Article

Online Platforms for Remote Immersive Virtual Reality Testing: An Emerging Tool for Experimental Behavioral Research

Tobias Loetscher [1,*], Nadia Siena Jurkovic [1], Stefan Carlo Michalski [1,2], Mark Billinghurst [3,4] and Gun Lee [3]

1   Cognitive Ageing and Impairment Neurosciences Lab., Justice & Society, University of South Australia, Adelaide, SA 5000, Australia
2   School of Psychology, University of Sydney, Sydney, NSW 2006, Australia
3   Australian Research Centre for Interactive and Virtual Environments, STEM, University of South Australia, Adelaide, SA 5000, Australia
4   Empathic Computing Laboratory, The University of Auckland, Auckland 1010, New Zealand
*   Correspondence: tobias.loetscher@unisa.edu.au

Abstract: Virtual Reality (VR) technology is gaining in popularity as a research tool for studying human behavior. However, the use of VR technology for remote testing is still an emerging field. This study aimed to evaluate the feasibility of conducting remote VR behavioral experiments that require millisecond timing. Participants were recruited via an online crowdsourcing platform and accessed a task on the classic cognitive phenomenon "Inhibition of Return" through a web browser using their own VR headset or desktop computer (68 participants in each group). The results confirm previous research that remote participants using desktop computers can be used effectively for conducting time-critical cognitive experiments. However, inhibition of return was only partially replicated for the VR headset group. Exploratory analyses revealed that technical factors, such as headset type, were likely to significantly impact variability and must be mitigated to obtain accurate results. This study demonstrates the potential for remote VR testing to broaden the research scope and reach a larger participant population. Crowdsourcing services appear to be an efficient and effective way to recruit participants for remote behavioral testing using high-end VR headsets.

Keywords: crowdsourcing; virtual reality; online testing; reaction time; Prolific

1. Introduction

This paper explores the potential of using virtual reality (VR) for remote data collection. VR technology is being increasingly utilized by researchers as a tool for studying human behavior [1,2]. One reason for its appeal is the technology's ability to simulate realistic interactions in a safe and controlled environment. Such simulation enables researchers to study behavior in a more naturalistic settings than traditional computer-based approaches permit, while maintaining control over the environment [3,4]. This control allows for simple manipulation of variables to test their effects on behavior. The potential of VR for researching human behavior has long been recognized [4,5], but a surge in VR research studies has only occurred in the last few years, facilitated by technological advances and a decrease in the cost of VR systems [1].

Access to a wide range of participants can be a bottleneck for conducting VR research. For non-VR research, an established approach to recruiting participants is to conduct experiments remotely, with participants taking part from their homes using their own equipment [6–8]. Surveys and web-based behavioral experiments are advertised on online crowdsourcing platforms such as mTurk (http://mturk.com, accessed on 11 March 2023) and Prolific (http://prolific.co, accessed on 11 March 2023), where registered participants can partake and receive reimbursement for their time. The advantages of remote testing are the ease of access to participants, the speed of data collection, and the ability to recruit from a diverse population [6–8]. However, concerns have been raised about data quality

when using online platforms [8–10]. Factors such as bots completing the task, participant inattention, and personal misrepresentation can impact the data, and mitigation strategies are required to improve data quality [8–10]. Despite this, studies on data quality typically converge that the data quality is satisfactory in most cases [11,12], particularly when compared to the quality obtained from student convenience samples [11].

Remote testing using VR technologies is still an emerging field [13–17], and studies on its use are limited compared to the extensive literature on online testing using traditional devices such as phones, tablets, and computers. Early attempts at remote VR testing relied on phone-based VR systems, such as Google Cardboard and Samsung Gear VR, and recruitment of participants directly through web pages and emails [18,19]. Recently, there have been efforts to use social VR platforms to recruit participants and run remote VR experiments [20]. The use of social VR platforms has shown promise, but there are challenges in terms of privacy and data protection laws, as well as limitations in terms of experimental control [21]. Asking participants to install applications on their headsets (e.g., using SideQuest, [22]) is another viable option for remote VR testing. However, participants may be hesitant to install applications on their devices due to security concerns, or they may encounter issues during the installation process. The deployment of VR applications and controlling data collection is an ongoing challenge for remote testing [14].

Crowdsourcing platforms are not yet embraced by VR researchers. Previous studies reported that only few users with high-end VR headsets could be recruited via crowdsourcing platforms [14,18,23], which casted doubts about the usefulness of these platforms. However, the number of high-end VR headset owners is steadily increasing. The platform Prolific, for example, reports having a pool of more than 9200 people who own a VR headset and have been active on the platform in the last 90 days (information as of 11 Jan 2023). Although there is no information on the type of headset, it stands to reason that it might now be possible to recruit adequate numbers of participants with non-phone-based VR headsets on Prolific.

A challenging aspect of investigating attentional processes is achieving highly accurate timing for online data collection. The first challenge is related to technical aspects, such as programming web-based experiment protocols and ensuring that the participant and experimenter browser systems support the same features. The second challenge is related to memory and timing, which are sensitive to the environmental factors of the testing environment that are difficult to control online. It has been demonstrated that accurate timing can be achieved in lab-based VR studies [24], but it is unclear whether this can also be achieved for remotely conducted experiments that depend on highly accurate timing for response recording and stimulus presentation.

The aim of this study is to assess the viability of remotely conducting behavioral experiments in VR that depend on millisecond timing. We will replicate a classic cognitive phenomenon called "Inhibition of Return" with crowdsourced participants who will be using their own VR headsets or computer to complete a web-based experiment hosted on a dedicated research platform (Pavlovia.org, accessed on 11 March 2023). The main novelty of this work is that it is the first example of studying a time-sensitive cognitive phenomenon using crowdsourced remote VR participants.

Inhibition of return (IOR) is a cognitive mechanism that suppresses the brain from revisiting recently attended locations or objects [25–27]. This mechanism is thought to promote the exploration of new information by facilitating the detection and selection of novel stimuli [25]. There are many experimental variations of how IOR can be elicited [25,26]. This study uses a paradigm described in a seminal paper [28] that validated the use of the crowdsourcing platform mTurk for online behavioral testing. The paradigm involves the presentation of a target stimulus at various intervals (100, 400, 800, or 1200 ms) after a white cue is briefly flashed. IOR is observed if participants show faster target detection when the target is presented on the opposite side of the cue compared to when the target is presented on the same side as the cue. This IOR pattern was observed for trials with a minimum interval of 400 ms between cue and target [28]. If the interval between the

cue and target is short (e.g., 100 ms), a cueing effect is observed where participants show faster reaction times when the cue and target are presented on the same side, as opposed to when the cue is presented on the opposite side of the target. We predicted to replicate this pattern of results in remote participants completing the task using their none-phone based VR headsets or computers.

2. Materials and Methods

2.1. Participants

A total of 150 participants recruited through Prolific (www.prolific.co, accessed on 11 March 2023) completed the experimental task. Eligibility criteria for all participants were ownership of a VR headset and being fluent in English. Additional eligibility criteria for the VR group were (1) complete inhibition of return task in a non-phone-based VR headset, (2) having more than 5 h of VR experience, (3) not being prone to motion sickness, and (4) confirmation of participating seated and in a cleared area around them. The additional eligibility criteria 2–4 were included to ensure the safety of the participants. Informed consent was obtained from all participants. Ethics approval was provided by the University of South Australia Human Research Ethics Committee (approval number: 204737).

After inspection of the data, 14 participants were excluded from further analyses. Exclusion reasons were the use of a phone-based VR headset such as google cardboard (n = 5), the use of a phone or tablet instead of a VR headset or desktop computer (n = 8), and revoked consent after participation (n = 1). This left 136 participants in total, with 68 participants in each condition (see Table 1 for demographics).

Table 1. Participant demographic information.

	VR Group	Desktop Group
n	68	68
Age (SD) in years	28.9 (9.8)	30.3 (9.5)
Gender	16 female, 51 male, 1 non-binary	30 female, 38 male
Handedness	10 left, 58 right	7 left, 61 right
Country of Residence (Top 4)	South Africa (22) Poland (11) UK (9) Italy (6)	South Africa (24) Poland (10) Portugal (7) Italy (7)
Used Headset (Top 4)	Oculus Quest 2 (27) PlayStation VR (16) Oculus Quest (7) Oculus Rift (7)	
Hours spent in VR	between 6 and 50 h (30) between 51–100 h (16) more than 100 h (22)	

2.2. Measures

2.2.1. Demographics

A brief online survey was created on Qualtrics to obtain informed consent and the participants' self-reported age, gender, and handedness. Participants in the VR group were also asked about the VR headset they were using, their VR experience (up to 5 h, 6 to 50 h, 51 to 100 h, or above 100 h), proneness to cybersickness (yes/no), and whether they participate in a seated and cleared area around them (yes/no).

2.2.2. Inhibition of Return Task

The experiment was created with the open-source software PsychoPy [29] and modelled on the IOR paradigm described in [28]. Each trial began with the presentation of the

white outlines of two lateral boxes against a black background and a white fixation cross in the center of the screen (see Figure 1). Normalized screen units were used to place the boxes due to participants' different screen resolutions, with the screen window's bottom left and top right corners having the coordinates (−1,−1) and (1,1), respectively. The fixation cross was placed in the center of the screen (0,0), and the boxes to the left (−0.4,0) and right (0.4,0) of the screen center. After 1000 milliseconds (ms), one of the two boxes turned white for 100 ms, acting as a cue in the form of a white flash. A target letter 'X' would then appear in either of the lateral boxes until a response was made with the controller or any button on a keyboard. The time between the cue and appearance of the target (Stimulus-onset asynchrony (SOA)) was 100 ms, 400 ms, 800 ms, or 1200 ms. Importantly, the cue was non-informative as the target was equally likely (50% chance) to appear on the same or opposite side of the cue. If the cue and the target were presented on the same side, it was considered a valid cue. If the cue and target were on opposite sides, it was an invalid cue.
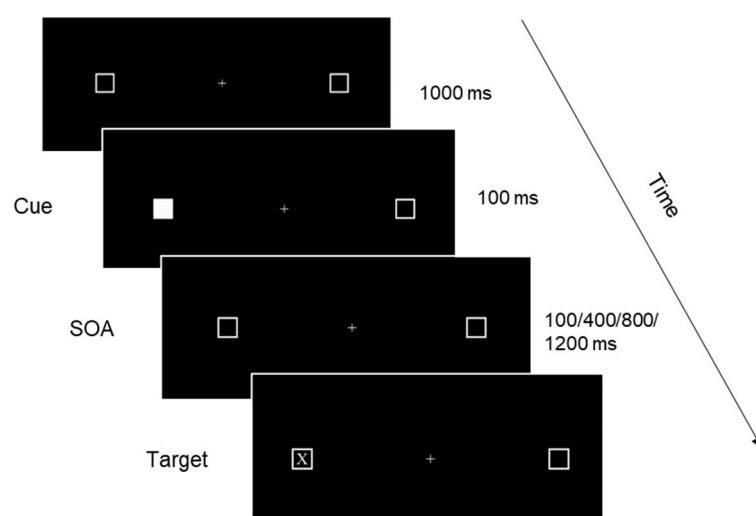

Figure 1. Stimulus train inhibition of return task. Example of valid cue trial with cue and target letter appearing on the same side. SOA = stimulus-onset asynchrony.

There were 96 trials in total, with the factorial combinations of SOA (4 levels: 100, 400, 800, 1200 ms), cue side (2 levels: left and right), and target side (2 levels: left and right) repeated six times. The dependent variable was reaction time (time from the appearance of the target letter to response; RT) and the independent variables were SOA and Cue Validity (valid, invalid cue). The presentation of the trials was randomized, and participants were given eight practice trials before starting the experimental trials.

2.3. Procedure

The experiment was advertised on Prolific for people who owned a VR headset and were fluent in English. Participants for the VR and desktop conditions were recruited separately, and those who participated in one condition (e.g., VR) could not participate in the other (e.g., desktop). Interested participants were directed to a Qualtrics survey to confirm their eligibility, provide consent to participate, and answer demographic questions (see Section 2.2.1). Eligible participants were then automatically directed to the experiment hosting platform Pavlovia (www.pavlovia.org, accessed on 11 March 2023), and returned to Prolific after completing the experimental task.

The median completion time for the VR and desktop conditions were 10.68 and 8.18 min, respectively. Participants were paid at a rate of around A$19 per hour.

3. Results

Data analysis and plotting were performed with Jamovi [30], R 4.2.2 [31], and a range of R packages, including tidyverse v.1.3.2 [32], trimr v.1.1.1 [33], and effectsize v.0.8.3.1 [34].

3.1. Data Preparation

To remove outliers, the modified recursive outlier procedure proposed by [35] was applied using the trimr package [33] with a minimum RT of 150 ms. This led to the exclusion of 4.3% of all trials.

3.2. Inhibition of Return

Initial visual inspection of the data distributions shows similar patterns for the VR and Desktop group (Figure 2) as a function of SOA and Cue Validity. However, the density curves are noticeably wider for the VR compared to the Desktop group (see Figure 2).
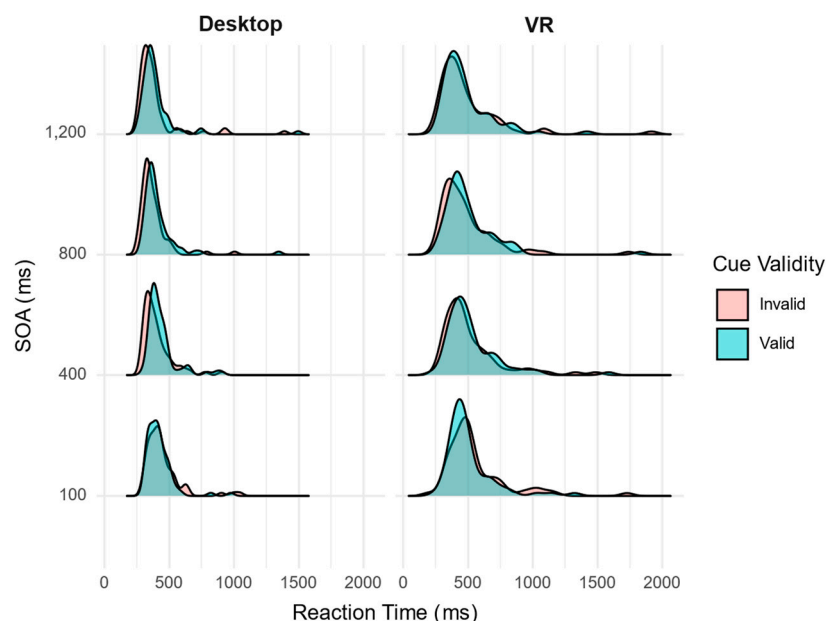

Figure 2. Density plots of valid and invalid cues with RT on the x-axis and the different SOAs (stimulus-onset asynchrony) on the y-axis. The left and right sides show the data for the Desktop and VR group, respectively.

To statistically assess Inhibition of Return, separate linear mixed-effects models were run for the VR and Desktop groups. In each model, RT was the dependent variable, SOA and Cue Validity were fixed factors with full interactions, and participant ID was modelled as random effects on the intercept. Simple effects were used in instances of significant interactions. Effect sizes were calculated from test statistics using [34].

For the Desktop group, there were main effects for Cue Validity ($F_{(1,6192)} = 19.4$, $p < 0.001$) and SOA ($F_{(3,6192)} = 30.6$, $p < 0.001$), which were qualified by a significant interaction between SOA and Cue Validity ($F_{(3,6192)} = 13.8$, $p < 0.001$). Simple effects analysis of the interaction revealed a significant cueing effect (faster RT after valid than invalid cue) for SOAs of 100 ms, and inhibition of return (faster RT after invalid than valid cue) for all other SOAs (see Table 2).

Table 2. Simple effects of Cue Validity (contrast valid -invalid) with SOA as a moderator level for the Desktop 95% CI provides estimates for lower and upper 95% Confidence intervals. * indicates a $p < 0.05$. Effect sizes were approximated with the use of test statistics [34].

SOA (ms)	Estimate	95% CI	t	p	Cohen's d
100	−22.7	−37.5 to −7.8	−3.00	0.003 *	0.04
400	33.9	19.1 to 48.8	4.48	<0.001 *	0.06
800	39.5	24.7 to 54.3	5.24	<0.001 *	0.07
1200	15.7	1.0 to 30.5	2.09	0.037 *	0.03

For the VR group, there was no main effect for Cue Validity ($F_{(1,6154)}$ = 2.39, p = 0.122), but a main effect of SOA ($F_{(3,6154)}$ = 22.24, $p < 0.001$), which was qualified by a significant interaction between SOA and Cue Validity ($F_{(3,6154)}$ = 11.63, $p < 0.001$). Simple effects analysis of the interaction revealed a significant cueing effect (faster RT after valid than invalid cue) for SOAs of 100 ms, and inhibition of return (faster RT after invalid than valid cue) for SOAs of 800 ms. No effects were found for SOAs of 400 and 1200 ms (see Table 3).

Table 3. Simple effects of Cue Validity (contrast valid -invalid) with SOA as moderator level for the VR Group. 95% CI provides estimates for lower and upper 95% Confidence interval. * indicates a $p < 0.05$. Effect sizes were approximated with the use of test statistics [34].

SOA (ms)	Estimate	95% CI	t	p	Cohen's d
100	−48.0	−64.6 to −31.4	−5.66	<0.001 *	0.07
400	5.6	−11.2 to 22.4	0.65	0.514	0.01
800	18.5	1.9 to 35.1	2.19	0.029 *	0.03
1200	−2.3	−18.9 to 14.2	−0.28	0.782	0.00

To compare RTs and RT variability between the Desktop and VR groups, two-sample unpaired rank-sum Mann–Whitney tests were used as there was a violation of the assumption of normality ($p < 0.001$ for Shapiro–Wilk tests). The VR group (Mdn 441 ms) had significantly slower RTs compared to the Desktop group (Mdn 374 ms, U = 1395, $p < 0.001$, r = 0.40). To assess timing variability, intra-individual RT variabilities (i.e., SDs) in the VR and Desktop groups were compared (see Figure 3). The individual SDs in the VR group (Mdn 80 ms) were significantly higher than in the Desktop group (Mdn 67 ms. U = 1808, $p < 0.028$, r = 0.22).
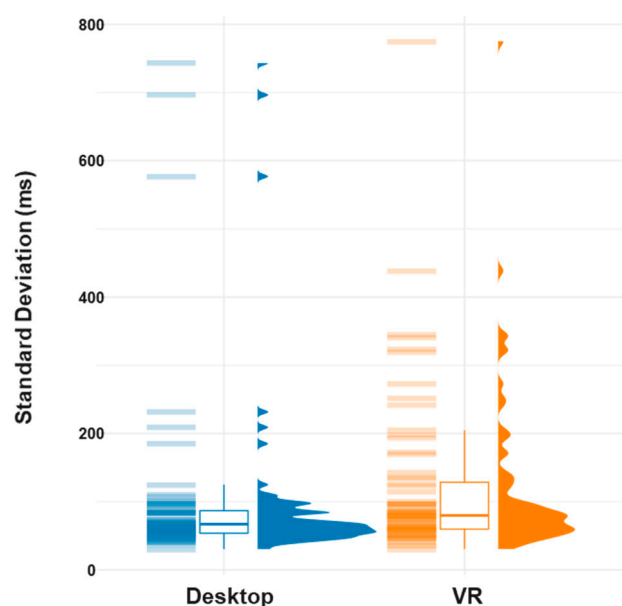

Figure 3. Distribution of intra-individual reaction time variability (SD) as a function of Group shown as strip plots (each line represents a participant), boxplots and density plots.

Having established a larger variability in the VR group, the intra-individual reaction time variability (SD) as a function of VR headsets was explored for headsets with at least five participants (Figure 4). A Kruskal–Wallis test revealed significant differences between headsets (χ^2 (3) = 17.6, $p < 0.001$, ε^2 = 0.32). Subsequent Dwass–Steel–Critchlow–Fligner pairwise comparisons revealed significant differences between Oculus Quest 2 and PlayStation VR (W = 1.61, $p < 0.001$). The comparison between Oculus Quest 2 and Oculus Quest did not reach significance (W = 3.58, $p < 0.055$), as did none of the other pairwise comparisons ($p > 0.15$).
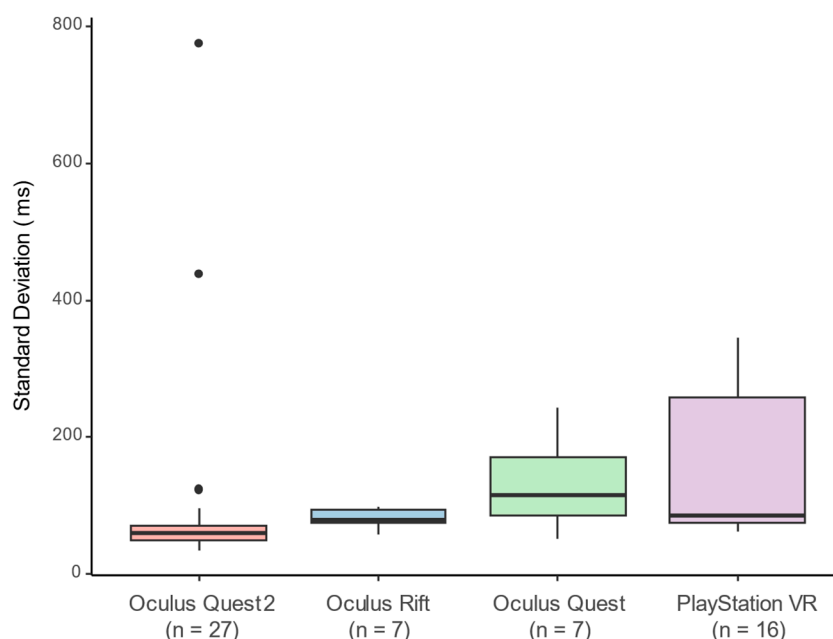
Figure 4. Boxplots of intra-individual reaction time variability (SD) as a function of headsets. Only headsets with five or more participants are shown.

An exploratory analysis of just the Oculus Quest 2 participants (n = 27) discovered a significant cueing effect for SOAs of 100 ms (p = 0.02) and an inhibition of return for SOAs of 800 ms (p = 0.04). SOAs of 400 and 1200 ms yielded no significant results. It is important to note that all exploratory headset analyses, including comparisons with Oculus Quest (n = 7) and Oculus Rift (n = 7) headsets, were underpowered and will need to be confirmed in a larger sample.

4. Discussion

The study aimed to evaluate the viability of conducting behavioral experiments in VR remotely with a primary focus on a cognitive effect that critically depends on timing precision. Results showed that the VR group exhibited the expected cueing effect at 100 ms and an IOR effect at 800 ms, but there was no statistically significant inhibitory effect at 400 ms and 1200 ms. The desktop group successfully replicated previous findings [28], with a cueing effect at 100 ms and IOR effects at all longer stimulus onset asynchronies. The study confirms that remote testing with desktop computers can be used to study time-critical cognitive phenomena, but the results with VR headsets were only partially replicated.

The overall pattern in the VR and desktop groups are very similar, but there was larger intra-individual reaction time variability in VR participants compared to desktop participants. This higher variability may have contributed to the failure to reach statistical significance for some SOAs in the VR group. The source of this greater variability in the VR group may stem from human factors such as variations in attention, motivation, or disruptions; or from technical factors including technical differences in headsets, software/browser differences; or speed and variability of internet connection. It is not possible to determine the specific contributions of these factors in this study. However, there is no apparent reason to believe that human factors would be more prevalent in the VR group than in the desktop group.

We suspect that the primary source of variability in reaction time is likely due to technical factors of the test environment, rather than human factors. This suspicion is drawn from the observation that the type of headset used had a significant impact on variability, with the lower-spec PlayStation VR headset exhibiting greater variability compared to the Oculus Quest 2.

Recommendations and Limitations

To reduce reaction time variability in future studies, it is recommended to take steps to minimize environmental factors. One practical approach could be to limit participation to certain headset models, even though this may reduce the number of available participants. This restriction could be particularly important when timing is a crucial aspect of the study. It should be noted that participants accessed the browser in different ways, with some reporting that they connected their headset to a computer and used Virtual Desktop to complete the experiment, while others completed the experiment without tethering the headset to a computer and using a web browser application in their headset. Unfortunately, we do not have information on the method used by all participants, as we only have anecdotal reports from some. It is recommended to consider providing explicit instructions on how to access the task in future studies to ensure consistency and reduce potential variability in time-sensitive studies.

Collecting further details of the hardware (e.g., whether using a link to PC) and software (e.g., type of web browser, or whether using a virtual desktop app) setup used would also be useful for further analysis. This can be partly automated through embedding software modules that can check and record the hardware and software configuration.

Another practical strategy to minimize the impact of variability on statistical outcomes is to carefully plan the number of trials and study participants [36]. With larger samples some variability might not impact the results. For instance, a simulation study with 158 participants showed that the addition of reaction time variability had minimal impact on the statistical outcomes [36,37]. However, it should be acknowledged that the generalizability of findings from this simulation study, which employed specific experimental parameters, to other studies, such as remote VR experiments, is uncertain.

A critical factor impacting timing is the software package and deployment method for the VR application. This study explored a web-based approach by hosting the VR experiment on the General Data Protection Regulation (GDPR) compliant research platform Pavlovia. This approach provides control over data collection and facilitates easy participant access to the study. The experiment was coded in Psychopy [29], a software package that can automatically generate a PsychoJS script, which is a JavaScript library that mirrors the PsychoPy Python library classes and functions.

The study employed a web-based approach, where participants were able to access the task through a standard URL and with the use of WebGL. This method allows for running behavioral experiments with good temporal precision. A recent timing study found inter-trial variability of under five milliseconds in almost all browsers when PsychoPy or PsychoJS was used [36]. This level of accuracy was achieved under near-optimal conditions (mid-spec computer, correctly configured graphics card, dedicated response box). While these conditions were not met in this study, we can confidently assert that the software packages used in this study were among the best available for time-critical online experiments.

It is important to note that while PsychoPy/PsychoJS and WebGL-based applications were adequate for the purpose of our study, they are not building true VR applications. However, with the increasing popularity of WebXR (https://immersive-web.github.io/, accessed on 11 January 2023), deploying VR applications on dedicated research and crowdsourced platforms may become an interesting option for testing some VR applications remotely. This is because WebXR enables developers to create VR and AR experiences that run on any device with a compatible web browser, making it more accessible and easy to use.

The study demonstrated that it is possible to recruit a significant number of participants with high-end VR headsets through the Prolific research pool. Data collection was completed within a few hours after advertising the study. Previous studies have raised concerns about the feasibility of recruiting participants with high-end VR headsets through crowdsourcing platforms [18,23]. However, this study suggests that there has been a recent increase in the number of participants with such VR headsets in the Prolific research pool, making the use of this crowdsourcing platform a viable option for VR researchers now.

Despite Prolific's reputation for good data quality compared to other platforms [11,38], it is critical to implement mitigation strategies to improve data quality [8–10]. In this study, 13 participants had to be excluded from the analysis due to non-compliance with the study's requirements (e.g., using a phone-based VR headset or a phone/tablet instead of a VR headset or desktop computer) despite clear instructions provided. Employing automated check of system configuration would help reducing the risk of participants continuing with non-compliance setup.

5. Conclusions

This study provides insight into the potential of using VR remote testing for cognitive behavioral research. Crowdsourcing services appear to be an efficient and effective way to recruit participants with high-end VR headsets for remote behavioral testing. Web-based deployment of VR applications on dedicated research platforms is an interesting option for remote VR testing.

However, it is crucial to mitigate sources of technical inter-trial variability. For example, by restricting participation to certain headsets when timing is a crucial aspect of the study. Investigating cognitive phenomena that are not as time-dependent will be better suited for remote VR testing. In any case, the potential for remote VR testing to expand the research scope and reach a wider participant population is evident from this study.

References

1. Vasser, M.; Aru, J. Guidelines for immersive virtual reality in psychological research. *Curr. Opin. Psychol.* **2020**, *36*, 71–76. [CrossRef] [PubMed]
2. Cipresso, P.; Giglioli, I.A.C.; Raya, M.A.; Riva, G. The Past, Present, and Future of Virtual and Augmented Reality Research: A Network and Cluster Analysis of the Literature. *Front. Psychol.* **2018**, *9*, 2086. [CrossRef] [PubMed]
3. Pan, X.; Hamilton, A.F.C. Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *Br. J. Psychol.* **2018**, *109*, 395–417. [CrossRef] [PubMed]
4. Tarr, M.J.; Warren, W.H. Virtual reality in behavioral neuroscience and beyond. *Nat. Neurosci.* **2002**, *5*, 1089–1092. [CrossRef]
5. Loomis, J.M.; Blascovich, J.J.; Beall, A.C. Immersive virtual environment technology as a basic research tool in psychology. *Behav. Res. Methods Instrum. Comput.* **1999**, *31*, 557–564. [CrossRef] [PubMed]
6. Follmer, D.J.; Sperling, R.A.; Suen, H.K. The role of MTurk in education research: Advantages, issues, and future directions. *Educ. Res.* **2017**, *46*, 329–334. [CrossRef]
7. Aguinis, H.; Villamor, I.; Ramani, R.S. MTurk Research: Review and Recommendations. *J. Manag.* **2020**, *47*, 823–837. [CrossRef]
8. Newman, A.; Bavik, Y.L.; Mount, M.; Shao, B. Data collection via online platforms: Challenges and recommendations for future research. *Appl. Psychol.* **2021**, *70*, 1380–1402. [CrossRef]
9. Chmielewski, M.; Kucker, S.C. An MTurk crisis? Shifts in data quality and the impact on study results. *Soc. Psychol. Personal. Sci.* **2020**, *11*, 464–473. [CrossRef]
10. Kennedy, R.; Clifford, S.; Burleigh, T.; Waggoner, P.D.; Jewell, R.; Winter, N.J. The shape of and solutions to the MTurk quality crisis. *Political Sci. Res. Methods* **2020**, *8*, 614–629. [CrossRef]
11. Peer, E.; Rothschild, D.; Gordon, A.; Evernden, Z.; Damer, E. Data quality of platforms and panels for online behavioral research. *Behav. Res. Methods* **2022**, *54*, 1643–1662. [CrossRef] [PubMed]

12. Walter, S.L.; Seibert, S.E.; Goering, D.; O'Boyle, E.H. A tale of two sample sources: Do results from online panel data and conventional data converge? In *Key Topics in Consumer Behavior*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 75–102.

13. Radiah, R.; Mäkelä, V.; Prange, S.; Rodriguez, S.D.; Piening, R.; Zhou, Y.; Köhle, K.; Pfeuffer, K.; Abdelrahman, Y.; Hoppe, M. Remote VR studies: A framework for running virtual reality studies remotely via participant-owned HMDs. *ACM Trans. Comput. Hum. Interact.* **2021**, *28*, 46. [CrossRef]

14. Ratcliffe, J.; Soave, F.; Bryan-Kinns, N.; Tokarchuk, L.; Farkhatdinov, I. Extended Reality (XR) remote research: A survey of drawbacks and opportunities. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; pp. 1–13.

15. Zhao, J.; Simpson, M.; Sajjadi, P.; Wallgrün, J.O.; Li, P.; Bagher, M.M.; Oprean, D.; Padilla, L.; Klippel, A. Crowdxr-pitfalls and potentials of experiments with remote participants. In Proceedings of the 2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Bari, Italy, 4–8 October 2021; pp. 450–459.

16. Beacco, A.; Oliva, R.; Cabreira, C.; Gallego, J.; Slater, M. Disturbance and plausibility in a virtual rock concert: A pilot study. In Proceedings of the 2021 IEEE Virtual Reality and 3D User Interfaces (VR), Lisboa, Portugal, 27 March–1 April 2021; pp. 538–545.

17. Huber, B.; Gajos, K.Z. Conducting online virtual environment experiments with uncompensated, unsupervised samples. *PLoS ONE* **2020**, *15*, e0227629. [CrossRef] [PubMed]

18. Mottelson, A.; Hornbæk, K. Virtual reality studies outside the laboratory. In Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology, Gothenburg, Sweden, 8–10 November 2017; pp. 1–10.

19. Steed, A.; Frlston, S.; Lopez, M.M.; Drummond, J.; Pan, Y.; Swapp, D. An 'In the Wild' Experiment on Presence and Embodiment using Consumer Virtual Reality Equipment. *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 1406–1414. [CrossRef] [PubMed]

20. Saffo, D.; Di Bartolomeo, S.; Yildirim, C.; Dunne, C. Remote and collaborative virtual reality experiments via social vr platforms. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; pp. 1–15.

21. Steed, A.; Izzouzi , L.; Brandstätter , K.; Friston, S.; Congdon , B.; Olkkonen, O.; Giunchi , D.; Numan , N.; Swapp , D. Ubiq-exp: A toolkit to build and run remote and distributed mixed reality experiments. *Front. Virtual Real.* **2022**, *3*, 912078. [CrossRef]

22. Mottelson, A.; Petersen, G.B.; Lilija, K.; Makransky, G. Conducting Unsupervised Virtual Reality User Studies Online. *Front. Virtual Real.* **2021**, *2*, 681482. [CrossRef]

23. Ma, X.; Cackett, M.; Park, L.; Chien, E.; Naaman, M. Web-based vr experiments powered by the crowd. In Proceedings of the 2018 World Wide Web Conference, Lyon, France, 23–27 April 2018; pp. 33–43.

24. Wiesing, M.; Steinkönig, H.; Vossel, S.; Fink, G.R.; Weidner, R. Transferring paradigms from physical to virtual reality: Can reaction time effects be replicated in a virtual setting? *PsyArXiv* **2022**. [CrossRef]

25. Klein, R.M. Inhibition of return. *Trends Cogn. Sci.* **2000**, *4*, 138–147. [CrossRef]

26. Lupiáñez, J.; Klein, R.M.; Bartolomeo, P. Inhibition of return: Twenty years after. *Cogn. Neuropsychol.* **2006**, *23*, 1003–1014. [CrossRef]

27. Posner, M.I.; Rafal, R.D.; Choate, L.S.; Vaughan, J. Inhibition of return: Neural basis and function. *Cogn. Neuropsychol.* **1985**, *2*, 211–228. [CrossRef]

28. Crump, M.J.C.; McDonnell, J.V.; Gureckis, T.M. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE* **2013**, *8*, e57410. [CrossRef] [PubMed]

29. Peirce, J.; Gray, J.R.; Simpson, S.; MacAskill, M.; Höchenberger, R.; Sogo, H.; Kastman, E.; Lindeløv, J.K. PsychoPy2: Experiments in behavior made easy. *Behav. Res. Methods* **2019**, *51*, 195–203. [CrossRef] [PubMed]

30. The Jamovi Project Jamovi (Version 2.2.5). 2022. Available online: https://www.jamovi.org (accessed on 5 January 2022).

31. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2022; Available online: https://www.R-project.org/ (accessed on 10 November 2022).

32. Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L.D.A.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J. Welcome to the Tidyverse. *J. Open Source Softw.* **2019**, *4*, 1686. [CrossRef]

33. Grange, J. Trimr: An Implementation of Common Response Time Trimming Methods. 2018. Available online: https://CRAN.R-project.org/package=trimr (accessed on 22 December 2022).

34. Ben-Shachar, M.S.; Lüdecke, D.; Makowski, D. Effectsize: Estimation of Effect Size Indices and Standardized Parameters. *J. Open Source Softw.* **2020**, *5*, 2815. [CrossRef]

35. Van Selst, M.; Jolicoeur, P. A solution to the effect of sample size on outlier elimination. *Q. J. Exp. Psychol. Sect. A* **1994**, *47*, 631–650. [CrossRef]

36. Bridges, D.; Pitiot, A.; MacAskill, M.R.; Peirce, J.W. The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ* **2020**, *8*, e9414. [CrossRef] [PubMed]

37. Brand, A.; Bradley, M.T. Assessing the Effects of Technical Variance on the Statistical Outcomes of Web Experiments Measuring Response Times. *Soc. Sci. Comput. Rev.* **2012**, *30*, 350–357. [CrossRef]

38. Peer, E.; Brandimarte, L.; Samat, S.; Acquisti, A. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *J. Exp. Soc. Psychol.* **2017**, *70*, 153–163. [CrossRef]