Int. J. Human–Computer Studies 152 (2021) 102649

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Expertise and Experience in VR-supported learning: Achieving a deep non-verbal comprehension of four-dimensional space

Jonny Collins *, Holger Regenbrecht, Tobias Langlotz

Information Science, University of Otago, 60 Clyde Street, Dunedin, New Zealand 9016

ARTICLE INFO

Keywords:
Virtual Reality
Immersion
Interaction
Experience
Learning
Constructivism

MSC:
68U01

ABSTRACT

Immersive Virtual Reality (VR) has shown to be effective in general training and learning applications, but whether it has potential in understanding and learning complex topics is not well researched. In addition, the role of users' expertise in VR-based learning is not really understood.

Here we present findings from our purpose-designed immersive VR system and investigation on whether experts with theoretical knowledge in a certain domain can develop deep non-verbal comprehension beyond pure understanding. Mathematically educated experts as well as non-experts are asked to interact with four-dimensional cubes projected into three-dimensional VR space (hypercubes)—an intentional task not found in real life.

In the first of two studies, we validate the feasibility of the principle subject matter, apparatus, and proposed measurements with 22 participants. This study is based on a seminal study proposal from the 1970's. We use the results of this first study to inform a second study based on a philosophical thought experiment known as Mary's Room. With 70 participants we investigate experience, interaction, and prior knowledge, in an immersive VR learning environment.

We can show that both experts and non-experts benefit from the immersive interaction with hypercubes. Both groups develop a non-verbal comprehension of this theoretical construct. Surprisingly though, experts, with prior theoretical knowledge, benefit stronger.

Our findings have implications for immersive VR learning environments and open a future research space on the importance of and relationships between immersive VR, interaction, understanding, comprehension, and constructivist learning.

1. Introduction

Immersive *Virtual Reality* (VR) provides us with a powerful tool for exploring how learners grasp complex concepts. Since the advent of personal computing, researchers from both technical and pedagogical backgrounds have explored the potential of computer-supported learning. There is a broad range of research with respect to computer-supported education, but one field of increasing popularity is that of immersive learning environments, especially with an emphasis on interactivity (Roussou, 2004; Roussou et al., 2006). VR technology is probably the most well known interface enabling immersive environments and is used in a multitude of application domains (Finkelstein et al., 2010; Goedicke et al., 2018; Harman et al., 2017; Nguyen et al., 2017). One of the key domains for VR is learning and education (Lee and Wong, 2008). Here, prior work has already shown VR to be an effective tool for education (Merchant et al., 2014) although some researchers have identified a general lack of pedagogical guidance in the implementation of VR learning systems (Fowler, 2015; Johnston et al., 2018). A common theme among systems which include pedagogy in their development is the identification of constructivist learning as a philosophy that is coherent with VR learning environments (Chee and Hooi, 2002; Huang et al., 2010; Winterbottom and Blake, 2008). A key element of constructivist learning philosophies is interaction with the world. This prompts us to raise the broader questions: what role does interaction and experience play for immersive VR learning environments? Similarly, how important is interaction and experience for learning in general? While those questions are well researched in general terms, they haven't been studied in-depth for very theoretical, abstract subject matters and also not for the understanding of the relationship between immersive VR and deep non-verbal

* Corresponding author.
E-mail addresses: jonnymcollins@gmail.com (J. Collins), holger.regenbrecht@otago.ac.nz (H. Regenbrecht), tobias.langlotz@otago.ac.nz (T. Langlotz).

https://doi.org/10.1016/j.ijhcs.2021.102649
Received 16 July 2019; Received in revised form 15 March 2021; Accepted 5 April 2021
Available online 16 April 2021

comprehension. This is the focus of our work and holds relevance not only for the Human-computer Interaction (HCI) and VR communities, but also for those concerned with education and computer-supported learning.

Constructivism presents an approach to education which places the learner at the center of the process, and states learning as a constructed knowledge gain based on the experiences of the learner —the result of an individuals interactions with the world (Hanna et al., 2010). There is a large body of work investigating the concept of constructivist learning, for instance the extensive work of Jean Piaget, one of the philosophy's original thinkers (Piaget, 1964). Another proponent of constructivist approaches was P. Arnold. In 1971 he proposed to utilize early forms of VR technology to investigate the practice of constructivist learning (Arnold, 1971; 1972). To our knowledge, he was the first proposing a very early conception of VR technology to present learners with interactive abstract content not otherwise attainable through real-world interaction (different object representations in four-dimensional space). His hypothesis was that if he allowed one learner to interact with the content by manipulating it using actions, and allowed a second learner only to view the interactions and the resulting manipulations of the content, that only the first learner would be able to achieve what he calls a "deep non-verbal comprehension" of the content, later referred to in German by Heinz Von Foerster as "begreifen" (Von Foerster et al., 1992). To make their argument scientifically robust, Arnold and v. Foerster suggest a task subject matter which hasn't been experienced by people before: four-dimensional space. Unfortunately, this work remained in a conceptual state and empirical results have not been reported.

In this work, (1) we report on a study to test the functionality of both our purpose-built research system and the measures proposed by Arnold (Collins et al., 2018). We also took inspiration from his proposed subject matter, four-dimensional space (or 4D space), as an abstract subject matter that participants are unlikely to have experienced before. The results and lessons learned are used to inform our second, main study. In this main study, (2) we investigate the relationship between interactive experience and prior knowledge (expertise) in immersive VR learning environments. We base our second study on a thought experiment known as "Mary's Room", or the knowledge argument, and we are able to demonstrate the value of interaction and experience. Again using the subject matter of 4D space, we show that experts are able to take more advantage of immersive interfaces than laypeople, hence both expertise and interaction are necessary to achieve that deep non-verbal comprehension ("begreifen") with implications for the HCI, VR, and education-technology communities.

2. Background

VR applications targeting the education domain have been studied at an increasing rate over the past two decades within the HCI, VR, and education communities.

This is probably due to VR's unique characteristics which now become affordable to implement in a meaningful way in training and education, amongst other areas. VR can be defined as a computer-generated, three-dimensional environment with which users interact in real-time with immediate feedback leading to a sense of presence in that environment.

Of particular interest to learning is the potential for embodied interaction, which can be decomposed into its cognitive and functional aspects (Salen et al., 2004) for VR. Cognitive interactivity in VR refers to the psychological participation and perception of the environment leading to presence and other experiences. Functional interactivity, mainly in the form of direct manipulation (Nielsen, 2000) of the elements of the environment refers to immersion, or the degree and fidelity of the technical "surroundedness" of the user/learner.

As interaction in general, embodied interaction comprises cognitive and functional characteristics. Kirsh's embodied cognition theory proposal is grounded in four ideas: (1) that interaction with tools influences perception and thinking, (2) that thinking is embodied, (3) that doing contributes more to knowing than seeing, and (4) that sometimes we literally think with things (Kirsh, 2013, pg 3:1). This view can be augmented by Kilteni's (Kilteni et al., 2012) considerations on embodiment as ownership, self-location, and agency. The latter, agency, as having global motor control, is directly contributing to the functional aspects of embodied interaction. For our research we consider both, the cognitive and the functional aspects of embodied interaction.

Training and education using VR systems. The literature can be divided into two categories: 1) training systems, and 2) education systems. Training systems are those designed to train users on specific real world tasks, e.g. within the health care domain (Andreatta et al., 2010; Gallagher and Cates, 2004; Regenbrecht et al., 2011; Seymour, 2002; Sorathia et al., 2017), safety- (Wyk, 2006), navigation- (Bliss et al., 1997), industry assembly task- (Boud et al., 1999), and pre-flight- (Stroud et al., 2005) training. In many cases these works have been able to establish robust skill transfer to the real world (Lehmann et al., 2005). The second category within the literature are those targeting education in the purer sense, rather than skill training.

A recent meta analysis of learning outcomes in K-12 and higher education when applying (desktop-) VR-based instructions analyzed 69 studies (Merchant et al., 2014) showing that systems were effective in terms of gains in learning outcomes. Such systems have targeted subjects such as mathematics and spatial thinking (Hauptman, 2010; Song and Lee, 2002; Yeh, 2004), health sciences (Nicholson et al., 2006; Shim et al., 2003), and language (Yang et al., 2010). However, e.g. Sun et al. presented a system for teaching arts curriculum material which students engaged with, but did not appear to improve learning gains compared to current methods (Sun et al., 2010). Examples for more immersive systems incorporating haptic feedback for teaching astro-physics concepts (Civelek et al., 2014) produced a positive effect on students' achievement as well as their motivation, autonomy, and encouragement. A further example of immersive VR is found in the work of Izatt et al. who present a CAVE system called Neutrino-KAVE to demonstrate neutrino physics to students (Izatt et al., 2014) but unfortunately this system was not evaluated with users. Limniou et al. also use a CAVE approach to teach college students chemistry concepts with students' understanding improving after use of the immersive system (Limniou et al., 2008). ScienceSpace is an HMD-based immersive VR project for teaching complex and abstract scientific concepts at secondary and college level with inconclusive results (Dede et al., 1996). Cheng et al. show HMD-based VR to be effective for teaching cultural interactions in the context of language learning (Cheng et al., 2017).

Constructivism in Immersive Virtual Reality. The inconclusive results of the examples above indicate that there is a lack of research into the underlying pedagogical questions. Researchers have identified a lack of pedagogical consideration in immersive VR learning developments (Fowler, 2015; Johnston et al., 2018). Of the projects that do integrate pedagogical approaches, constructivism and experiential learning are common themes. Particular focus has been put on which aspects of constructivist learning are afforded by VR technologies, and how aspects of learning can be practically implemented in VR systems. E.g. Chee et al. present a collaborative, simulation-based desktop VR system designed on both experiential and constructivist/socio-constructivist principles though the system was not evaluated with users (Chee and Hooi, 2002). Peruzza et al. focus on the learner centered aspects of constructivist principles as a guide for the implementation of a modular desktop VR system demonstrated with physics material which, similarly, was not evaluated (Melchiori Peruzza and Zuffo, 2004).

Recent work (Huang et al., 2010) uses web-based desktop VR technologies for implementation of educational systems and proposes constructivist learning strategies that can be applied when developing VR learning environments. They use two case studies to evaluate VR learning environments concluding with a discussion of their insights. They discovered issues such as environment fidelity (compared to real

world), difficulty to implement, and 3D user interface usability issues. Recent fully immersive technologies and state of the art computational systems are able to address many of these issues. Schwienhorst provides a detailed discussion on the concept of learner autonomy and VR in the context of computer-assisted language learning (Schwienhorst, 2002). Three primary elements are identified as being naturally facilitated by VR: 1) awareness, 2) interaction and collaboration, and 3) experimental, learner centered environments. The second element here is particularly relevant for us. Similarly, Winterbottom identified constructivist "practical values" such as *atomic simplicity, multiplicity, practical exploration, control*, and *reflective process*, which might inform the design of immersive VR learning environments (Winterbottom and Blake, 2004; 2008). Collins et al. used a VR-based learning environment as the use case for a methodology they developed. The methodology leverages emotional response measures to evaluate users' cognitive load and spontaneous moments of insight while conducting novel learning tasks (Collins et al., 2019).

The most relevant related work for us is that of Roussou et al. where user interaction in immersive virtual learning environments is investigated (Roussou, 2004; Roussou et al., 2006). The conclusion of the work suggests that interactivity facilitates (childrens') problem-solving abilities, but conceptual formations and changes were not effected. Rather, passive environments were shown to be more effective for conceptual changes.

Our work follows this path and contributes additional insights on immersive virtual learning environments by investigating experience and interaction in an immersive Virtual Reality learning environment. We investigate the general importance of interaction in immersive virtual learning environments using an abstract subject matter - the 4^{th} spatial dimension or 4D space - while also investigating the effect for people with expert knowledge compared to laypeople. The findings highlight the importance of the interaction component and answer questions regarding their feasibility for users of different expertise which is important for the future development of computer-supported learning environments and the interactions within those environments.

3. Concept and Implementation

When investigating learning in empirical studies, we need to consider the effect of participants' existing knowledge to mitigate any effect in user studies. In 1971, Arnold proposed to use the rather abstract idea of 4D space as subject matter (Arnold, 1971). The assumption is that not many people have a firm spatial understanding of 4D space making it an ideal subject matter. He proposed an apparatus with a stereoscopic VR system as its' core component. The system should be capable of visualizing a 3D projection of a hypercube (a cube in 4D) and further allow for direct manipulation of the visualized hypercube in real-time. In his original proposal the hypercube, or more precisely its' rotation in 4D, is controlled with six dials and is then rendered on a stereoscopic CRT monitor (see Fig. 1 (left)) (Arnold, 1972). Although Arnold's work was revisited at a later date (Von Foerster et al., 1992), the actual apparatus and study appear never to have been realized.

In this work we look specifically at the impact of interaction on learning experiences in immersive VR learning environments. Therefore we have the following research questions:

RQ1: Does VR-based, embodied interaction with a specific, theoretical, never-experienced before subject matter lead to an understanding of it?
RQ2: What influence does the relationship between interaction and expertise have on learning in VR environments?

We have implemented two systems to facilitate our investigation. The first system we implemented closely to Arnold's originally described system (see Fig. 1 (middle)) and the second system is implemented on the same principle, but uses modern immersive VR technology as the interactive and visual interface (see Fig. 1 (right)).

3.1. Subject Matter - Four-dimensional Space

Most of us are capable of understanding one- (1D), two- (2D), and three- dimensional (3D) concepts. We are able to imagine three dimensions which exist perpendicular to each other, however, to attempt imagining a fourth dimension which maintains perpendicularity with the previous three seems near impossible. A hypercube is a 4D cube; all edges are the same length, all internal angles are 90 degrees, but where a regular cube consists of 8 vertices, 12 edges, and 6 faces, a hypercube consists of 16 vertices, 32 edges, and 24 faces (see Fig. 2). A vertex in 4D-space is described by 4 coordinates (x,y,z,w).

We also define transformations, in our case rotation operations, in 4D. In 3D we use Euler angles or Quaternions to describe rotations. Quaternions are robust against the common problem of gimble lock but are mathematically more complex. They encode an axis-angle representation in four values which can be applied to a 3D point. As object representations change between the third and fourth dimensions, so too does rotation. Quaternions are only applicable to 3D space, however

Fig. 1. Experimental Setups - Historic and Modern *Left: the original setup proposed by Arnold has a user sitting at a station with dials for manipulating a 4-Dimensional cube and a stereoscopic viewing platform for visualization of the resulting user manipulations. Middle: our implementation of Arnold's original proposal. Right: our implementation of the original principle but with modern immersive technology.*

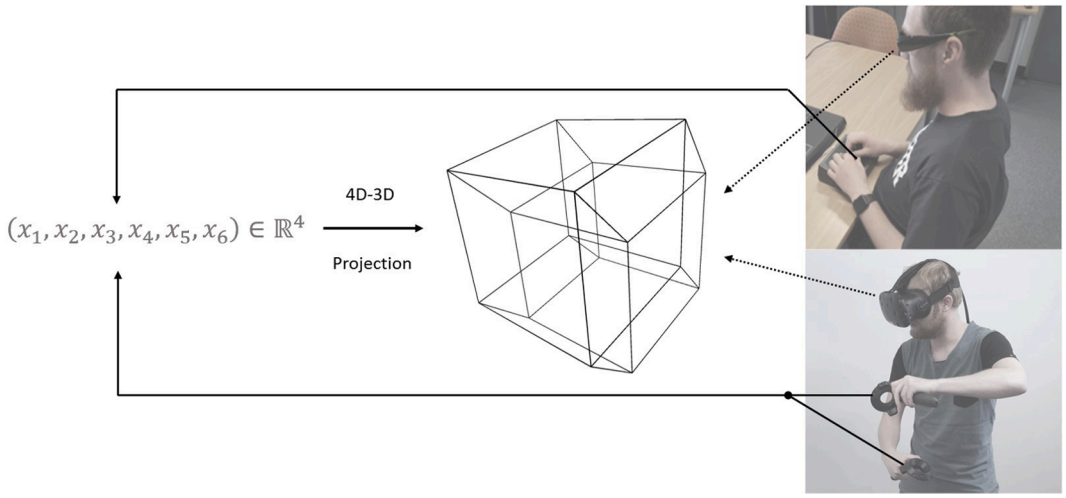

Fig. 2. The Interactive Feedback Loop *Concept of the Hypercube system: An operator (right) is controlling the rotations of the 4D mathematical cube either by rotating the hands in real space (2x3DOF), or turning dials in a board. The manipulated hypercube rotation is then projected into 3D space to be observed by the operator either in an HMD, or with 3D glasses on a screen. This implements a closed interactive feedback loop.*

there are 4D equivalents called Octonions. Octonions are similar to quaternions in principle but contain an additional four imaginary components (resulting in seven values + one real number). However, 4D rotations (Octonions) can be decomposed into a double quaternion representation (Cayley, 1894; Perez-Gracia and Thomas, 2016) which we take advantage of for our implementation.

3.2. System

Our main system for investigating learning in VR is built in Unity3D. The main components are the visualization component and the interaction component.

Unity provides different rendering techniques used in our *visualization component*. We use a line renderer to draw the projection of the hypercube in 3D. Once the 16 vertices of the hypercube are projected from 4D to 3D space (using an orthographic projection), we can draw the 24 edges by connecting the appropriate vertices in 3D.

We implement two visual mediums: (1) stereoscopic display, and (2) HTC Vive. For the first (originally proposed system (Arnold, 1971; 1972)), Unity3D's native stereoscopic support handles the rendering of the graphics on a 3D monitor as a stereo-image which the user can view by wearing stereoscopic glasses. Our second visual medium is the HTC Vive head-mounted display (HMD) which gives operators an immersive 3D viewing experience of our projected hypercube. This provides an advantage over the stereoscopic display approach in that the user can observe different perspectives of the hypercube. These two visualization techniques form one part of the interactive experience.

Corresponding to the two visual mediums described above, we have implemented two different forms of control for manipulating the hypercube: (1) a six dial input device (originally proposed system (Arnold, 1971; 1972)), and (2) two HTC Vive controllers. Hence, this forms the second part of the interactive experience and closes the interactive feedback loop (see Fig. 2). Users are able to interact with a hypercube by mathematically rotating it, and can view the resulting manipulations in 3D.

The six dial device (shown in Fig. 1 (middle)) is driven using a Freetronics Leostick (http://www.freetronics.com.au/) which takes 6 analogue potentiometers as input. Unity3D's input manager allows us to receive the raw input from the Leostick as joystick axes (due to it's configuration) which is then fed into a range transformation operation which maps the raw potentiometer values to Euler angles. Unity3D contains it's own quaternion library for converting Euler angles into quaternions. We have six input dials that are mapped as Euler angles allowing us to generate two separate quaternions which are combined to form an octonion (4D rotation matrix) (Perez-Gracia and Thomas, 2016). The rotation is then applied to the hypercube vertices resulting in direct rotational manipulation.

The second bi-manual interface is the standard HTC Vive controllers. Given the precise tracking of the Vive controllers, we are able to access the orientation of both controllers in space providing us directly with two separate quaternions.

Based on the implementations above, users are able to manipulate a hypercube in two different ways. By turning one of six dials in the first approach, a user will be manipulating the hypercube's rotation about one plane in isolation (where there are the six planes). In our second approach using VR controllers, each axis of a controller is the equivalent of one dial on the board (three axes per controller). Therefore, as the user rotates the controllers in their hands, the rotation of the hypercube is altered.

4. Validation and Feasibility: Study 1

The proposed works of Arnold and v. Foerster (Arnold, 1971; 1972; Von Foerster et al., 1992) provided us with the starting point for our investigation. As mentioned earlier, we decided to use the subject matter, apparatus, and measures proposed by Arnold for our study of interaction and experience in immersive VR learning environments. Arnold, and later v. Foerster, only presented their work as proposals given the limits of technology at the time of publishing. For this reason we needed to validate the elements of their work that we intended to use for our studies. Throughout previous sections we have elaborated on the concept of the fourth spatial dimension and its qualification as subject matter for our experiment because it is not a concept that has likely been taught to, or learned by the average person. It is certainly not a concept we can conceive of just by interacting with the reality around us, however, this also makes it more difficult to measure one's comprehension of such a concept. Arnold's original proposal suggested presenting participants with a set of tasks in order to determine whether they had acquired either a "partial or complete mastery of the situation" (Arnold, 1972), though the task descriptions lacked detail. We used these broad descriptions to guide the design and implementation of our measures. Therefore, the main purpose of this first study was to validate our implementations (stereo and immersive) and the measures to inform a further study.

We had the following hypothesis regarding the study: (1) Our implementation of Arnold's measures will be a valid measure of

comprehension. We specifically focused on the concept of comprehension because it is the part of the learning process referred to by Arnold and is indicative of positive learning outcomes.

4.1. Study Design

The experiment was a single factor between-participant design with two conditions. Our two systems comprised our between-participant conditions: (1) the six-dial interface while viewing on a 3D stereoscopic display (3D glasses) - *desktop experience*, and (2) the Vive controller interface while viewing with an HMD - *immersive experience*. Participants' conditions are pre-randomized. We recruited our participants from the university student and staff populations from a range of disciplines with the only inclusion criterion being age (between 18 and 65). We excluded students affiliated with our research laboratory.

4.2. Measures

In the following we describe the collected data and the tools used for analysis.

Demographics Questionnaire. A demographics questionnaire collecting data including age, gender, ethnicity, vision, and prior VR experience.

Self-assessment Knowledge Questionnaire. We designed a short knowledge questionnaire presenting two questions intended as a self-assessment of the participants' current understanding of 4D space and understanding of the Hypercube (or Tesseract). The items have been: 1) "I understand the concept of the fourth spatial dimension", and 2) "I know what a Hypercube is." Both items were answered on a Likert-like scale from -3 to 3 mapping to "Not at all", and "Very Much", respectively. The purpose of this short questionnaire was only to get an indication of how participants feel about how their experiences impact their "knowledge" or "comprehension" of the subject matter.

Hypercube Assessment Questionnaire. The first of Arnold's original measures is the hypercube assessment questionnaire. The hypothesis was, if a participant is able to distinguish correct and incorrect hypercubes from each other, we can say they have likely formed a 'correct' internal representation of the construct. Therefore, as a measure, we created snapshots of hypercubes rotated in various different ways which comprise a set of correct hypercubes. We then snapshot various different rotated hypercubes with different obscurities (rendering impossible hypercubes). Overall, we created 36 total snapshots which we printed (six per page) to form the paper-based hypercube assessment. The form asks participants to tick only hypercubes they believe to be correct.

Ghostcube Matching Task. The second of Arnold's measures was the ghostcube matching task. A user is presented with two hypercubes in a system, and they had to manipulate one to match the other. It was hypothesized that effective performance on this task is further indicative of whether a participant had a 'correct' internal representation of a hypercube. We implemented this task in Unity and set an upper time limit of 10 minutes to solve one hypercube.

4.3. Procedure

Upon arrival, we welcomed participants, introduced them to the study, gave them a consent form and if agreed, presented the demographics questionnaire. Before beginning the study, participants had to answer the knowledge questionnaire in order to help determine their current knowledge of our subject matter—4D space and the hypercube (Fig. 3-KQ(1)). We then presented to each participant a short explanatory video (Fig. 3-A) introducing participants to the context of the subject matter, 4D space and the hypercube. Assuming little prior knowledge among the participants the video aimed to establish a common base knowledge among the participants as the video went beyond what can be assumed common or general knowledge of 4D space. After watching the video, we again presented the self-assessment knowledge questionnaire (Fig. 3-KQ(2)). At this stage participants diverged into their pre-randomized condition in which they experienced the hypercube on their assigned system, either desktop experience (Fig. 3-B1) or immersive experience (Fig. 3-B2), for a five minute period where they were asked to interact with the hypercube with the aim of 'grasping' the construct. Participants were then asked, for the third and final time, to fill out the self-assessment knowledge questionnaire (Fig. 3-KQ(3)) followed by the hypercube assessment questionnaire described earlier (Fig. 3-C). The final task for participants before being released was the ghostcube matching task (Fig. 3-D). All participants performed this task in the immersive system (HTC Vive).

There was an upper time limit of 10 minutes to complete the task. Time limits were required because: 1) we wanted to mitigate for simulator sickness, 2) the study cannot run for too long due to ethical constraints, and 3) if it turned out that most users of the system could not complete one matching task in 10 minutes, we would have needed to rethink the task itself. Based on prior observations of the system in use, it was decided that 10 minutes was a good starting point. Upon completion, our test environment automatically stopped and stored their participant ID, and the completion time. We asked the participants to spend at least two minutes trying to complete the task after which they were allowed to give up if they so chose. Finally, we thanked the participants, and compensated them for their time.

Upon closing the final experiment, we decided to improve the validity testing of our paper-based hypercube assessment. We recruited an

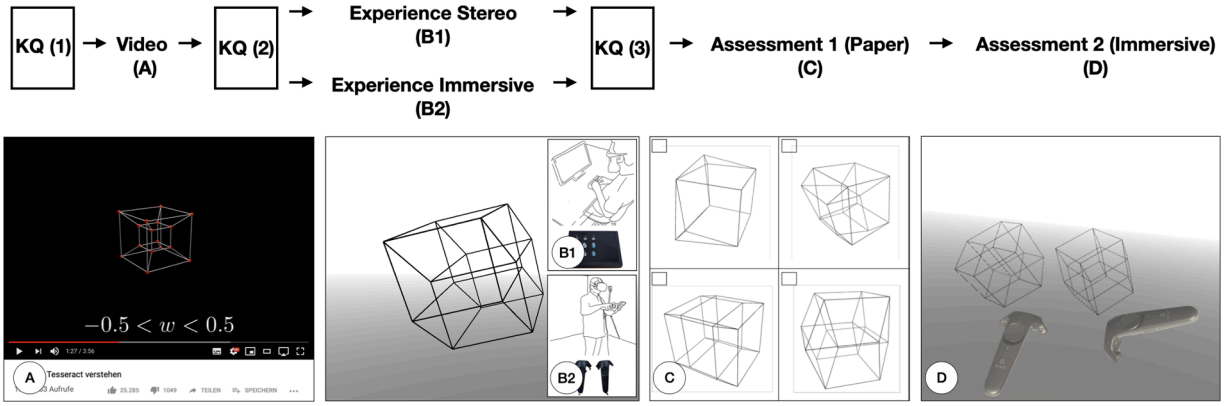

Fig. 3. Study One - Procedure *This figure demonstrates the experimental procedure for Study 1. Participants watched a short video, diverged into their conditions, completed the hypercube assessment, then all participants attempted the ghostcube matching task in the immersive system. KQ (1), KQ (2), and KQ (3) are the instances at which knowledge questionnaires were given to participants.*

additional 11 participants for 10 minutes to fill out the hypercube assessment without having any experience at all. We only asked the participants to distinguish correct and incorrect 4D cubes (hypercubes) from each other. We used the resulting data as a baseline result for our hypercube assessment.

4.4. Results

Demographics Questionnaire. There were 22 participants (15 male, 7 female) in total with a mean age of 25.9. All participants completed the experiment in full with only one participant giving up in the final ghostcube task.

Self-assessment Knowledge Questionnaire. Likert-scales were mapped to a 1-7 range. Questionnaire data was of a non-parametric distribution (Shapiro-Wilk). Wilcoxon Signed-rank tests showed a significant increase in reported ratings between the first and second instances for both questions (Q1: first instance median = 3.0, second instance median = 5.0, $p < 0.01$; Q2: first instance median = 1.0, second instance median = 6.0, $p < 0.01$). No significance was shown between the second and third instances (Q1: third instance median = 6.0; Q2: third instance median = 6.0; $p > 0.05$).

Significance tests for independent groups revealed the stereo group increased significantly between the first and second instances for both questions ($p < 0.05$) but not between second and third instances (Stereo Q1: first instance median = 3.0, second instance median = 5.0, third instance median = 6.0; Stereo Q2: first instance median = 4.0, second instance median = 6.0, third instance median = 6.0). These results were supported by a Friedman's ANOVA (First Question: $X^2(2) = 13.15$, $p < 0.01$, Second Question: $X^2(2) = 16.424$, $p < 0.01$). The immersive group reported significant increases across the first and second instances for both questions ($p < 0.05$) and also increases across the second and third instances ($p < 0.05$) for Q2 only (Immersive Q1: first instance median = 2.0, second instance median = 5.0, third instance median = 6.0; Immersive Q2: first instance median = 1.0, second instance median = 5.0, third instance median = 6.0). For the rest, no significance was found ($p > 0.05$). A Friedman's ANOVA confirmed these results (First Question: $X^2(2) = 17.684$, $p < 0.01$, Second Question: $X^2(2) = 17.897$, $p < 0.01$).

Hypercube Assessment Questionnaire. The assessment data is analyzed in terms of Positive Prediction Power (PPP) and Negative Prediction Power (NPP) which is considered to be a measure of accuracy (Szalma et al., 2006). Fig. 4 (left) and Fig. 4 (middle) show the mean PPP and NPP values respectively, for each condition.

A participant could tick any given hypercube they perceive to be correct whether it is correct or incorrect. This is accounted for in a proposal (Szalma et al., 2006) where it is stated that performance can be measured in terms of Positive Predictive Power (PPP) and Negative Predictive Power (NPP). A participant that ticked all the correct cubes, and none of the incorrect cubes received a PPP value of 1.0 and similarly, a participant that left all incorrect hypercubes unticked and left no correct hypercubes unticked received a NPP value of 1.0.

A Shapiro-Wilk test revealed a parametric distribution. A t-test did not reveal any significant differences in the PPP/NPP results of the stereo ($M=0.65/0.80$, $S.D=0.22/0.08$) and immersive ($M=0.52/0.79$, $S.D=0.15/0.09$) groups ($p > 0.05$).

The additional 11 participants recruited after the initial part of the study were asked to tick pictures they believed were possible/correct 4D cubes. This formed, in essence, an additional participant group for this assessment only. The reason we did this is to further inform us of the sensitivity of the assessment as a measure. The untrained group scored lower mean scores than both original groups for each of PPP and NPP ($M=0.46/0.72$, $S.D=0.09/0.05$). A t-test revealed significance between the desktop (stereo) and untrained group ($p < 0.05$) for both PPP and NPP. No significance was found between the immersive and untrained groups ($p > 0.05$). We applied an ANOVA which confirmed a significant effect of our conditions on participants' PPP performance on the assessment ($F(2,30) = 3.53$, $p < 0.05$, $\omega = 0.37$), but not for NPP ($F(2,30) = 2.71$, $p > 0.05$, $\omega = 0.28$).

Ghostcube Matching Task. All but two participants completed the ghostcube matching task. One participant gave up early and one ran out of the allotted time. These times were excluded from the completion time analysis. Both of these participants were in the stereo condition stream. Upon completion, participants' completion times are recorded in seconds. Shapiro-Wilk tests revealed a non-parametric distribution. A Mann-Whitney U test did not reveal significant differences in the completion times between the groups (Stereo median time = 281.0, immersive median time = 220.0, $p > 0.05$). Fig. 4 (right) presents the ghostcube results.

4.5. Discussion of Results

The main purpose of this study was to validate the subject matter, implementation, and measurements proposed by Arnold. Arnold's measures seemed to be effective, although they should be altered to improve sensitivity. We were able to apply our insights from the study to improve the measures for our next experiment.

For the self-assessed knowledge questionnaire, we found an expected increase in self-perceived understanding of both 4D-space and the hypercube construct for all users. Both groups reported significant increases between the beginning of the study and after watching the video.

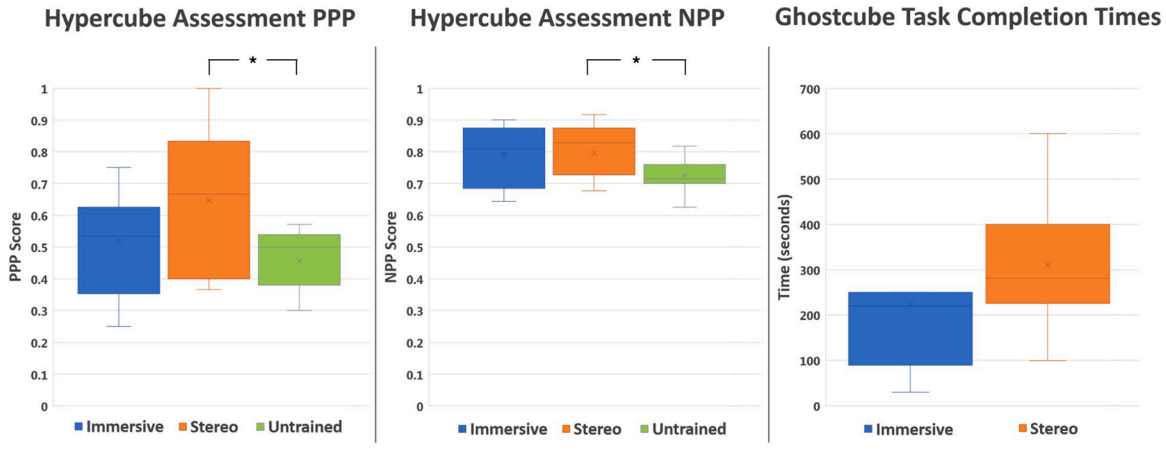

Fig. 4. Study One - Results *Positive and Negative Prediction Power (PPP/NPP) scores for the first and second hypercube assessments (left and right respectively). Significance is found between the stereo (desktop) and untrained groups but no significance is found in other group comparisons. No significant differences found between immersive and desktop groups' ghostcube completion times (right).*

The immersive group also reported a significant increase on the third instance after the exposure for the statement "I know what a hypercube is". We attribute this to the immersiveness of the system which provides participants with higher visualization fidelity and perhaps more importantly, a more "intuitive" or "embodied" interaction metaphor as compared to the desktop system. This measure provided us with broad insight which is useful to give an indication of a participant's perception of their own comprehension.

The hypercube assessment was one of our primary validation targets. While the desktop group scored an observably higher positive prediction power (see Fig. 4 (left)), no significant differences were found between the two groups for either of positive or negative prediction power. The desktop group did however score significantly higher than the added untrained group, but no significance was found between the untrained and the immersive groups. The hypercube questionnaire was presented to participants on paper (2D), and the 2.5D (2 X 2D) nature of the desktop visual medium means that group are seeing a correct hypercube representation rendered closer to what they are presented on the paper questionnaire. This demonstrated that the tool is measuring comprehension, albeit we should mention that it leaves room for a confounding effect. Thus for future studies, we recommend to do the hypercube assessment using similar visual medium as the expertise to reduce the effect of the similarity between the experience and the assessment.

We had all participants attempt the ghostcube matching task in the immersive system to assess how one groups experience would impact their performance or apparent comprehension on a different medium. The expected outcome from this was that the immersive group would be significantly faster at completing the task (given their experience condition), but while they had a slightly lower mean time, no significance was found. This measure is likely a valuable tool for measuring comprehension, but like the hypercube questionnaire, it requires reworking to improve sensitivity. In particular, we realised that we should show different ghostcubes with increasing difficulty to have a better sensitivity as only showing one ghostcube might give only binary results in the worst case (when not solved). We furthermore also witnessed during some demonstrations of our prototype, that some people attempted to solve the task by quickly performing many random orientations. Showing many ghostcubes decreases the chance of being lucky.

In summary, we have found support for our hypothesis that our implementation of Arnold's measures are indicative of subject matter comprehension. Although the measures require tweaking for improvement, the result has been a successful validation of the proposed system. Participants were also able to successfully operate our systems for the purposes of the experiment providing validation of the implementation.

The hypercube assessment questionnaire should be moved from paper-based to VR-based presentation and more hypercubes should be presented of both correct and incorrect states. Similarly, the ghostcube matching task should change to add more hypercubes of varying difficulties. Most importantly, more time should be given to participants for the whole study, in particular for the experience phase. Having participant's report on their own perceptions of their understandings is a valuable tool. Our self-assessment knowledge questionnaire did not provide enough data, so we believe an open ended interview form would be more beneficial for allowing participants to express themselves.

Besides the limitations in the self-assessment knowledge questionnaire, we want to point out here that both groups were sampled from the same student cohort and we did not directly compare both groups but only compared the performance in the assessments over time within each group. Furthermore, we did not see any ceiling or flooring effects in each group that could potentially hide certain effects (e.g. because one group has too much prior knowledge). Thus, we are confident that the effects are actual effects.

In the following study, we propose an approach and study design based on a philosophical thought experiment known as the knowledge argument (Jackson, 1982) for which we utilize our previous implementation, revised measurements, and lessons learned from our first study.

5. Expertise and Interaction: Study 2

The knowledge argument, often referred to as the Mary's room thought experiment, explores the idea that certain knowledge exists such that it is only attainable through conscious experience (Jackson, 1982). A scenario is proposed in which a woman named Mary lives in a black and white room with only black and white belongings for her entire life without ever seeing any colour. During her time in this black and white world she learns everything there is to know about colour. Mary studies all of the biological, chemical, and physical theory. Once she has learned all there is to know about colour, suddenly a colourful red apple appears in her black and white world and she experiences colour for the first time in her life. The question is, does Mary learn anything new about colour having had this experience?

We decided to integrate the knowledge argument into our work to help us investigate the value of the VR experience. We used the same scenario as is described in Mary's room, but had to adapt several variables. We proposed to take a participant that possesses mathematical expertise such that they theoretically understood all there is to know about 4D space and the essence of a Hypercube. This participant would be the equivalent of Mary once she had learned of all the aspects of colour. Then, just as Mary was exposed to a red apple, we would allow our experts to experience 4D space through our system and concurrently observe it. We could then ask the same question - did our mathematical expert attain any new understanding of 4D space through their experience?

5.1. Study Design

Our version of Mary's Room only required one stream of expert participants. However, we were also interested in investigating interaction in more depth, so we decided to provide the experience not only to expert participants, but to layperson participants as well. The final study design was a single factor between-participant design with two conditions where the independent variable was participant expertise. The conditions were: 1) *theoretical subject experts*, and 2) *no theoretical subject knowledge*. The process for each condition was identical; we assessed participants before giving them an experience of 4D space, and then reassessed.

By using this experimental design, we could address our two research questions outlined in section 3. From our questions, we draw the following hypotheses: (H1) Subject experts will perform more effectively on assessments post-exposure, (H2) Subject experts will perform more effectively than subject laypeople, and (H3) Subject experts will report having increased comprehension of the hypercube.

Given that our conditions were based on subject knowledge, we needed to recruit from a specific population. Our subject experts were recruited primarily from mathematics and physics departments and here primarily from staff or students working towards a higher degree. We recruited laypeople randomly from various non-mathematical disciplines. Our inclusion criterion was the same as in the first study —age (between 18 and 65).

5.2. Implementation Changes

For this study, we only used the fully immersive hypercube system from the first study. We did not need the stereo system as we did not require the additional condition due to our focus particularly on interaction and experience in fully immersive VR learning environments. We modified our system in several ways based on our lessons learned and the improvements made to Arnold's measures.

Training Scene. To ensure our participants understood their capabilities we had created a "VR training scene" where participants were

presented with a regular 3D cube and were asked to point at each side of the cube with a controller (see Fig. 5A). This served two purposes. Firstly, it prompted participants to look at objects from different angles hence demonstrating they were able to move around. Secondly, it familiarized them with the interface they would use in later parts of the study.

Hypercube Assessment. In the first study we identified the need to shift the hypercube assessment from paper-based to VR-based. Consequently, for this study we created a VR-based assessment that generated 50 hypercube forms, half are possible/correct hypercube forms, the other half are obscured, impossible/incorrect hypercube forms. We randomly selected the hypercubes and presented the selected hypercube one at a time to the user who had to indicate the validity of the hypercube by pointing at either a green 'yes' or a red 'no' button (see Fig. 5B).

Ghostcube Task. We adapted this task from the previous study. We presented participants with a fixed list of eight hypercubes of varying difficulties. Once a user completed one matching task, the next one was presented (see Fig. 5C). Participant's had a maximum time limit of eight minutes and their completion times were recorded in seconds.

5.3. Measures

We implemented various measures for this experiment which we detail below.

Entry Test. We introduced a test which participants take upon entry which evaluated two primary abilities: 1) their spatial reasoning and mental rotation abilities, and 2) their logical reasoning abilities. The purpose of the test was to support our categorization of participants as experts. The entry test was developed using questions from several sources. We took eight questions of varying degrees of difficulty from the New Zealand Mensa online IQ test (mensa.org.nz), and six spatial rotation questions from fibonicci.com to assess participant's capability at mental rotations and geometric operations.

Demographics Questionnaire. Data collected included age, gender, ethnicity, vision, and prior VR experience.

Semi-structured Interview. Recordings were made of a brief discussion at the beginning and at the end of the study. The purpose was to gauge participants' self-perception of their knowledge of 4D space and of the hypercube. The interviews were semi-structured, with the two primary questions being asked at the beginning: 1) Do you know what 4D space is? And 2) Do you know what a Hypercube is? And the two primary questions asked at the conclusion of the study: 1) Do you feel that your understanding of 4D space or the Hypercube construct has increased? And 2) Do you feel that you attained any measure of grasping 4D space or the Hypercube? Participants were encouraged to express their ideas. Interviews were all recorded and later transcribed to ensure anonymity.

Hypercube Assessment and Ghostcube Task. The assessment and ghostcube task were conducted in the system as described above. The system stored participants' answers to the Hypercube assessments and completion times from the ghostcube matching tasks.

Simulator Sickness Questionnaire (SSQ). Finally, due to long exposure times in our VR system, we used a SSQ to collect data regarding user's potential simulator sickness symptoms.

5.4. Procedure

Upon arrival participants were greeted, introduced to the study, and given a consent form which they signed. They were then presented with the demographics questionnaire followed by the first interview. Participants were introduced to the VR training scene and asked to complete the training task (subsection 5.2). Participants then entered the first hypercube assessment followed by their first ghostcube matching task. Once participants either finished all eight cubes or ran out of time, they entered the experience phase where they were asked to be seated and then had an interactive experience with a single hypercube for a time of 10 minutes (Fig. 5D). They were told there was no specific goal other than to experience the hypercube. We seated participants to mitigate for any fatigue they may experience though they were allowed to move around on the chair within the tracked space to view the scene from different perspectives. After they have completed the experience phase, they stood once again to complete the hypercube assessment followed by the identical ghostcube matching task (eight minute time limit). Upon completion of the second ghostcube matching phase, participants were asked to complete the SSQ followed by a final interview asking the follow-up questions described above. Participants were compensated for their time and released.

5.5. Results

In total we had 70 participants complete the study (40 males, 30 females) with a mean age of 24.1. We were only able to recruit 22 expert participants resulting in an unbalanced sample of 22 experts and 48 laypeople.

Entry Test. Test scores were out of a total of 15. The expert group yielded a higher mean score on the entry test (M=8.0, S.D=1.67) than the layperson group (M=7.21, S.D=2.42) but a t-test found no significance ($p > 0.05$).

Hypercube Assessments. We applied the same analysis of positive and negative prediction power as in the first study (analysis of the paper-based assessment). Shapiro-Wilk tests revealed the PPP results to have a parametric distribution, while the NPP results had a non-parametric distribution. PPP/NPP values achieved for all participants increased between the first ($M=0.58/0.63$, $S.D=0.11/0.12$) and the second ($M=0.67/0.81$, $S.D=0.12/0.17$) hypercube assessments. Fig. 6 shows a graph of the overall PPP/NPP means for the first and second Hypercube assessment. Significance was found for both PPP (*t-test*, $p < 0.01$) and NPP (first assessment median = 0.61, second assessment median = 0.87,

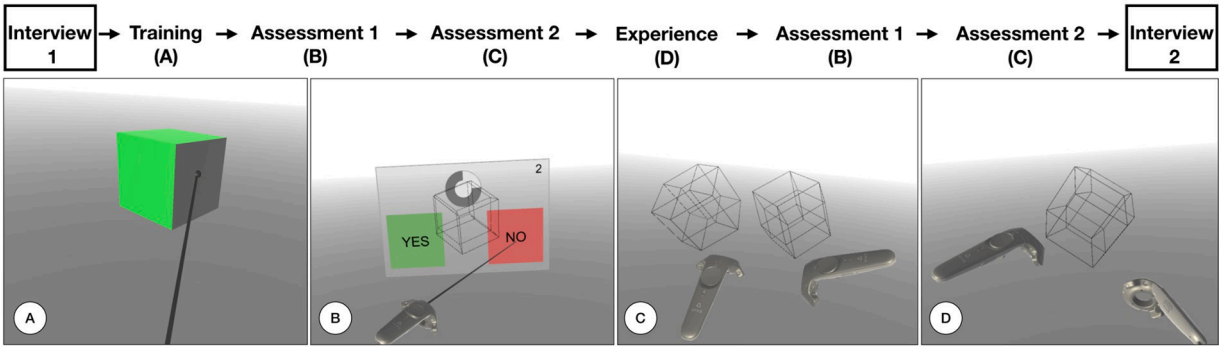

Fig. 5. Study Two - Procedure *This figure demonstrates the procedure for the second study. The images (from left to right) depict the training scene, the revised hypercube assessment, the ghostcube task, and the experience task. The procedure order is shown above the images.*

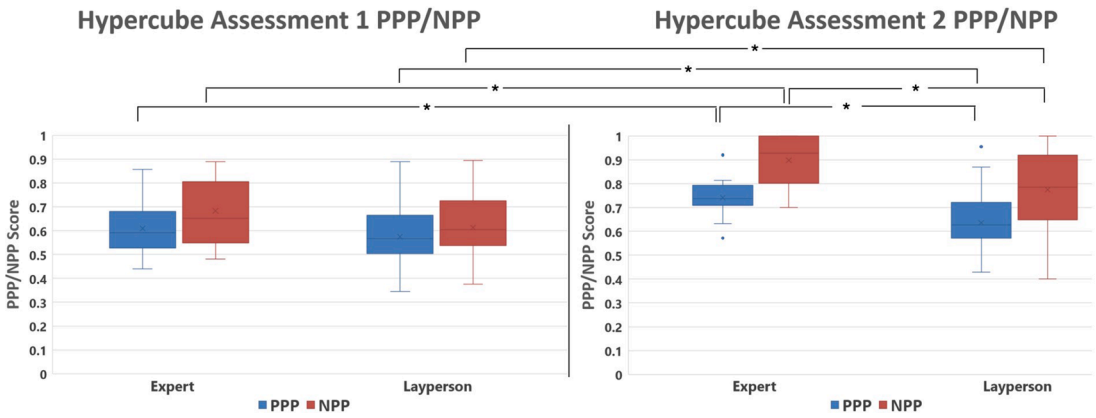

Fig. 6. Study Two - Hypercube Assessment Results *Hypercube Assessment results between expert and layperson groups for the first and second assessments broken down into Position (blue) and Negative (orange) Prediction Power (PPP/NPP). Both groups made significant improvements (especially for NPP). The expert group's gain was also significantly higher than the layperson group.*

$p < 0.01$ *(Wilcoxon)*). Significance testing between expert and layperson groups revealed no significance for either PPP (layperson mean/s.d = 0.57/0.11, expert mean/s.d = 0.61/0.11, $p > 0.05$ *(t-test)*) or NPP (layperson median = 0.57, expert median = 0.61, $p > 0.05$ *(Man-Whitney U)*). Significance tests for the second assessment revealed the expert group scored higher than the laypersons group on both PPP (layperson mean/s.d = 0.64/0.12, expert mean/s.d = 0.74/0.07, $p < 0.01$ *(t-test)*) and NPP (first assessment median = 0.61, second assessment median = 0.87, $p < 0.01$ *(Wilcoxon)*) scores.

Ghostcube Task. Results were analyzed in two ways: 1) number of cubes solved and 2) time taken for each cube solution. There were a total of 8 cubes to be solved during the ghostcube matching task. All data pertinent to the number of cubes solved was non-parametric so either Wilcoxon or Mann-Whitney U tests were used for significance testing. For all participants, there were no significant differences found ($p > 0.05$) for the number of hypercubes solved between the first (median no. cubes = 3.0) and second (median no. cubes = 3.0) task instances.

Significance tests between expertise groups for the first ghostcube task did not reveal significant differences ($p > 0.05$) in the number of cubes solved (expert group: $M=3.05$, $S.D=0.57$; layperson group: $M=2.85$, $S.D=1.17$). Significant differences were found ($p < 0.05$) for the second ghostcube task between the expert group ($M=3.82$, $S.D=1.5$) and the layperson group ($M=2.94$, $S.D=1.02$). When testing within groups from the first to second task instance, significance was found only for the expert group ($p < 0.05$).

Completion time data is shown in Fig. 7. We only considered the first three target hypercubes (1, 2, 3) in the matching task due to the low number of participants that solved any further cubes. Median completion times dropped significantly between the first (19.45, 34.65, and 96.45) and second (19.00, 24.95, and 48.30) ghostcube tasks for the second and third target cubes ($p < 0.01$, *Wilcoxon*), but not for the first cube where the mean rose slightly (first GC mean = 37.37, second GC mean = 44.44, $p > 0.05$, *Wilcoxon*).

The expert group's median completion times reduced for all three cubes between the first (19.95, 24.15, and 76.50) to the second (15.95, 9.05, and 40.95) ghostcube tasks where Wilcoxon tests revealed significant reductions for the 2nd and 3rd cubes ($p < 0.05$) but not for the 1st ($p > 0.05$). The same test revealed a similar trend for the layperson group with means mostly dropping from the first (18.50, 41.95, and 99.20) to the second (19.30, 27.80, and 52.20) task, with significant reductions found only for the 3rd hypercube ($p < 0.01$) but not for the 1st or 2nd ($p > 0.05$).

Simulator Sickness Questionnaire. Participants reported negligible SSQ scores with the mean reported score being 0.29 (on a scale of 0-3). The highest ratings were placed on "eye strain" and "fullness of the head", with scores of 0.70 and 0.64 respectively.

5.6. Discussion of Results

Our first hypothesis (H1) was that expert participants would perform more effectively between the first and second assessments (after having had the experience) despite their theoretical expertise. This hypothesis is supported in our findings for both the hypercube assessment (visual), and the ghostcube task (interactive). The layperson participant group

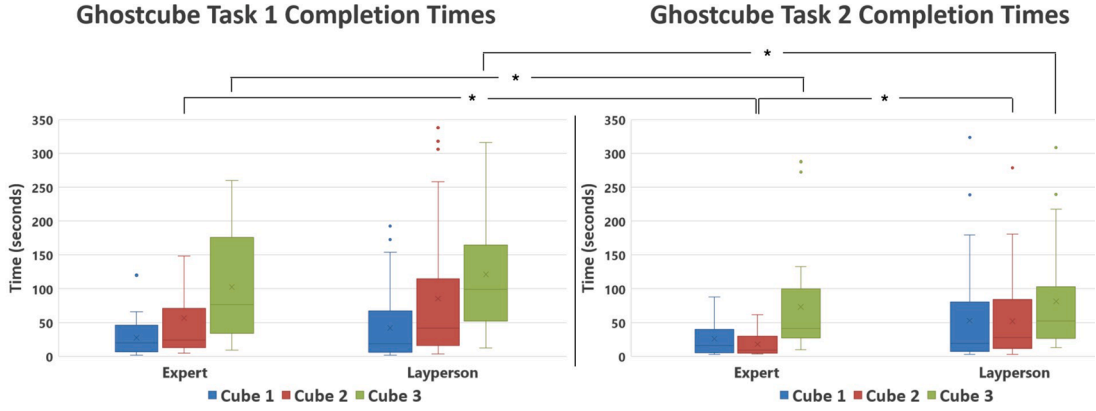

Fig. 7. Study Two - Ghostcube Results *Ghostcube task completion times for the first three target ghostcubes. The difficulty of the cubes is reflected in the completion times. Similarly to the hypercube assessment, both groups mostly improved, namely for the 2nd and 3rd cubes, but the expert group have the most significant completion time reductions.*

also improved significantly which is an expected outcome, but when testing the second assessment differences between expert and layperson groups, the expert groups performance was significantly better. This is the case for the hypercube assessment, and the number of cubes solved between the first and the second ghostcube assessments. This also contributes to our second hypothesis (H2) that experts would perform more effectively than layperson participants. Support for H1 and H2 is also found in the analysis of the ghostcube completion times.

The overall mean time taken to solve the first target ghostcube actually rose from 37.37 to 44.44 seconds. This was also the easiest of the target cubes to solve and the rise is likely due to the starting position of (nervous) participants being very close to the solution. This time actually dropped slightly for the expert group from 27.78 to 26.17 seconds, and for the layperson group, rose from 41.95 to 52.99 seconds which is the reason for the overall rise again supporting H2.

Our 3rd hypothesis (H3) is that experts would report on achieving an improved comprehension from having had the experience. The general consensus among the expert participants was similar to that of laypeople. They reported the experience as useful and while they often reported on gaining something, it was intangible or, for them, indescribable. This could be hinting towards Arnold's "deep non-verbal comprehension" but given the lack of complete competency in the tasks, it is unlikely this was comprehensively achieved. Experts made statements such as "... I think being able to see it and manipulate it yourself really helps...". There were reports of still using trial and error instead of knowing what to do: "... I had no idea how to do it so I just had to try several axes slowly and see which one did it and which did not...". Hypothesis H3 can not be supported through statistical methods, and by iterating the interview transcriptions and highlighting answers to pertinent questions, we can not say that we found evidence of complete certainty from any participant with regards to their comprehension of the subject matter. This is expected given the limited time of exposure.

Overall, there are a couple of observations. Firstly, it was a bit surprising to see that despite having theoretical knowledge in the subject matter, the experts did not score significantly better in the entry tests when compared to laypeople. We argue that this is another indicator for the concept of "begreifen" or the deep non-verbal comprehension which is different from general knowledge that the experts had before. Or to use the Mary's room analogy, knowing everything about colour does not fully replace actually seeing colour and experts can still learn something despite existing conceptual knowledge. Furthermore, both groups significantly benefit from the experience and thus the experience can add in general to a better understanding of the subject matter. However, people with existing knowledge (experts) benefit more from the expertise when compared to laypeople. This has a couple of potential consequences. Firstly, it demonstrates that VR systems like ours would benefit everyone but in particular its benefiting users when complementing other forms of learning, thus when people had already acquired conceptual knowledge before the experience.

There are also a few limitations of our current approach. Firstly, our groups of experts and laypeople were assigned based on background (e.g. working or studying towards a higher degree in mathematics or physics versus recruitment from a cohort with no mathematical background). Despite an informal interview to check for understanding of the subject matter, the experts had of course a different level of understanding and it is not necessarily encompassing all knowledge as in the original Mary's room experiment. Another aspect to highlight again is the fact that the groups were not balanced, thus we had more laypeople than experts because the latter were harder to find.

Finally, we would like to point out that we have used an inside-VR approach to questionnaire use. As Schwind et al. (2019) pointed out, completing questionnaires in VR can increase the consistency of the variance while producing the same presence scores. It also mitigates the dependency on retrospective judgement of feelings and perceptions.

6. Conclusion

Constructivist learning can benefit from Virtual Reality learning and Virtual Reality Learning can benefit from constructivist learning. Within our scope of design, implementation, and study, namely geometric-mathematical comprehension (or "begreifen" to stick with the term from the original study), we could show this reciprocal relationship identifying the complexities of constructivist learning with and in VR. For instance, the challenges associated with assessing comprehension became evident in Study 1 by presenting participants with 2D images of complex 4D geometries however the use of VR for learning and assessment as such is working. Study 2 could show that "Mary" would gain a learning experience from seeing the red apple —the immersive, interactive experience of a projected 4D cube improves the deep non-verbal comprehension ("begreifen"). In addition, we could show that not only experts but non-experts benefit from such an experience, and that experts benefit more than non-experts. Such a finding has implications for (a) the design of interactive VR learning environments and consequently the HCI and VR research communities and (b) for computer supported learning and learning in general. It raises the question of how much expertise is needed and sufficient for effective interaction with the subject matter but it answers the question whether both, expertise and interaction, are needed. They are.

In that sense, Arnold and Von Foerster triggered the right questions about "seeing and doing", even if they couldn't really tackle them appropriately because of the lack of technological advancement. Nowadays, we are in a position to actually put constructivist learning practices into an immersive VR experience and with this potentially open up a wide field of research and practice on future forms of learning.

There is certainly little disagreement amongst educationalists and teachers, especially the major group of people who are following approaches inspired by constructivist learning, that experience and conceptual knowledge are paramount for developing understanding. However, our research goes beyond that in three ways: (1) Mere and pure understanding can be extended towards deep non-verbal comprehension ("begreifen") by way of immersive VR interaction. (2) Not only non-experts in a certain domain can benefit from immersive VR, but also experts. (3) Experts can benefit even more from immersive VR than non-experts. While we could only show this for the very specialised example of understanding and "begreifen" of hypercubes, we would assume that our findings can be extended to other, theory-grounded, abstract matters.

Also, we cannot claim that any arbitrary form of interaction with an immersive virtual environment would lead to understanding and "begreifen". However, if the mode of interaction, here the interplay of the two manual 3D rotations with the specific feedback of the 4D projection, is designed in a meaningful way, novel learning experiences can be delivered.

What we can possibly claim is that our domain experts connected their theoretical knowledge with a novel experience to develop a new understanding, or as we would argue, a deeper non-verbal comprehension of a subject matter they already had a good understanding of. If we project this finding onto other areas of expert education then we might conclude that the "knowledge of the book" can always be amended by the "experience of the hand", figuratively speaking.

There is much room for future work in this space to focus on interesting and important questions around: what efficacy actually means and how it can be measured, the different and novel ways we might learn with VR in the future, the individual's approach towards abstract problem solving, or the range and areas of learning applications VR might be suitable for. How much and what forms of understanding and experience are needed for Mary to deeply comprehend a multi-coloured, multi-dimensional apple?

CRediT authorship contribution statement

Jonny Collins: Conceptualization, Methodology, Software, Writing - original draft, Writing - review & editing, Visualization, Investigation, Formal analysis. **Holger Regenbrecht:** Conceptualization, Methodology, Writing - review & editing, Visualization, Resources, Supervision. **Tobias Langlotz:** Conceptualization, Methodology, Writing - review & editing, Visualization, Resources, Supervision.

Declaration of Competing Interest

None

Acknowledgments

We would like to thank all participants in our studies, the Human-Computer Interaction Lab people, and our colleagues, in particular Russel Butson. We would also like to thank the Biological Computer Laboratory (BCL) at the University of Illinois for fruitful conversations and the granted permission to use their illustration for Fig. 1 (Left).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ijhcs.2021.102649.

References

Andreatta, P.B., Maslowski, E., Petty, S., Shim, W., Marsh, M., Hall, T., Stern, S., Frankel, J., et al., 2010. Virtual Reality Triage Training Provides a Viable Solution for Disaster-preparedness. Academic Emergency Medicine 17 (8), 870–876. https://doi.org/10.1111/j.1553-2712.2010.00728.x.

Arnold, P., 1971. Experiencing the Fourth Spatial Dimension. Accomplishment Summary 70 (71), 201–215.

Arnold, P., 1972. A Proposal for a Study of the Mechanisms of Perception of, and Formation of Internal Representations of, the Spatial Fourth Dimension. Accomplishment Summary 71 (72), 223–235.

Bliss, J.P., Tidwell, P.D., Guest, M.A., 1997. The Effectiveness of Virtual Reality for Administering Spatial Navigation Training to Firefighters. Presence: Teleoperators and Virtual Environments 6 (1), 73–86. https://doi.org/10.1162/pres.1997.6.1.73.

Boud, A.C., Haniff, D.J., Baber, C., Steiner, S.J., et al., 1999. Virtual reality and augmented reality as a training tool for assembly tasks. 1999 IEEE International Conference on Information Visualization (Cat. No. PR00210), pp. 32–36. https://doi.org/10.1109/IV.1999.781532.

Cayley, A., 1894. The collected mathematical papers of Arthur Cayley, Vol. 7. The University Press.https://books.google.co.nz/books?hl=en&lr=&id=UfQ7AQAAMAAJ&oi=fnd&pg=PR1&dq=The+collected+mathematical+papers+of+Arthur+Cayley&ots=zlbdiIU-T2&sig=RliO38IkaJEhEoaIYKFWpnBhdUE.

Chee, Y.S., Hooi, C.M., 2002. C-VISions: socialized learning through collaborative, virtual, interactive simulations. International Society of the Learning Sciences. http://dl.acm.org/citation.cfm?id=1658616.1658789.

Cheng, A., Yang, L., Andersen, E., 2017. Teaching Language and Culture with a Virtual Reality Game. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 541–549. https://doi.org/10.1145/3025453.3025857.

Civelek, T., Ucar, E., Ustunel, H., Aydın, M.K., 2014. Effects of a Haptic Augmented Simulation on K-12 Students' Achievement and their Attitudes towards Physics. Eurasia Journal of Mathematics, Science and Technology Education 10 (6), 565–574. https://doi.org/10.12973/eurasia.2014.1122a.

Collins, J., Regenbrecht, H., Langlotz, T., 2018. Back to the future: Constructivist learning in virtual reality. Adjunct Proceedings of the IEEE International Symposium for Mixed and Augmented Reality 2018 (Unpublished).

Collins, J, Regenbrecht, H, Langlotz, T, Can, YS, Ersoy, C, Butson, R, 2019. Measuring cognitive load and insight: A methodology exemplified in a virtual reality learning context. 2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, pp. 351–362.

Dede, C., Salzman, M.C., Loftin, R.B., et al., 1996. ScienceSpace: virtual realities for learning complex and abstract scientific concepts. Proceedings of the IEEE 1996 Virtual Reality Annual International Symposium, pp. 246–252. https://doi.org/10.1109/VRAIS.1996.490534.

Finkelstein, S., Nickel, A., Barnes, T., Suma, E.A., et al., 2010. Astrojumper: Motivating Children with Autism to Exercise Using a VR Game. CHI '10 Extended Abstracts on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 4189–4194. https://doi.org/10.1145/1753846.1754124.

Fowler, C., 2015. Virtual reality and learning: Where is the pedagogy? British Journal of Educational Technology 46 (2), 412–422. https://doi.org/10.1111/bjet.12135.

Gallagher, A.G., Cates, C.U., 2004. Virtual reality training for the operating room and cardiac catheterisation laboratory. The Lancet 364 (9444), 1538–1540. https://doi.org/10.1016/S0140-6736(04)17278-4.

Goedicke, D., Li, J., Evers, V., Ju, W., 2018. VR-OOM: Virtual Reality On-rOad Driving siMulation. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 165:1–165:11. https://doi.org/10.1145/3173574.3173739.

Hanna, D., David, I., Francisco, B., 2010. The Nature of Learning: Using Research To Inspire Practice. OECD publishing.https://books.google.co.nz/books?hl=en&lr=&id=306PApBeLTwC&oi=fnd&pg=PA3&dq=The+Nature+Of+Learning+Dumont&ots=-JxgFLLJnM&sig=zMvWBZve_u7jPMQEjKOSbbh3AuI.

Harman, J., Brown, R., Johnson, D., 2017. Improved Memory Elicitation in Virtual Reality: New Experimental Results and Insights. Conference on Human-Computer Interaction. IFIP.https://link-springer-com.ezproxy.otago.ac.nz/chapter/10.1007/978-3-319-67684-5_9.

Hauptman, H., 2010. Enhancement of spatial thinking with Virtual Spaces 1.0. Computers & Education 54 (1), 123–135. https://doi.org/10.1016/j.compedu.2009.07.013.http://www.sciencedirect.com/science/article/pii/S0360131509001894.

Huang, H.-M., Rauch, U., Liaw, S.-S., et al., 2010. Investigating learners' attitudes toward virtual reality learning environments: Based on a constructivist approach. Computers & Education 55 (3), 1171–1182. https://doi.org/10.1016/j.compedu.2010.05.014. http://www.sciencedirect.com/science/article/pii/S0360131510001466.

Izatt, E., Scholberg, K., Kopper, R., et al., 2014. Neutrino-KAVE: An immersive visualization and fitting tool for neutrino physics education. 2014 IEEE Virtual Reality (VR), pp. 83–84. https://doi.org/10.1109/VR.2014.6802062.

Jackson, F., 1982. Epiphenomenal Qualia. The Philosophical Quarterly (1950-) 32 (127), 127–136. https://doi.org/10.2307/2960077.

Johnston, E., Olivas, G., Steele, P., Smith, C., Bailey, L., et al., 2018. Exploring Pedagogical Foundations of Existing Virtual Reality Educational Applications: A Content Analysis Study. Journal of Educational Technology Systems 46 (4), 414–439. https://doi.org/10.1177/0047239517745560.

Kilteni, K., Groten, R., Slater, M., 2012. The sense of embodiment in virtual reality. Presence: Teleoperators and Virtual Environments 21 (4), 373–387.

Kirsh, D., 2013. Embodied cognition and the magical future of interaction design. ACM Transactions on Computer-Human Interaction (TOCHI) 20 (1), 1–30.

Lee, E.A.-L., Wong, K.W., 2008. A Review of Using Virtual Reality for Learning. In: Pan, Z., Cheok, A.D., Müller, W., El Rhalibi, A. (Eds.), Transactions on Edutainment I. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 231–241. https://doi.org/10.1007/978-3-540-69744-2_18.Lecture Notes in Computer Science.

Lehmann, K.S., Ritz, J.P., Maass, H., Çakmak, H.K., Kuehnapfel, U.G., Germer, C.T., Bretthauer, G., Buhr, H.J., et al., 2005. A Prospective Randomized Study to Test the Transfer of Basic Psychomotor Skills From Virtual Reality to Physical Reality in a Comparable Training Setting. Annals of Surgery 241 (3), 442–449. https://doi.org/10.1097/01.sla.0000154552.89886.91.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1356982/.

Limniou, M., Roberts, D., Papadopoulos, N., et al., 2008. Full immersive virtual environment CAVETM in chemistry education. Computers & Education 51 (2), 584–593. https://doi.org/10.1016/j.compedu.2007.06.014.http://www.sciencedirect.com/science/article/pii/S0360131507000747.

Melchiori Peruzza, A.P.P., Zuffo, M.K., 2004. ConstruiRV: constructing knowledge using the virtual reality. ACM, pp. 180–183. https://doi.org/10.1145/1044588.1044624.

Merchant, Z., Goetz, E.T., Cifuentes, L., Keeney-Kennicutt, W., Davis, T.J., et al., 2014. Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis. Computers & Education 70, 29–40. https://doi.org/10.1016/j.compedu.2013.07.033.http://www.sciencedirect.com/science/article/pii/S0360131513002108.

Nguyen, C., DiVerdi, S., Hertzmann, A., Liu, F., et al., 2017. Vremiere: In-Headset Virtual Reality Video Editing. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 5428–5438. https://doi.org/10.1145/3025453.3025675.

Nicholson, D.T., Chalk, C., Funnell, W.R.J., Daniel, S.J., et al., 2006. Can virtual reality improve anatomy education? A randomised controlled study of a computer-generated three-dimensional anatomical ear model. Medical Education 40 (11), 1081–1087. https://doi.org/10.1111/j.1365-2929.2006.02611.x.

Nielsen, J., 2000. A layered interaction analysis of direct manipulation. Unpublished online paper 9 (28), 1992.Retrieved from http://www.useit.com/papers/directmanipulation.html.

Perez-Gracia, A., Thomas, F., 2016. On Cayley's Factorization of 4d Rotations and Applications. Advances in Applied Clifford Algebras 1–16. https://doi.org/10.1007/s00006-016-0683-9.

Piaget, J., 1964. Part I: Cognitive development in children: Piaget development and learning. Journal of Research in Science Teaching 2 (3), 176–186. https://doi.org/10.1002/tea.3660020306.

Regenbrecht, H., McGregor, G., Ott, C., Hoermann, S., Schubert, T., Hale, L., Hoermann, J., Dixon, B., Franz, E., et al., 2011. Out of reach? - A novel AR interface approach for motor rehabilitation. 2011 10th IEEE International Symposium on Mixed and Augmented Reality, pp. 219–228. https://doi.org/10.1109/ISMAR.2011.6092389.

Roussou, M., 2004. Interactivity and Conceptual Learning in Virtual Environments for Children. CHI '04 Extended Abstracts on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 1049–1050. https://doi.org/10.1145/985921.985973.

Roussou, M., Oliver, M., Slater, M., et al., 2006. The virtual playground: an educational virtual reality environment for evaluating interactivity and conceptual learning. Virtual Reality 10 (3), 227–240. https://doi.org/10.1007/s10055-006-0035-5.

Salen, K., Tekinbaş, K.S., Zimmerman, E., 2004. Rules of play: Game design fundamentals. MIT press.

Schwienhorst, K., 2002. Why Virtual, Why Environments? Implementing Virtual Reality Concepts in Computer-Assisted Language Learning. Simulation & Gaming 33 (2), 196–209. https://doi.org/10.1177/1046878102332008.

Schwind, V., Knierim, P., Haas, N., Henze, N., 2019. Using presence questionnaires in virtual reality. Proceedings of the 2019 CHI conference on human factors in computing systems, pp. 1–12.

Seymour, N.E., 2002. Virtual Reality Training Improves Operating Room Performance. Annals of Surgery 236.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1422600/.

Shim, K.-C., Park, J.-S., Kim, H.-S., Kim, J.-H., Park, Y.-C., Ryu, H.-I., et al., 2003. Application of virtual reality technology in biology education. Journal of Biological Education 37 (2), 71–74. https://doi.org/10.1080/00219266.2003.9655854.

Song, K.S., Lee, W.Y., 2002. A virtual reality application for geometry classes. Journal of Computer Assisted Learning 18 (2), 149–156. https://doi.org/10.1046/j.0266-4909.2001.00222.x.

Sorathia, K., Sharma, K., Bhowmick, S., Kamidi, P., 2017. Pragati - A Mobile Based Virtual Reality (VR) Platform to Train and Educate Community Health Workers. Conference on Human-Computer Interaction. IFIP. https://doi.org/10.1007/978-3-319-68059-0_51.

Stroud, K. J., Harm, D. L., Klaus, D. M., 2005. Preflight Virtual Reality Training as a Countermeasure for Space Motion Sickness and Disorientation. http://www.ingentaconnect.com/content/asma/asem/2005/00000076/00000004/art00006.

Sun, K.-T., Chan, H.-T., Meng, K.C., 2010. Research on the application of virtual reality on arts core curricula, pp. 234–239. https://doi.org/10.1109/ICCIT.2010.5711063.

Szalma, J.L., Hancock, P.A., Warm, J.S., Dember, W.N., Parsons, K.S., 2006. Training for vigilance: Using predictive power to evaluate feedback effectiveness. Human factors 48 (4), 682–692. https://doi.org/10.1518/001872006779166343.

Von Foerster, H., von Glasersfeld, E., Hejl, P.M., 1992. Einführung in den Konstruktivismus. Piper.http://tocs.ulb.tu-darmstadt.de/5718433X.pdf.

Winterbottom, C., Blake, E., 2004. Designing a VR interaction authoring tool using constructivist practices. ACM, pp. 67–71. https://doi.org/10.1145/1029949.1029961.

Winterbottom, C., Blake, E., 2008. Constructivism, virtual reality and tools to support design. ACM, pp. 230–239. https://doi.org/10.1145/1394445.1394470.

Wyk, E.v., 2006. Improving Mine Safety Training Using Interactive Simulations. Association for the Advancement of Computing in Education (AACE), pp. 2454–2459.https://www.learntechlib.org/primary/p/23352/.

Yang, J.C., Chen, C.H., Chang Jeng, M., et al., 2010. Integrating video-capture virtual reality technology into a physically interactive learning environment for English learning. Computers & Education 55 (3), 1346–1356. https://doi.org/10.1016/j.compedu.2010.06.005.http://www.sciencedirect.com/science/article/pii/S03601315100016 5X.

Yeh, A., 2004. VRMath: Knowledge Construction of 3d Geometry in Virtual Reality Microworlds. CHI '04 Extended Abstracts on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 1061–1062. https://doi.org/10.1145/985921.985979.