frontiers | Frontiers in Virtual Reality

TYPE Perspective
PUBLISHED 12 April 2023
DOI 10.3389/frvir.2023.1069423

OPEN ACCESS

EDITED BY
Jiayan Zhao,
Wageningen University and Research, Netherlands

REVIEWED BY
Mahda M. Bagher,
The Pennsylvania State University (PSU), United States
Jack Shen-Kuen Chang,
National Cheng Kung University, Taiwan

*CORRESPONDENCE
Jascha Grübel,
jgruebel@ethz.ch

SPECIALTY SECTION
This article was submitted to Virtual Reality and Human Behaviour, a section of the journal Frontiers in Virtual Reality

RECEIVED 13 October 2022
ACCEPTED 27 March 2023
PUBLISHED 12 April 2023

CITATION
Grübel J (2023), The design, experiment, analyse, and reproduce principle for experimentation in virtual reality. *Front. Virtual Real.* 4:1069423. doi: 10.3389/frvir.2023.1069423

The design, experiment, analyse, and reproduce principle for experimentation in virtual reality

Jascha Grübel[1,2,3,4,5]*

[1]Chair of Cognitive Science, Department of Humanities, Social and Political Sciences, ETH Zürich, Zurich, Switzerland, [2]Game Technology Center, Department of Computer Science, ETH Zürich, Zurich, Switzerland, [3]Visual Computing Group, School of Engineering and Applied Sciences, Harvard University, Zurich, Switzerland, [4]Center for Sustainable Future Mobility, Department of Civil, Environmental and Geomatic Engineering, ETH Zürich, Zurich, Switzerland, [5]Chair of Geoinformation Engineering, Department of Civil, Environmental and Geomatic Engineering, ETH Zürich, Zurich, Switzerland

Conducting experiments in virtual reality (VR) requires a complex setup of hardware, software, experiment design and implementation, and data collection which is supported by frameworks that provide pre-determined features for scientists to implement their experiment in VR. These VR frameworks have proliferated exponentially since the start of the millennia, and unfortunately, they both only differ slightly from one another and often miss one or more of the key features required by the researcher. Therefore, it has become less clear to researchers which framework to choose for what task and to what benefit. I introduce the design, experiment, analyse, and reproduce (DEAR) principle to develop a new perspective on VR frameworks through a holistic approach to experimentation (i.e., the process of conducting an experiment). The DEAR principle lays out the core components that future frameworks should entail. Most previous VR frameworks have focussed on the design phase and sometimes on the experiment phase to help researchers create and conduct experiments. However, being able to create an experiment with a framework is not sufficient for wide adoption. Ultimately, I argue that it is important to take reproducibility seriously to overcome the limitations of current frameworks. Once experiments are fully reproducible through automation, the adaptation of new experiments becomes easier. Hopefully, researchers can find ways to converge in the use of frameworks or else frameworks may become a hindrance instead of a help.

KEYWORDS

virtual reality, digital twin, framework, reproducibility, experiments, experimentation

1 Introduction

Virtual reality (VR) has become one of the most promising venues to conduct behavioural research because it allows us to control stimuli, the environment, the modes of perception, and interaction, to systematically record a variety of responses (e.g., physiological measurements and eye-tracking), and more to design and implement experiments. Over the last decades, experiments in VR have become a main staple for any serious research topic on human behavioural analysis (HBA) ranging from psychology (Gaggioli, 2001) to architecture (Whyte, 2003) and economics (Innocenti, 2017). At the same time, the experimentation, or the process of how we conduct experiments, has received less scrutiny. Using VR has become easier as technology progressed from the clunky Sword of Damocles (Sutherland, 1968) to the Oculus, the VIVE, and more modern head-mounted

displays (HMDs), with improvements in computation power, size, knowledge in optics, and interfaces. Nonetheless, using VR in experiments has one drawback that has haunted behavioural sciences in general—reproducibility (Collaboration, 2015). Early work (Dalton, 2003) only presented a high-level summary of the methodology that would be hard to accurately reproduce, thus leaving us with insufficient traces of the experimentation that happened in order to perform the experiment.

The practical questions of experimentation are compounded with the meta-level trade-offs between the cost of conducting an experiment and the control that can be achieved within the experiment. Without frameworks, each new experiment leads to a full development cycle that is both time-consuming and expensive. However, generalising the steps of experimentation effectively into a framework that supports a large array of experiments is costly.

Ultimately, this gave rise to frameworks to support VR experiments across the board. Frameworks contrast with templates and platforms as an intermediate-level solution to complexity by providing predefined features for some tasks but still requiring external software and hardware to run. On the one hand, templates usually provide a pre-written starting point for a program, class, or function that pre-defines common features to solve similar tasks (Vandevoorde and Josuttis, 2002). The application level in writing a code is usually too low for most experimental researchers as they are not computer scientists and require a more accessible programming platform to design their experiment. Some frameworks arguably use templating for creating tasks, but overall, they also try to manage participants, data, and more which goes beyond the scope of a usual template. On the other hand, platforms provide an execution environment for other software (Evans et al., 2008). This usually consists of high-level definitions of what kind of activities can be performed like operating systems or game platforms like Unity. The line between frameworks and platforms is somewhat blurry as a game platform relying on an operating system could still be considered a framework as it does not run independently. However, the general consensus is to accept platforms that rely on other platforms to work. Platforms have a more general goal than frameworks and usually provide generalised services beyond the scope of a single type of application. Modern VR frameworks usually rely on Unity or Unreal game platforms as a basis and are compiled within the respective platforms and not in a stand-alone way (Aguilar et al., 2022).

Today, the number of frameworks is booming (Aguilar et al., 2022), and it has become more difficult than at any point in time before to decide which framework to use. VR frameworks are currently cannibalising each other as they are only marginally different but still crucially different for the users across different disciplines and sub-disciplines. Therefore, in most cases, it is more useful for researchers to develop a new framework than reuse an older one. In this perspective paper, I discuss why this is the case and how it can be addressed through the development of the design, experiment, analyse, and reproduce (DEAR) principle.

2 A short history of VR frameworks for behavioural experiments

The multiplication of frameworks can be historically grouped into three generations (Aguilar et al., 2022). In the first generation in the early 2000s, the framework mostly took care of the hardware (Tramberend, 1999; Allen et al., 2000; Ayaz et al., 2008; Annett and Bischof, 2009; Mossel et al., 2012). It enabled the addition of human interface devices (HIDs) and organised the graphics regarding some display platforms such as head-mounted displays (HMDs; Sutherland, 1968) and cave automatic virtual environment (CAVE; Cruz-Neira et al., 1993). The onerousness of creating an experiment in the virtual environment remained with the researcher. A great discrepancy caused by this is that ultimately creating experiments was more tedious than necessary. The users were not computer scientists but behavioural scientists who learnt about VR in a self-taught manner. A trained expert could often resolve the tasks within days or weeks, whereas a novice user could take years. In a personal anonymous conversation, a behavioural researcher informed me that re-implementing their 4 years' of work for a PhD thesis during their postdoctoral research, with the help of a trained computer scientist, took about 1 month. This difference in time is not to be attributed to the researcher's capabilities but should serve as a stark reminder of the limitations that accompanies experimenting in VR.

This plight led to the development of the second-generation of VR experiments that produced higher-level abstractions for experiment components from 2010 onwards (Grübel et al., 2016; Moulec et al., 2017; Schneider et al., 2018; Zhao et al., 2018; Brookes et al., 2019; Watson et al., 2019; Alsbury-Nealy et al., 2020; Bebko and Troje, 2020; Starrett et al., 2020; Wang et al., 2020; Ugwitz et al., 2021; Schuetz et al., 2022). These frameworks focus on the comfort of the researcher by providing templates for certain tasks and infrastructure to design typical sequences of experimental tasks such as repetitions and vignettes. They also provided prepared experimental packages for important tasks such as participant training in the controls which is crucial to differentiate between the task performance and the HID usability performance (Grübel et al., 2017). Some frameworks extended their scope beyond design helpers and offered analysis tools (Grübel et al., 2016; Starrett et al., 2020). In terms of data management, there were two broadly different approaches, hiding data complexity from the user (e.g., Grübel et al., 2016; Brookes et al., 2019; Wang et al., 2020) and providing a very flat and simple data hierarchy (Zhao et al., 2018; Alsbury-Nealy et al., 2020; Starrett et al., 2020). However, the specialisation of frameworks for very particular tasks despite supporting generalised experiment tools increased the number of frameworks over the last years exponentially.

This cannibalisation can be traced back to two major problems. First, despite the simplifications compared to first-generation frameworks, expert programmers are still required to create new experiments. Expert programmers often do not like to learn a new technology stack, especially if it is still in the early stage and has numerous gaps and bugs in the implementation. Ultimately, these experts (the author included) often conclude that their research partners are better served if they (re-) develop a framework. The functionality of VR frameworks such as EVE (Grübel et al., 2016), Landmarks (Starrett et al., 2020), and NVR-DeSciL (Zhao et al., 2018) is remarkably similar. Cooperative efforts to create a single framework could have resulted in a more stable base-framework and less competition between frameworks. Nonetheless, there are subtle differences documented by Starrett et al. (2020) that ultimately lead to the decision of starting from scratch rather than reusing the existing work. Similarly, VREVAL (Schneider et al., 2018) and vexptoolbox (Schuetz et al., 2022) preferred a different technology platform. In contrast, USE (Watson et al., 2019), bmlTUX (Bebko and Troje, 2020), UXF (Brookes et al., 2019), and VO (Howie and Gilardi, 2020) required the management of high-frequency data and opted for restarting from scratch rather than figuring out how this can be

integrated in the existing work. Lastly, VR-Rides (Wang et al., 2020) came from a different discipline and were not aware of work in other fields.

Another impediment to the actual use of these second-generation frameworks is end-user usability. To apply any of these frameworks and especially to change them often requires advanced knowledge in computer science, knowledge of the game engines, and programming in general. Therefore, the second-generation frameworks do not achieve the wide-range application which their potential for reuse could imply. A notable exception is the niche framework virtual SILCton (Schinazi et al., 2013; Weisberg et al., 2022) which continues to be used in a variety of spatial cognition experiments over the last decade. SILCton offers very few options for configuration and is specialised for a specific set of spatial cognition tasks. However, SILCton provides a simple browser-based user interface (UI) that can be used with no expert knowledge and allows the reuse, re-configuration, and export of data by novices and non-experts. Despite the tremendous progress in usability, most second-generation frameworks are still too difficult to become the base for behavioural research in VR that could be reused across disciplines.

Second, the setup of a framework often requires many steps that are seldom easy to perform and often impact the reproducibility of a system and ultimately of experiments conducted with a platform. Modern content delivery systems have made the use easier but still require interactions with complex software platforms such as the Unity game engine (Starrett et al., 2020). Users need to install databases, drivers, game engines, and statistical software. Users need to learn interfaces for different programs and understand engineering intentions to modify and adjust frameworks according to their needs. The sheer number of tasks can already deter researchers from trying to develop and usually leads to the addition of an expert programmer bringing us back to the first problem.

A notable exception in second-generation frameworks is the PsychoPy package 3 developed for non-VR stimuli in neuroscience (Peirce, 2009) which has gained tremendous popularity even beyond neuroscience. Neuroscientists often use python-based packages to interact with their measurement devices. PsychoPy is a simplifying agent in the complex task and therefore is quickly adopted. However, it is not applicable to VR setups. The problem for VR frameworks is the large number of options for each component from data collection to visualisation hardware and software. Only recently major players such as the Unity game engine have unified access to hardware through standardised interfaces such as the Unity XR manager. However, the quick development rate of hardware often means that standard interfaces can only use a partial instruction set of HIDs, ranging from HMDs to user input based on gaze, physiological reactivity, or hand motion. For instance, newer Oculus products support hand-tracking, and the Omnicept hardware supports physiological data collection, yet neither feature is supported in standard interfaces of Unity, but each hardware developer provides unity-based libraries for interaction. Producing a holistic framework based on the wide variety of hardware-specific libraries is not fruitful. Developers are left to either only maintain one piece of hardware that may become obsolete or only support general features that are not as state of the art as the specialised libraries. These choices in second-generation frameworks are often made intransparently and make it impossible to understand their requirements, which ultimately limits the take-up by users.

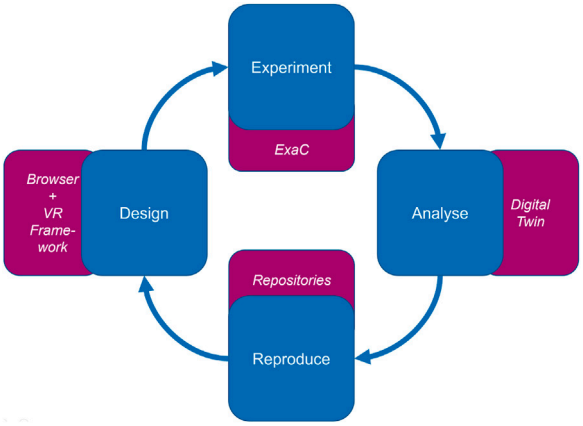

FIGURE 1
Overview of the DEAR principle. The design phase happens in a browser with an existing VR framework as a backup. The experiment phase uses an implementation of the Experiment as Code (ExaC) paradigm to provision, deploy, and run the experiment and enable the collection of data. The analyse phase uses a digital twin to represent the experiment and enable real-time management, as well as statistical evaluation of results. Lastly, the reproduce phase ensures that the experiment setup, the collected data, and the proposed analysis are available in repositories. These repositories can be used as an input to the next design phase.

Third, not every step of an experiment is part of the frameworks and can easily be missed without an appropriate protocol that documents what needs to happen when (Weibel et al., 2018). Ultimately, second-generation frameworks mostly focus on facilitating the design step and may provide some guidance on data collection and analysis. These three reasons together have condemned most VR frameworks to languish in a very niche position and not to be picked up across disciplines despite the expected gain of reusing experimental software.

The third-generation of frameworks is only emerging now and is shifting the focus away from the design to reproduction (Aguilar et al., 2022; Colombo et al., 2022; Colombo et al., 2023) under the guidance of preproducibility (Stark, 2018). Reproducibility has become the Achilles' heel of modern science, with several experiments not being able to continuously produce the same results (Ioannidis et al., 2015; Camerer et al., 2018) and consequently putting the claims made based on the experiments performed on shaky grounds (Collaboration, 2015). The practice is not yet common in VR frameworks but has found some application in more general stimuli frameworks (Cherrueau et al., 2018; Almaatouq et al., 2020). What the second generation was still missing is the ability to quickly reproduce original experiments and possibly modify them to gain new insights and helping with reproducibility. The third generation helps provision the hardware for an experiment and deploy all the software, including the required background software for data collection, the experiment application, the data storage, the data management, and the data analysis through the Experiments as Code (ExaC) paradigm (Aguilar et al., 2022). Protocols about the experimental process and other documentations are explicitly part of the framework. Furthermore, experiment management and data analysis are included in the software stack of an experiment in order to address the reproducibility crisis on the statistics level (Gosselin, 2020). This view helps better understand the data collection as well as

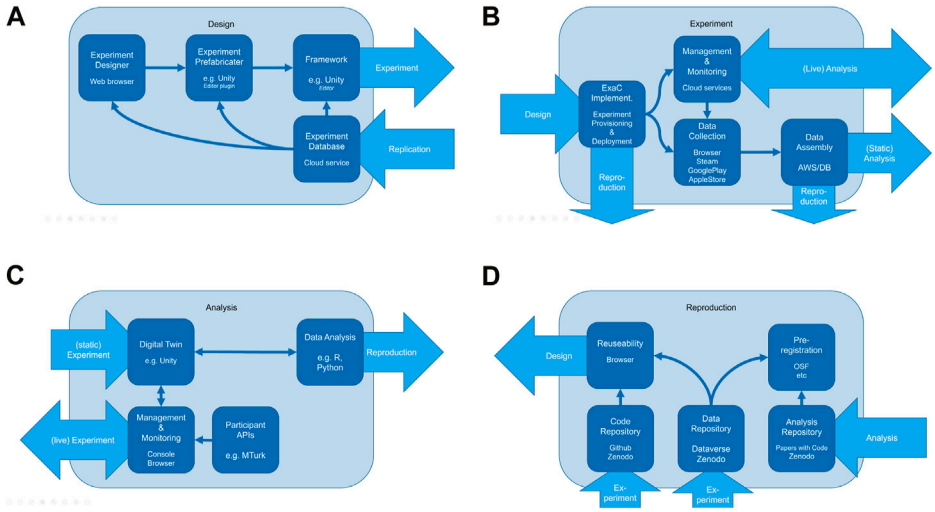

FIGURE 2
DEAR principle has four distinct faces that work in a circular pattern. **(A)** Design phase provides users with a simplified interface to set up standard variants of known experiments from a database. It prepares the experiment and offers the opportunity to advanced users to adjust the experiment. **(B)** Experiment phase uses the Experiments as Code paradigm to automatically deploy the experiment to a suitable platform, engage participants, manage the experiment, and collect data. **(C)** Analyse phase of an experiment employs a digital twin of the experiment to monitor the progress of the experiment, recruit participants into the experiment, and statistically evaluate the data. **(D)** Reproduce phase gathers the experiment code, the collected data, and the analysis scripts into long-term repositories. These repositories are in turn indexed in the experiment database for the design phase.

the reproduction of experimental results. Third-generation frameworks can also be considered meta-frameworks as they provide ample space for the implementation of the actual experiment with any second-generation framework and focus on embedding this experiment in best practices for Open Research Data (ORD; Burgelman et al., 2019). Ultimately, the third generation of experimental frameworks for VR is only getting started as the required technology becomes easily available, and it remains to be seen whether it will be befallen by the same ills as the second generation.

3 Design, experiment, analyse, and reproduce principle

The DEAR principle can be understood as a guideline to properly implement third-generation frameworks for experiments in VR, as shown in Figure 1. Many previous attempts at VR experimentation frameworks have partially implemented the DEAR principle. Most efforts can be categorised into the design phase and the experiment phase with some ventures into the reproduce phase. Crucially, no framework has covered all aspects and ultimately neither does support circular reproducibility. I define circular reproducibility as the ability to reuse previous experiments, their design, and analysis as an input to a new generation of experiments. Circular reproducibility entails the preproducibility of experiments as all information to instantiate an experiment or a new variant thereof must be available at any time.

The principle is grouped into four distinct phases. Each phase covers different technology platforms that are required to implement a third-generation framework, as shown in Figure 2. In the following, I will discuss how an implementation of the DEAR paradigm could work. Notably, what I propose here is a minimal working example of which technologies must be joined to satisfy the DEAR principle and enable circular reproducibility. However, already partial implementation of the DEAR principle will increase reproducibility and hopefully can lead the way to better science. Therefore, the focus here is on the minimum requirements for usability as well as some advanced suggestions on the implementation.

3.1 Design phase

Conceptually, the design phase enables researchers to specify their experiment, as shown in Figure 2A. An important aspect is usability, and it is a key requirement to lower the threshold for interaction. Researchers are occupied with questions on the order and elements of the protocol, sufficient statistical power, identifiability of the main effects, and soundness of the design. In this context, it would be preferable not to have to limit the design based on the technical capability of the researcher. Therefore, it is important that in the first step, an experiment template can be selected as simply as possible, without having to learn complex software like Unity. A website to browse experiments like goods in a store helps establish first contact. After an experiment has been chosen, a good template will provide the researchers to manipulate key quantities of interest in a simple form. Once the researchers have configured their experiment, they can create the required software at the push of a button. Unless the researchers want to change the template itself, it is not required to interact with the underlying source code.

Behind the front-end for the user, a database lists all possible experiments and their framework called the ExaC (Aguilar et al., 2022). ExaC are stored in public online repositories that are used to instantiate a configured template, with the experiment prefabricator for the required framework. The template configuration is also stored in the database for later reuse. On a server, the ExaC is cloned and loaded by

the target platform on which the framework was developed, such as Unity. It would be possible to use any second-generation framework for an experiment and deploy accordingly with a correctly configured ExaC. Furthermore, any third-generation framework is theoretically agnostic to the code deployed and could also be used for classic stimuli experiments such as from PsychoPy or surveys based on Qualtrics or other survey providers. The settings the researchers chose on the website are applied on the instantiated template. Here, experienced users could step into the generation process and modify the template according to their needs.

3.2 Experiment phase

The experiment phase focusses on conducting the experiment and collecting data, as shown in Figure 2B. This phase should look mostly automated from the perspective of the researcher. The ExaC implementation should provision for and deploy the experiment (Aguilar et al., 2022). In case of online experiments, this would mean automatic recruitment of participants through online platforms, the distribution of the program and other required software to the participants, checking hardware requirements, opening databases for data assembly, and setting up management and monitoring software. In case of offline experiments, some variations would be required. Participants may automatically be recruited, but researchers will have to manage the introduction to the hardware and software.

Nonetheless, the automatisation can take care of many otherwise tedious steps. Once participants are ready, the data are collected during an experiment run. Depending on the deployment mode, it may use renowned platforms like Steam, Google Play, the Apple App Store, a browser, or a local machine. During data collection, it is possible for researchers to manage and monitor experiments. So far, management and monitoring has been an afterthought for most experiment frameworks. However, it is possible to actually intervene during the process and trigger events or take other actions.

3.3 Analyse phase

The analyse phase provides researchers with the necessary support to manage and monitor ongoing experiments and conduct statistical analysis, as shown in Figure 2C. The two most important aspects are a digital twin of the experiment (Grübel, 2023) and a management and monitoring console (Aguilar et al., 2022). Experiments as digital twins offer a new perspective on taking the experiment process as part of the data analysis. A digital twin is a digital representation of a physical twin that allows deeper insights into processes (Grieves and Vickers, 2017; Grübel et al., 2022). The idea behind applying digital twins to experiments is to also capture the process of experimentation that is often underreported in scientific applications (Grübel, 2023). Often, results are only reported on variables of interest, which were collected during an experiment. However, these variables may not be independent of how the experiment was conducted. Timing between tasks, order of tasks, environmental properties of the experiment, and many more may have had an impact that researchers did not account for. In previous experiments, getting these impacts in retrospect is impossible without conducting the experiment again, which may run into issues of behavioural drifts (Reardon, 2016) such that the experiment was conducted properly, but the sampled population has changed unbeknown to the researcher. The different results could accidentally be attributed to newly observed variables. Experiments such as digital twins resolve this issue by capturing the experiment process as well and consequently allowing for additional analysis that the original research plan did not entail.

In the context of VR, the digital twin for the analysis has to replicate both the physical twin (i.e., the participants) and so-called virtual twin (i.e., the VR environment; Grübel, 2023). For implementation, the digital twin requires a replica of the VR environment used in the experiment that allows for the overlay of experiment information for the researcher. Information can range from overview information of the participants' current performance to preliminary statistical analysis. This information can be displayed in real-time or *post hoc* in a replay mode. The environment can also be used to gather participants from crowd-sourcing platforms such as MTurk or similar. The management and monitoring tools (MMTs) may rely on digital twin but implements a simpler front-end that allows researchers to quickly adjust the experiment without having to enter the digital twin. The data analysis can be both run from within the digital twin or within the typical statistical environments such as Python and R. The data analysis can either be prepared in advance with templates (from previous experiments), pilots, and pre-registered analysis designs (Gonzales and Cunningham, 2015) or be assembled *in situ* in the digital twin, leaving a trail of applied analyses for the posterior world.

3.4 Reproduce phase

The reproduce phase takes all steps a researcher took and forms them into persistent records of the experiment, as shown in Figure 2D. There are two main goals here: first, producing records of all important components to comply with preproducibility (Stark, 2018) and second enabling automated preregistration of studies based on data collection models and chosen analysis. The choice of repository should be free and not be predetermined by the framework, although the framework will have to implement the repository API for automated processing. Here, it is important to account for delayed publications to enable researchers to publish their results before going public. Ultimately, these repositories can be registered in the experiment database for the design phase and be used as an input for the circular reproducibility.

4 Discussion

Experiment frameworks in VR have proliferated exponentially in the 2020s, and at the moment it looks like this trend is not abating. This trend is cannibalising the frameworks as the potential user base is split across frameworks, and lots of parallel implementation is happening that ultimately reduces the effectiveness of developing frameworks in the first place. To overcome this dilemma, VR frameworks need to become compatible in a meta-framework.

At the same time, another hindrance is reducing the uptake of experiment frameworks. While each framework tremendously reduces the workload for its application scenario, the low extendability of the frameworks or the high costs of extending a framework often stop other researchers with different application scenarios from taking up the framework. The cost around the take-up is evaluated as being too

high. Ultimately, this evaluation leads to yet another framework, resulting in a multiplication of standards. While there is no direct solution to this dilemma, it is possible to facilitate all other steps of the experiment in a meta-framework.

Lastly, as the number of frameworks increases, it becomes harder to understand the advantages and disadvantages of each framework. Most frameworks are developed in scientific niches and are even competing within the same disciplines for the researchers' attention. Some frameworks might have been useful but cannot make their way to the optimal user because the link is not available. A meta-framework for selecting the appropriate framework for a specific experiment could resolve the issue by providing experiments implemented across all platforms and offering a single go-to place.

Ultimately, reproducibility can only bring us so far when assessing the quality of experiments. To go beyond mere repetition, we need the infrastructure to produce intelligent variations of experiments. Only then we can validate theoretical concepts under varying dependent and independent variables to triangulate real effects (Munafò and Smith, 2018). Furthermore, interactive experiments where the researcher uses the management and monitoring tools to interact with the participants were simply not available before and offer new opportunities to define variables. A meta-framework that protocols all steps of an experiment including the process itself allows us to understand and develop variations of experiments that are comparable.

Currently, there is no full implementation of the DEAR paradigm. However, the Spatial Performance Assessment for Cognitive Evaluation (SPACE) framework (Colombo et al., 2022; 2023) covers important aspects of all four phases. A user-friendly experiment designer is available that instantiates experiments from a limited set of tasks related to spatial cognition. The app is deployed on tablets and reads in the configuration remotely. There is no live monitoring, but the resulting data are gathered and analysed with a standardised evaluation script. The source code will be made publicly available after completion. In a sense, this framework offers a partial DEAR platform to run spatial cognition experiments related to path integration.

Attaining experiment frameworks that satisfy the DEAR principle for general purpose experiments is still some time off. Meanwhile, it would be applaudable to contribute to existing works rather than re-inventing experiment frameworks even if that happens at the cost of personal fame. For instance, ExaC (Aguilar et al., 2022) provides a framework which can exchange the experiment core and provides provisioning and deployment. It would be great if other frameworks were adapted to this core to enable many more DEAR features to end users. Unanswered is the question of a front-end market place and the direct pipelines from the experiment to archiving repository, which might also be solvable in small packages that can be added to any framework similar to ExaC. Ultimately, it probably requires a larger ORD effort (Burgelman et al., 2019) from the whole community of experimenters to address these short-comings.

5 Conclusion

The DEAR principle offers guidelines for a meta-framework that helps us understand VR frameworks in a larger context. Each of the phases in DEAR addresses the dilemma in the current world of VR experiments that needs addressing. The design phase allows for lower barriers to entry and could enable researchers to pick the right line of experiments from where to expand the body of knowledge. The experiment phase implements the Experiments as Code paradigm and thereby ensures that the reproducibility of an experiment is enabled to the highest possible degree. At the same time, tedious steps can be automated that previously require many manual steps. The analyse phase offers basic analyses for experiments and allows others to validate the statistical approach while also giving the researcher extended control over the flow of an experiment. Lastly, the reproduce phase automates the required steps to feed the experiment back into the loop of circular reproducibility. If correctly implemented, the DEAR principle can become the basis for high quality research that opens the doors for what comes after it. Newer frameworks would perform best to implement DEAR as far as possible and to become interoperable with other phases of other frameworks to advance science beyond the myopic scope of a single paper or experiment.

Author contributions

JG wrote and edited the whole article.

Funding

This research was funded by ETH Zurich (grant number ETH-15 16-2), and JG was supported by an ETH Zurich Doc.Mobility Fellowship. Open access funding was provided by ETH Zurich.

Acknowledgments

JG would like to thank every colleague at the Chair of Cognitive Science at ETH Zürich with whom JG implemented VR experiments and whose interactions were fundamental in gathering the ideas behind this article.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Aguilar, L., Gath-Morad, M., Grübel, J., Ermatinger, J., Zhao, H., Wehrli, S., et al. (2022). Experiments as code: A concept for reproducible, auditable, debuggable, reusable, & scalable experiments. arXiv preprint arXiv:2202.12050

Allen, B., Christopher, J., Patrick, H., Kevin, M., Baker, A., and Carolina, C. (2000). "Vr juggler: A virtual platform for virtual reality application development," in Proceedings of IEEE, Japan, 13-17 March 2001 (IEEE).

Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., and Whiting, M. E. (2020). Empirica: A virtual lab for high-throughput macro-level experiments. arXiv preprint arXiv:2006.11398

Alsbury-Nealy, K., Wang, H., Howarth, C., Gordienko, A., Schlichting, M., and Duncan, K. (2020). *Openmaze: An open-source toolbox for creating virtual environment experiments*. Available at: https://psyarxiv.com/bsj47/.

Annett, M., and Bischof, W. F. (2009). Vr for everybody: The snap framework. *Proc. IEEE VR 2009 Workshop Softw. Eng. Archit. Realt. Interact. Syst.* 144, 131. doi:10.3233/978-1-60750-017-9-61

Ayaz, H., Allen, S. L., Platek, S. M., and Onaral, B. (2008). Maze suite 1.0: A complete set of tools to prepare, present, and analyze navigational and spatial cognitive neuroscience experiments. *Behav. Res. methods* 40, 353–359. doi:10.3758/brm.40.1.353

Bebko, A. O., and Troje, N. F. (2020). bmltux: Design and control of experiments in virtual reality and beyond. *i-Perception* 11, 204166952093840. doi:10.1177/2041669520938400

Brookes, J., Warburton, M., Alghadier, M., Mon-Williams, M., and Mushtaq, F. (2019). Studying human behavior with virtual reality: The unity experiment framework. *Behav. Res. methods* 1–9, 455–463. doi:10.3758/s13428-019-01242-0

Burgelman, J.-C., Pascu, C., Szkuta, K., Von Schomberg, R., Karalopoulos, A., Repanas, K., et al. (2019). Open science, open data, and open scholarship: European policies to make science fit for the twenty-first century. *Front. Big Data* 2, 43. doi:10.3389/fdata.2019.00043

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., et al. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nat. Hum. Behav.* 2, 637–644. doi:10.1038/s41562-018-0399-z

Cherrueau, R. A., Simonin, M., and Van Kempen, A. (2018). Enosstack: A lamp-like stack for the experimenter. In IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops 15-19 April 2018, USA, (INFOCOM WKSHPS) (IEEE, 336

Collaboration (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* 349, aac4716. doi:10.1126/science.aac4716

Colombo, G., Grübel, J., Hölscher, C., and Schinazi, V. R. (2023). The spatial performance assessment for cognitive evaluation (space): A novel game for the early detection of cognitive impairment. *CHI 2023 Stud. Game Compet.* 8, 1.

Colombo, G., Grübel, J., Minta, K., Wiener, J. M., Avraamides, M., Hölscher, C., et al. (2022). Spatial performance assessment for cognitive evaluation (space): A novel tablet-based tool to detect cognitive impairment. In. *4th Interdiscip. Navig. Symp. (iNAV 2022. Virtual Meet.* 2022, 14. doi:10.3929/ethz-b-000594027

Cruz-Neira, C., Sandin, D. J., and DeFanti, T. A. (1993). "Surround-screen projection-based virtual reality: The design and implementation of the cave," in Proceedings of the 20th annual conference on Computer graphics and interactive techniques, Germany, September 1993 (IEEE), 135–142.

Dalton, R. C. (2003). The secret is to follow your nose: Route path selection and angularity. *Environ. Behav.* 35, 107–131. doi:10.1177/0013916502238867

Evans, D. S., Hagiu, A., and Schmalensee, R. (2008). *Invisible engines: How software platforms drive innovation and transform industries*. Cambridge: The MIT Press.

Gaggioli, A. (2001). Using virtual reality in experimental psychology. *Towards Cyberpsychology*. Available at: https://books.google.ch/books?hl=enandlr=andid=c9UnQ0OKL_wCandoi=fndandpg=PA157anddq=Using+virtual+reality+in+experimental+psychology+gaggioliandots=MKgMIKkalqandsig=ovfb68L7HNf3uiYXk424cyM5TUY#v=onepageandq=Using%20virtual%20reality%20in%20experimental%20psychology%20gaggioliandf=false, 157–174.

Gonzales, J. E., and Cunningham, C. A. (2015). The promise of pre-registration in psychological research. Available at: https://www.apa.org/science/about/psa/2015/08/pre-Registration and archived on OSF: https://osf.io/k2q4d/. *Psychol. Sci. Agenda* 29, 2014. –2017.

Gosselin, R.-D. (2020). Statistical analysis must improve to address the reproducibility crisis: The access to transparent statistics (acts) call to action. *BioEssays* 42, 1900189. doi:10.1002/bies.201900189

Grieves, M., and Vickers, J. (2017). "Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems," in *Transdisciplinary perspectives on complex systems* (Germany: Springer), 85–113.

Grübel, J. (2023). *Handbook of digital Twins*. Florida: CRC Press.

Grübel, J., Thrash, T., Aguilar, L., Gath-Morad, M., Chatain, J., Sumner, R. W., et al. (2022). The hitchhiker's guide to fused twins: A review of access to digital twins *in situ* in smart cities. *Remote Sens.* 14, 3095. doi:10.3390/rs14133095

Grübel, J., Thrash, T., Hölscher, C., and Schinazi, V. R. (2017). Evaluation of a conceptual framework for predicting navigation performance in virtual reality. *PloS one* 12, e0184682. doi:10.1371/journal.pone.0184682

Grübel, J., Weibel, R., Jiang, M. H., Hölscher, C., Hackman, D. A., and Schinazi, V. R. (2016). "Eve: A framework for experiments in virtual environments," in *Spatial cognition X* (Germany: Springer), 159–176.

Howie, S., and Gilardi, M. (2020). Virtual observations: A software tool for contextual observation and assessment of user's actions in virtual reality. *Virtual Real.* 1–14, 447–460. doi:10.1007/s10055-020-00463-5

Innocenti, A. (2017). Virtual reality experiments in economics. *J. Behav. Exp. Econ.* 69, 71–77. doi:10.1016/j.socec.2017.06.001

Ioannidis, J. P., Fanelli, D., Dunne, D. D., and Goodman, S. N. (2015). Meta-research: Evaluation and improvement of research methods and practices. *PLoS Biol.* 13, e1002264. doi:10.1371/journal.pbio.1002264

Mossel, A., Schönauer, C., Gerstweiler, G., and Kaufmann, H. (2012). Artifice-augmented reality framework for distributed collaboration. *Int. J. Virtual Real.* 11, 1–7. doi:10.20870/ijvr.2012.11.3.2845

Moulec, G. L., Argelaguet, F., Gouranton, V., Blouin, A., and Arnaldi, B. (2017). Agent: Automatic generation of experimental protocol runtime. In Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology. November 2017, New York City, IEEE, 1

Munafò, M. R., and Davey Smith, G. (2018). Robust research needs many lines of evidence. *Nature* 553(7689), 399-401

Peirce, J. W. (2009). Generating stimuli for neuroscience using psychopy. *Front. neuroinformatics* 2, 10. doi:10.3389/neuro.11.010.2008

Reardon, S. (2016). A mouse's house may ruin experiments. *Nat. News* 530, 264. doi:10.1038/nature.2016.19335

Schinazi, V. R., Nardi, D., Newcombe, N. S., Shipley, T. F., and Epstein, R. A. (2013). Hippocampal size predicts rapid learning of a cognitive map in humans. *Hippocampus* 23, 515–528. doi:10.1002/hipo.22111

Schneider, S., Kuliga, S., Weiser, R., Kammler, O., and Fuchkina, E. (2018). "Vreval-a bim-based framework for user-centered evaluation of complex buildings in virtual environments," in *Proceedings of the 36th eCAADe conference* (CUMINCAD), 10.

Schuetz, I., Karimpur, H., and Fiehler, K. (2022). vexptoolbox: A software toolbox for human behavior studies using the vizard virtual reality platform. *Behav. Res. Methods* 55, 570–582. doi:10.3758/s13428-022-01831-6

Stark, P. B. (2018). Before reproducibility must come preproducibility. *Nature* 557, 613–614. doi:10.1038/d41586-018-05256-0

Starrett, M. J., McAvan, A. S., Huffman, D. J., Stokes, J. D., Kyle, C. T., Smuda, D. N., et al. (2020). Landmarks: A solution for spatial navigation and memory experiments in virtual reality. *Behav. Res. Methods* 1–14, 1046–1059. doi:10.3758/s13428-020-01481-6

Sutherland, I. E. (1968). "A head-mounted three dimensional display," in Proceedings of the December 9-11, 1968, fall joint computer conference, Pushkina, December 9 (AFIPS).

Tramberend, H. (1999). "Avocado: A distributed virtual reality framework," in Proceedings IEEE Virtual Reality (Cat. No. 99CB36316) (IEEE, USA, 13-17 March 1999 (IEEE), 14–21.

Ugwitz, P., Šašinková, A., Šašinka, Č., Stachoň, Z., and Juřík, V. (2021). Toggle toolkit: A tool for conducting experiments in unity virtual environments. *Behav. Res. methods* 53, 1581–1591. doi:10.3758/s13428-020-01510-4

Vandevoorde, D., and Josuttis, N. M. (2002). *C++ templates: The complete guide, portable documents*. United States: Addison-Wesley Professional.

Wang, Y., Ijaz, K., Yuan, D., and Calvo, R. A. (2020). Vr-rides: An object-oriented application framework for immersive virtual reality exergames. *Softw. Pract. Exp.* 50, 1305–1324. doi:10.1002/spe.2814

Watson, M. R., Voloh, B., Thomas, C., Hasan, A., and Womelsdorf, T. (2019). Use: An integrative suite for temporally-precise psychophysical experiments in virtual environments for human, nonhuman, and artificially intelligent agents. *J. Neurosci. methods* 326, 108374. doi:10.1016/j.jneumeth.2019.108374

Weibel, R. P., Grübel, J., Zhao, H., Thrash, T., Meloni, D., Hölscher, C., et al. (2018). Virtual reality experiments with physiological measures. *JoVE J. Vis. Exp.* 1, e58318. doi:10.3791/58318-v

Weisberg, S. M., Schinazi, V. R., Ferrario, A., and Newcombe, N. S. (2022). Evaluating the effects of a programming error on a virtual environment measure of spatial navigation behavior. *J. Exp. Psychol. Learn. Mem. Cognition* 2022, 0001146. doi:10.1037/xlm0001146

Whyte, J. (2003). Industrial applications of virtual reality in architecture and construction. *J. Inf. Technol. Constr. (ITcon)*. Available at: https://itcon.org/paper/2003/4 8, 43–50.

Zhao, H., Thrash, T., Wehrli, S., Hölscher, C., Kapadia, M., Grübel, J., et al. (2018). A networked desktop virtual reality setup for decision science and navigation experiments with multiple participants. *JoVE J. Vis. Exp.* 1, e58155. doi:10.3791/58155-v