided student's t-test ($p = .966$ and $p = .483$), we did not find any significant differences.

Regarding the factor perceived enjoyment, the four items interest, enjoyment, excitement, and fun showed a sufficiently high Cronbach's alpha value of .939. This value legitimizes not only to compare the individual items between the groups, but also the construct enjoyment in form of the mean values of the four items. In the left-oriented scale, the participants perceived all four items interest (mean = 2.2), enjoyment (mean = 2.15), excitement (mean = 1.4), and fun (mean = 1.8) more strongly in the VR than in the laboratory group (means = -.10, -.750, -.30, and -.350). In addition, the applied one-sided t-tests showed for all four items significant differences at the one percent level (see Table 2). In relation to the mean values of the four items, the construct perceived enjoyment is more pronounced in the VR group (mean = 1.64) than in the Lab group (-.375). Furthermore, applying the one-sided student's t-test showed significant differences at the one percent level.

A Cronbach's alpha value of .620 was computed for the seven of the sixteen item-sized SSQ used in this survey. Since we did not calculate a general simulator sickness value from the seven items in this analysis and Cronbach's alpha values tend to underestimate the actual reliability, we considered this value to be acceptable. All seven items general discomfort, fatigue, headache, eyestrain, nausea, blurred vision and dizziness were perceived more strongly on average in the VR group. In addition, by applying one-sided t-tests significant differences at the five percent level could be shown for the items nausea ($p = .048$) and dizziness ($p = .028$), as well as significant differences at the one percent level for the items general discomfort ($p = .002$), eyestrain ($p = .002$) and blurred vision ($p < .001$).

Table 2 Results of the analysis of the questionnaire

	Questionnaire/Items	Cronbach's α	Mean_{Lab}	Mean_{VR}	Normally distributed?	Variance homgeneity?	t-Test	p-value (two-sided)	p-value ($\mu_{Lab} > \mu_{VR}$)	p-value ($\mu_{Lab} < \mu_{VR}$)
	Perceived Telepresence (mean value of 4 items)	.662	-.713	-.0375	yes	yes	Student's t	.098	-	.049*
NASA-TLX	Mental demand	-	21.3	13.8	no	no	Welch's t	.169	.085	-
NASA-TLX	Temporal demand	-	6.5	14.5	no	yes	Mann-Whitney	.167	-	.083
NASA-TLX	Frustration	-	15.3	17	no	yes	Mann-Whitney	.48	-	.24
	SUS-Score (composed of 10 items)	-	85.6	85.5	yes	yes	Student's t	.966	.483	-
Perceived Enjoment	Interest	.939	-.10	2.2	no	no	Welch's t	< .001**	-	< .001**
Perceived Enjoment	Enjoyment		-.750	1.15	no	yes	Mann-Whitney	.002**	-	.001**
Perceived Enjoment	Excitement		-.30	1.4	yes	yes	Student's t	.003**	-	.001**
Perceived Enjoment	Fun		-.350	1.8	no	yes	Mann-Whitney	< .001**	-	< .001**
Perceived Enjoment	Perceived Enjoment (mean value of 4 items)	-	-.375	1.64	yes	yes	Student's t	< .001**	-	< .001**
Simulator Sickness	General discomfort	.620	1.0	1.35	no	no	Welch's t	0.005**	-	.002**
Simulator Sickness	Fatigue		1.05	1.2	no	no	Welch's t	.162	-	.081
Simulator Sickness	Headache		1.0	1.1	no	no	Welch's t	.163	-	.081
Simulator Sickness	Eyestrain		1.1	1.8	no	no	Welch's t	.003**	-	.002**
Simulator Sickness	Nausea		1.0	1.25	no	no	Welch's t	.096	-	.048*
Simulator Sickness	Blurred vision		1.0	2.1	no	no	Welch's t	< .001**	-	< .001**
Simulator Sickness	Dizziness		1.0	1.25	no	no	Welch's t	.056	-	.028*

*The symbols * and ** indicate a significance level at the 5, respectively 1 percent level.*

Regarding the comparison between Cardboard HMD (N = 15) and traditional VR glasses (N = 5), it must first be said that due to the different and small samples, no significant test is suitable for application and biases regarding the comparison of means must be accepted. On average, the traditional VR glasses produced a higher perceived telepresence (mean = .150) and perceived enjoyment (mean = 2.45) than the Cardboard HMD glasses (means = -.10 and 2.33). Regarding the items of the SSQ as well as the items of the NASA-TLX, no clear conclusions can be drawn. The SUS score was higher for the Lab Group (mean = 86.2) than for the VR Group (mean = 83.5). Further information on the results of the questionnaires can be found in Table A.1 in the Appendix.

Referring to the free comments submitted by the subjects, it was initially noticeable that only the VR group provided significant feedback in terms of content and quantity. Only two Lab subjects, but ten VR subjects left further comments. On the part of the Lab group the questions referred only to the sense of the experiment and the structure of the questionnaire. In the VR group, most of the comments consisted of more extensive questions regarding the meaning of the experiment as well as the optimal behavior in the Dictator Game. However, two test persons also criticized that the texts were difficult to read or that the focus control did not function optimally. In both cases the participants used Cardboard HMD glasses.

Discussion, Limitations and Further Research

Discussion. To answer the question, to what extent the results from a classical lab experiment can be replicated in VR, we showed that the execution of the Dictator Game in VR provides the same results as an execution of the Dictator Game in a desktop version. So, we demonstrated at one example, that it is possible. This insight fulfills a basic requirement for the usage of VR in Experimental Economics. Note that the comparison of the decisions between VR and the physical lab does not give insights into how the results of the treatments differ from the results in a real-life environment. However, as the results of the VR group are equal to the results of the laboratory group, they are equally good at explaining reality. Therefore, this experiment gives first indications that Experimental Economics in virtual labs might be a possibility in the future. Furthermore, our results indicate, that being in a VR environment does not change the amount of selfishness compared to decisions in front of a desktop.

The results did not confirm our expectation, that the attention of participants in the VR laboratory is higher than that of participants in the traditional laboratory. Thus, our results dismiss the supposition that conducting an experiment in VR leads to higher attention. An explanation for this finding could be that the higher temporal demand and frustration of the VR group leads to lesser attention, even though we found no significant differences between the two groups regarding these parameters. Furthermore, almost half of the participants in the VR group had not had any experience with VR up to that point. The attention check was on the first page in VR. It is, therefore, possible that the participants first had to familiarize themselves with the unfamiliar environment and that their attention was therefore not high enough for the attention check.

Regarding the user experience the significantly higher general discomfort, eyestrain, and blurred vision in the VR group are a sign of a higher physical discomfort through to the VR environment. This is a disadvantage for the usage of VR technology in Experimental Economics. Those factors might also be a cause for the higher, though not significantly higher, values of the VR group for temporal demand and frustration. However, this seems either not to have affected the results or other positively influencing factors offset those negative factors. One factor that could be relevant here is the significantly higher enjoyment of the VR group. Thus, a VR experiment is more attractive for the participants, making it more likely that they will attend future experiments. Although a significant comparison between the users of the Cardboard HMD and the traditional VR glasses is not possible, it was noticeable that the mean scale values mostly differed only to a small extent. This suggests that the Cardboard HMD glasses are a real, economical alternative to traditional VR glasses for future experiments.

Limitations and Further Research. We invited the participants into a physical lab and conducted the experiment inside a cabin. Therefore, they were physically in the same room as shown in VR. There might be changes in behavior if the virtual room differs from the physical room the participant is in. We did not investigate this factor. Furthermore, the invitation of all participants to the laboratory leads to the same amount of control in both treatments. If the subjects participate in the experiment in a VR environment from home, this control is not that high anymore. The influence of the lower control on the behavior should be evaluated before experiments are conducted without a physical lab. Moreover, the participants were only a few minutes inside the VR during the experiment. In more complex experiments it is necessary, that the subjects are much longer in a VR. This might have a significant impact on parameters such as the mental demand of the participants. Our results do not imply that such more complex experiments can also be carried out in VR. In addition, we only evaluated if we could replicate the results in the Dictator Game in a virtual lab. That does not implicate, that one can replicate other economic experiments. For example, there could be a lack of trust in a virtual environment and therefore, the decisions in a Trust Game could differ from the results in a physical laboratory. In the procedure of the experiment, the participants played three rounds of the Dictator Game and thus made three decisions. We assumed these decisions to be independent because they played the game in each round with a different person, and they did not see the decisions from the other players before they finished all rounds. However, this assumption is a limitation of our results. We also need to mention that many participants in the experiment are not used to the interface of the virtual lab. Therefore, they make their decisions in a less familiar environment than the participants of the control group. This circumstance can influence the results of the experiment. Hence, it could be interesting to conduct the experiment again with people who already have experience with VR. Moreover, they used a sample mostly consisting of students for the experiments. In General, students are affine to modern technologies which might influence the results and therefore, we cannot generalize the results to other demographic samples (Hartl & Berger, 2017). Besides, the possibility to change parameters of the virtual lab such as the size of the cabins or the color of the wall can be an advantage. We have not evaluated the influence of these parameters on the decisions. This can be an objective of investigation in the future. The next steps could be to conduct an experiment with a VR environment from home or changing parameters of the virtual laboratory.

Lessons Learned. During the experiment, it became clear that preparation and planning pays off. Because we had to supervise both treatments at the same time, we had to clearly distribute the tasks. In the days before the experiment, we discussed details and scenarios. This meant that in the event of minor problems during the experiment, the procedure was still clear, so the data collection went smooth. For the first sessions, we assigned participants to the treatments based on time of arrival. This meant that the participants who arrived first went into the VR group until four participants were in the group. After that, we filled the laboratory group. However, it became apparent that mostly women were among the earlier arriving participants, resulting in some of the laboratory groups being all male. We considered this in the upcoming sessions to achieve a more balanced gender distribution in the treatments. Therefore, we recommend for future experiments to assign the participants alternately to the treatments (first participant in treatment A, second in B, third again in A, ...). Furthermore, we noticed that there could be problems with the compatibility of VR glasses with normal eyeglass frames since you place the VR glasses over the normal glasses. This can lead to participants putting on the VR glasses incorrectly for large frames and the image becoming blurred. For further experiments, it can therefore be advantageous to ask the participants to wear contact lenses when inviting them. On the other hand, this problem can also occur in the "normal" use of VR glasses and experiments like this can highlight this problem. In addition, we noticed that we scheduled too much time for the experiment itself. In both treatments the duration was shorter than expected. A pretest with participants who do not know the experiment can help to have a more realistic time planning in the future. We noticed that it is much more strenuous to read the text in VR than on the Desktop. Researchers should keep this in mind when designing experiments in VR. They should use large font size, not too much text on a page, and keep the time participants spend in VR as short as possible.

Conclusion

We conducted a Dictator Game with participants either in a physical lab on the desktop or in a virtual lab in VR. Hereby, our research paper provides a first approach for the usage of VR in Experimental Economics and provides a basis for further research. We show that using a VR environment inside a physical laboratory does not change the results of a Dictator Game. That indicates that it is possible to replicate the results of classical Economic standard games in VR. However, as this is an explorative study, a lot of research is still necessary to certainly determine the effects of the usage of VR in Experimental Economics. This study lays a foundation for those investigations. If further research can approve our results, it would give great opportunities to experimental economic research. Researchers can conduct experiments in a VR environment, and they might not need a physical laboratory at all. It also indicates that it might be possible to take experiments into the Metaverse in the future.

References

Amin, A., Gromala, D., Tong, X., & Shaw, C. (2016). Immersion in Cardboard VR Compared to a Traditional Head-Mounted Display. In S. S. Lackey, *Virtual, Augmented and Mixed Reality* (pp. 269-276). Springer. doi: 10.1007/978-3-319-39907-2_25

Bashiri, A., Ghazisaeedi, M., & Shahmoradi, L. (2017). The opportunities of virtual reality in the rehabilitation of children with attention deficit hyperactivity disorder: a literature review. *Korean J Pediatr., Volume 60*, pp. 337-343. doi: 10.3345/kjp.2017.60.11.337

Bortz, J., & Schuster, C. (2005). *Statistik für Human- und Sozialwissenschaftler* (Vol. 7). Berlin: Springer. doi: 10.1007/978-3-642-12770-0

Brañas-Garza, P. (2007). Promoting helping behavior with framing in dictator games. *Journal of Economic Psychology, Volume 28*, pp. 477-486. doi: 10.1016/j.joep.2006.10.001

Brooke, J. (1996). SUS - A 'Quick and Dirty' Usability Scale. In J. Brooke, *Usability Evaluation In Industry* (pp. 189–194). Boca Raton: CRC Press. doi: 10.1201/9781498710411-35

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance, Volume 9*, pp. 88-97. doi: 10.1016/j.jbef.2015.12.001

Chesney, T., Chuah, S.-H., & Hoffmann, R. (2009). Virtual world experimentation: An exploratory study. *Journal of Economic Behavior & Organization, Volume 72*, pp. 618-635. doi: 10.1016/j.jebo.2009.05.026

Cho, B. H., Ku, J., Jang, D. P., Kim, S., Lee, Y. H., Kim, I. Y., & Kim, S. I. (2002). The effect of virtual reality cognitive training for attention enhancement. *CyberPsychology & Behavior, 5*(2), 129-137.

De Witt, D. (2018, 06. 13). Virtual reality introduction and basics. doi: n.a.

Engel, C. (2010). Dictator Games: A Meta Study. *Experimental Economics, Volume 14*, pp. 583-610. doi: 10.1007/s10683-011-9283-7

Fiedler, M., & Haruvy, E. (2009). The lab versus the virtual lab and virtual field—An experimental investigation of trust games with communication. *Journal of Economic Behavior & Organization, Volume 72*, pp. 716-724. doi: 10.1016/j.jebo.2009.07.013

Fiedler, M., Haruvy, E., & Xin Li, S. (2011). Social distance in a virtual world experiment. *Games and Economic Behavior, Volume 72*, pp. 400-426. doi: 10.1016/j.geb.2010.09.004

Fox, J., & Bailenson, J. N. (2019). Virtual Self-Modeling: The Effects of Vicarious Reinforcement and Identification on Exercise Behaviors. *Media Psychology, Volume 12*, pp. 1-25. doi: 10.1080/15213260802669474

Ghani, J. A. (1991). The experience of flow in computer-mediated and in face-to-face groups. *Proceedings of the of the International Conference on Information Systems (ICIS)*, (pp. 229-237). New York, USA.

Gürerk, Ö., & Kasulke, A. (2018). Does Virtal Reality Increase Charitable Giving? An Experimental Study. *SSRN Electronic Journal*. doi: 10.2139/ssrn.3072002

Hartl, E., & Berger, B. (2017). Escaping Reality: Examining the role of presence and escaping in user adoption of virtual reality glasses. *25th European Conference on Information Systems (ECIS)* (pp. 2413-2428). Guimarães, Portugal: Association for Information Systems AIS Electronic Library (AISeL). doi: n.a.

Höfler, N., & Kroller, H. (2022, 04. 28). *Was hinter dem Metaverse-Hype steckt*. doi: n.a.

Human Performance Research Group. (1988). *NASA TASK LOAD INDEX (TLX)*. California: NASA Ames Research Center. doi: n.a.

Innocenti, A. (2017). Virtual reality experiments in economics. *Journal of Behavioral and Experimental, Volume 69*, pp. 71-77. doi: 10.1016/j.socec.2017.06.001

Karabasz, I. (2020, 02. 19). *Im Körper eines anderen: So soll Virtual Reality die Empathie stärken*. doi: n.a.

Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology, Volume 3*, pp. 203-220. doi: 10.1207/s15327108ijap0303_3

Kornmeier, M. (2007). *Wissenschaftstheorie und wissenschaftliches Arbeiten*. doi: 978-3-7908-1918-2

Leder, J., & Schütz, A. (2018). Dictator Game. In *Encyclopedia of Personality and Individual Differences*. doi: 10.1007/978-3-319-28099-8_652-1

Lombard, M., & Ditton, T. (1997). At the Heart of It All: The Concept of Presence. *Journal of Computer-Mediated Communication, Volume 3*. doi: 10.1111/j.1083-6101.1997.tb00072.x

Mehrfard, A., Fotouhi, J., Taylor, G., Forster, T., Navab, N., & Fuerst, B. (2019). A Comparative Analysis of Virtual Reality Head-Mounted Display. doi: 10.48550/arXiv.1912.02913

Meißner, M. P. (2020). How virtual reality affects consumer choice. *Journal of Business Research, Volume 117*, pp. 219-231. doi: 10.1016/j.jbusres.2020.06.004

Nah, F. F.-H. (2011). Enhancing brand equity through flow and telepresence: A comparison of 2D and 3D virtual worlds. *MIS Quarterly, 35*(3), 731-744.

Nesbitt, K., & Nalivaiko, E. (2018). Cybersickness. In N. Lee, *Encyclopedia of Computer Graphics and Games* (pp. 1-6). Springer International Publishing. doi: 10.1007/978-3-319-08234-9_252-1

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, Volume 45*, pp. 867-872. doi: 10.1016/j.jesp.2009.03.009

Peukert, C., Pfeiffer, J., Meißner, M., Pfeiffer, T., & Weinhardt, C. (2019). Shopping in Virtual Reality Stores: The Influence of Immersion on System Adoption. *Journal of Management Information Systems, Volume 36*, pp. 755-788. doi: 10.1080/07421222.2019.1628889

Pirker, J., & Dengel, A. (2021). The Potential of 360° Virtual Reality Videos and Real VR for Education—A Literature Review. *IEEE Computer Graphics and Applications, Volume 41*, pp. 76-89. doi: 10.1109/MCG.2021.3067999

Schnack, A., Wright, M. J., & Holdershaw, J. L. (2019). Immersive virtual reality technology in a three-dimensional virtual simulated store: Investigating telepresence and usability. *Food Research International, Volume 117*, pp. 40-49. doi: 10.1016/j.foodres.2018.01.028

Schuirmann, D. J. (1987). A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, Volume 15*, pp. 657-680. doi: 10.1007/BF01068419

Thunström, L., Cherry, T., McEvoy, D., & Shogren, J. (2016). Endogenous context in a dictator game. *Journal of Economic Behavior & Organization, Volume 65*, pp. 117-120. doi: 10.1016/j.socec.2016.08.001

Venkatesh, V. (2000). Determinants of Perceived Ease of Use: Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model. *Information Systems Research, Volume 11*, pp. 342-365. doi: 10.1287/isre.11.4.342.11872

Wilson, C. J., & Soranzo, A. (2015). The Use of Virtual Reality in Psychology: A Case Study in Visual Perception. *Computational and Mathematical Methods in Medicine*, pp. 1-7. doi: 10.1155/2015/151702

Zuckerberg, M. (2022, 03. 15). *Mark Zuckerberg: NFTs are coming to Instagram before the metaverse is a thing*. doi: n.a.

Appendix

Table A.1 Results of the applied questionnaires

Construct	Code	Item	Mean Lab (N=20)	Mean VR Overall (N=20)	Mean VR Cardboard HMD (N=15)	Mean VR Traditional HMD (N=5)
		7-point Likert Scale from -3 (not at all) to +3 (completely)				
Perceived Telepresence (Nah, 2011)	PT01	I forgot about my immediate surroundings during the experiment.	-.250	.150	-.0667	.80
	PT02	When the experiment ended, I felt like I was coming back to the "real world" after a journey.	-1.25	0.6	.667	.40
	PT03	During the experiment, I forgot that I was in the middle of an experiment.	-2.1	-1.55	-1.40	-2.0
	PT04	The experimental environment seemed to be “somewhere I visited” rather than “something I saw.”	0.75	0.65	.40	1.4
	Construct	Mean value of the Items PT01 - PT04	-.713	-.0375	-.10	.150
		21-point Likert scale from 0 (low) to 100 (high)				
NASA-TLX (Human Performance Research Group, 1988)	TL01	Mental demand	21.3	13.8	12.0	19.0
	TL02	Temporal demand	6.50	14.5	15.7	11.0
	TL03	Frustration	15.3	17.0	19.7	9.0
		5-point Likert scale from -2 (not at all) to +2 (completely)				
SUS (Brooke, 1996) + Attention Check (AC) (Oppenheimer, Meyvis, & Davidenko, 2009)	SU01	I think that I would like to use the experienced experimental environment frequently.	3.15	3.7	3.73	3.6
	SU02	I found the experience experimental environment unnecessarily complex.	1.40	1.80	1.87	1.60
	SU03	I thought the experienced experimental environment was easy to use.	4.90	4.75	4.87	4.40
	SU04	I think that I would need the support of a technical person to be able to use this experienced experimental environment	1.05	1.3	1.33	1.20
	SU05	I found the various functions in this experienced experimental environment were well integrated.	3.6	4.25	4.33	4.00
	SU06	I thought there was too much inconsistency in this experienced experimental environment.	1.45	1.45	1.40	1.60
	SU07	I would imagine that most people would learn to use this experienced experimental environment very quickly.	4.65	4.5	4.60	4.20

	AC	This is an attention check, check the rightmost answer option for this question.	5.00	5.00	5.00	5.00
	SU08	I found the experienced experimental environment very cumbersome to use.	1.30	1.70	1.73	1.60
	SU09	I felt very confident using the experienced experimental environment	4.35	4.50	4.47	4.60
	SU10	I need to learn a lot of things before I could get doing with this system	1.20	1.25	1.20	1.40
	SUS Score	SUS Score computed according to Brooke (1996)	85.6	85.5	86.2	83.5
Perceived Enjoyment (Ghani, 1991)		7-point left-oriented Likert scale from +3 (appropriate) to -3 (inappropriate)				
	EJ01	Interest	-.10	2.2	2.27	2.0
	EJ02	Enjoyment	-.750	1.15	1.13	1.2
	EJ03	Excitement	-.30	1.4	1.47	1.2
	EJ04	Fun	-.350	1.8	1.8	1.8
	Construct	Mean value of the Items EJ01 - EJ04	-.375	1.64	1.67	1.55
Simulator Sickness (Kennedy, Lane, Berbaum, & Lilienthal, 1993)		4-point scale: (none), 2 (slight), 3 (moderate), 4 (severe) symptoms				
	SS01	General discomfort	1.0	1.35	1.4	1.2
	SS02	Fatigue	1.05	1.20	1.2	1.2
	SS03	Headache	1.0	1.1	1.07	1.20
	SS04	Eyestrain	1.1	1.8	1.73	2.0
	SS05	Nausea	1.0	1.25	1.27	1.20
	SS06	Blurred vision	1.0	2.1	2.33	1.40
	SS07	Dizziness	1.0	1.25	1.20	1.40