A Comparison of Virtual and Physical Training Transfer of Bimanual Assembly Tasks

María Murcia-López and Anthony Steed

Abstract—As we explore the use of consumer virtual reality technology for training applications, there is a need to evaluate its validity compared to more traditional training formats. In this paper, we present a study that compares the effectiveness of virtual training and physical training for teaching a bimanual assembly task. In a between-subjects experiment, 60 participants were trained to solve three 3D burr puzzles in one of six conditions comprised of virtual and physical training elements. In the four physical conditions, training was delivered via paper- and video-based instructions, with or without the physical puzzles to practice with. In the two virtual conditions, participants learnt to assemble the puzzles in an interactive virtual environment, with or without 3D animations showing the assembly process. After training, we conducted immediate tests in which participants were asked to solve a physical version of the puzzles. We measured performance through success rates and assembly completion testing times. We also measured training times as well as subjective ratings on several aspects of the experience. Our results show that the performance of virtually trained participants was promising. A statistically significant difference was not found between virtual training with animated instructions and the best performing physical condition (in which physical blocks were available during training) for the last and most complex puzzle in terms of success rates and testing times. Performance in retention tests two weeks after training was generally not as good as expected for all experimental conditions. We discuss the implications of the results and highlight the validity of virtual reality systems in training.

Index Terms—Learning transfer, virtual reality, assembly, training

1 Introduction

The availability of consumer virtual reality technology has raised the manufacturing industry's interest in virtual training for manual assembly tasks. Virtual environments could deliver cost-efficient, safe and potentially effective training. If proven adequate, virtual training would also allow for the completion of operator instruction prior to the installation of physical workstations, tools and components. This would accelerate the end-to-end manufacturing process and, consequently, increase efficiency of production. However, more evidence is needed to ascertain the effectiveness of virtual environments for training as opposed to more traditional forms of training.

In this paper, we present a study that compares the effectiveness of virtual and traditional paper- and video-based training transfer of a bimanual assembly task, motivated by previous research [3, 10]. In a between-subjects experimental design, participants were trained to solve three six-piece burr puzzles in a virtual training environment or a physical training environment. The conditions were designed to account for situations in which the physical puzzle blocks are available or not during training. The conditions were also devised to include static instructions (paper) or combinations of static and animated instructions (video or 3D animations). Table 1 introduces the experimental condition types, acronyms and definitions. Table 2 shows a classification of the experimental conditions according to instruction type and block availability during training.

Following training, participants were asked to solve physical versions of the puzzles (referred to as immediate testing). Participants then completed a retention session, two weeks after the training (referred to as retention testing). During the course of the study, participants answered mental rotations tests and questionnaires measuring several aspects of the experience.

We tested three hypotheses about the effectiveness of training in each of the conditions being compared in the study. The first hypothesis (H1) was that conditions in which the physical blocks were available during training (PB and PV_IB) would yield a higher number of successful puzzle completions during immediate and retention testing. The second hypothesis (H2) was that the conditions in which both static and animated instructions were available during training (PV_I, PV_IB and V_EA) would result in lower assembly times during immediate and retention testing. The third hypothesis (H3) was that condition PV_IB, with animated instructions (video) and physical blocks, would yield the highest performance as measured by immediate and retention success rates and assembly testing times. Although we expected some conditions to deliver worse or better performance, we had no hypothesis on the full order so all the analysis presented in this paper is two-tailed.

Immediate testing results showed some support for the first hypothesis, some support for the second hypothesis and some support for the third hypothesis. Retention performance was lower than expected for all conditions both in terms of success rates and completion times and did not provide evidence to support any of the three hypotheses.

The remainder of this paper is organised as follows. In Section 2 we review related work on learning transfer in immersive mixed reality systems. Section 3 presents the experimental design and hypotheses. In Section 4 we introduce the methodology and experimental setup. In Section 5 we report the results of the study. Section 6 discusses the results, limitations and future work. Section 7 concludes.

2 Related work

Previous research has highlighted the effectiveness of immersive mixed reality training in different disciplines, including military training, medical training and vehicle driving simulators [17,21], as well as navigation and spatial knowledge training [8, 23], amongst others. Despite the recognised success in the aforementioned fields, studies on immersive virtual training transfer of procedural and assembly tasks have reported contrasting results.

Hall and Horwitz compared retention of procedural knowledge of equipment operation in an immersive virtual environment and in a 2D computer environment and found no significant differences [7]. They claimed that virtual reality training may not be superior to conventional electronic media for training certain skills. Gavish et al. evaluated the use of virtual reality and augmented reality technology for industrial maintenance and assembly task training [5]. They concluded that an augmented reality platform was more suitable for training of this type of tasks and encouraged further evaluation of virtual reality based training.

- *María Murcia-López is with University College London. E-mail: maria.murcia.13@ucl.ac.uk.*
- *Anthony Steed is with University College London. E-mail: a.steed@ucl.ac.uk.*

Manuscript received 11 Sept. 2017; accepted 8 Jan. 2018. Date of publication 19 Jan. 2018; date of current version 18 Mar. 2018. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TVCG.2018.2793638

Fig. 1. One of the three 3D printed burr puzzles used in the study.

In a more recent study Gonzalez-Franco et al. compared collaborative conventional face-to-face training with a mixed reality training setup for a manufacturing procedure of an aircraft door [6]. Their results indicated that performance levels yielded by the immersive mixed reality training system were not significantly different from the conventional face-to-face training format. Rose et al. evaluated the transfer from a virtual environment to the real world of a simple sensorimotor task [16]. Overall, virtual training resulted in equivalent or even better real world performance than real or physical training for the task. However, they advise that their findings may not apply to other types of training tasks.

Sowndararajan et al. found an effect of level of immersion in memorising a complex procedure [20]. In their study, participants trained in the system with the higher level of immersion (a large L-shaped projection display) completed tasks significantly faster and with fewer errors than participants trained in the system with lower level of immersion (using a typical laptop display).

Other studies have shown effective learning transfer in virtual environments with the addition of haptic force-feedback devices. For instance, Adams et al. conducted a study to explore the benefits of haptic feedback for virtual training of a manual task [1]. They reported that force-feedback was a requirement for higher learning transfer in virtual environments.

Our study is inspired by the work of Carlson et al. in 2015 [3], itself motivated by previous work [10, 13, 19]. In a between-subjects experimental design, Carlson et al. compared the effectiveness of virtual bimanual haptic training versus traditional physical training of an assembly task consisting of a six-piece burr puzzle. Their results indicated that physically trained participants initially outperformed virtually trained participants. However, virtually trained participants improved their testing times after two weeks. Results also showed that virtual training was enhanced by using coloured blocks as they helped participants remember the assembly process. We run a similar task comparing paper- and video-based training with virtual training in the absence of a haptic force-feedback device.

We agree with Carlson et al. in that 3D burr puzzles are suitable proxy tasks or abstractions of context-specific manual assembly tasks, such as engine assembly operations at vehicle manufacturing plants. We therefore decided to use the same type of task in our study. Following their reported methods, we complemented the training task with a series of mental rotation tests to distribute participants amongst the condition groups in our between-subjects experimental design [2, 14, 22]. We also decided to colour-code the puzzle blocks and instructions as well as to use a semi-transparent virtual representation of the hands in the virtual environment [11, 12], amongst other recommendations made by the authors which are further explained in Section 3.

Our study extends and builds on previous work by comparing a number of virtual and physical training formats, the latter representing the most common formats (video and paper instructions) in current assembly process training programmes. The main aim of this research is to verify whether exposure to a virtual training environment is sufficient for effective training. We are specifically interested in situations in which haptic devices are not available and when the physical components and tools used in the process are not accessible during training.

3 Experimental Design and Hypotheses

Inspired by previous research [3], in our study we used three different colour-coded versions of a six-piece burr puzzle for the assembly task (see Figure 1). Burr puzzles have been commonly used for assembly task training studies in the past because they provide a recognisable and adequately complex model in which participants must follow a specific procedure in order to solve them [3, 10]. However, our study differs from previous work in that no haptic devices were used. In addition, we are interested in whether consumer virtual reality systems are sufficient for effective training.

In our study, participants were trained and tested in assembling three versions of a six-piece burr puzzle. To provide increasing difficulty, the first three blocks had been preassembled for the first puzzle, the first two for the second and none for the third. This meant that participants had to remember a higher number of steps in the assembly process over the course of the experimental task for each puzzle.

Following a between-subjects experimental design, participants were trained to solve each puzzle by adding the corresponding unassembled blocks in one of six experimental conditions (see Table 1). Experimental conditions were designed to account for scenarios in which blocks are not available (P and PV_I), physical blocks are available (PB and PV_IB) or virtual blocks are available (V_E and V_EA) during training (see Table 2 for a classification of the experimental conditions). The physical experimental conditions (P, PB, PV_I and PV_IB) were designed to encompass combinations of paper- and video- based instructions. The virtual experimental conditions (V_E and V_EA) involved a virtual version of the paper instructions, with or without 3D animations showing how to correctly assemble the puzzle, and always with virtual blocks to practice during training. All instructions (static and animated) were colour-coded to match the physical puzzle blocks.

Following training and after a short break, participants were asked to assemble a 3D printed physical version of the corresponding puzzle within a given time. Participants were asked to attend a retention session, two weeks after the training, in which they were asked to solve the same puzzles in the same order and within the same time constraints. We measured success rates as well as training and testing times. Sessions were complemented by a series of mental rotations tests as well as questionnaires and debrief interviews.

As part of their recommendations for future work, Carlson et al. suggested adding a snap-to-fit function or constraint system [18] to alleviate the time that virtually trained participants spent attempting to fit and assemble the virtual blocks [3]. We followed this recommendation and added such functionality in the virtual training environment. We also followed their recommendation to make the selection of a block in the virtual environment to cause a change of colour instead of just causing a change in transparency, as participants in their study reported that it was difficult to discern transparent pieces against the transparent virtual representation of the glove. In their discussion they mentioned individual differences for interaction between the two hands, as some participants showed a preference for the haptic device or the glove for predominant use. We therefore decided to make interaction ambidextrous, meaning all operations were designed to be performed equally by the left hand and the right hand.

We made the following hypotheses:

- H1: The conditions in which the physical blocks were available during training (PB and PV_IB) would yield a higher number of successful puzzle completions during immediate and retention testing. This relates to the experience (or lack of) built around manipulating and assembling the physical blocks during training.
- H2: The conditions in which static and animated instructions (video or 3D animations) were available during training (PV_I, PV_IB and V_EA) would result in lower assembly times during immediate and retention testing, as participants would have received richer visualisation on how to assemble the blocks during training.
- H3: Condition PV_IB, with physical blocks and animated instructions (video), would yield the best performance as measured by immediate and retention success rates and assembly testing times. This hypothesis is based on H1 and H2.

Table 1. Experimental condition types, acronyms and definitions. See Table 2 for a classification of the experimental conditions according to instruction type and block availability during training. Please note the choice of acronym V_I to represent *video* and V_E to represent *virtual environment* to avoid any confusion in making reference to the experimental conditions throughout the paper.

Type	Acronym	Definition
Physical	P	Paper instructions
	PB	Paper instructions and physical blocks
	PV_I	Paper instructions and assembly process video
	PV_IB	Paper instructions, assembly process video and physical blocks
Virtual	V_E	Virtual paper instructions and virtual blocks
	V_EA	Virtual paper instructions and virtual blocks, with assembly process animations

Table 2. Classification of the experimental conditions according to instruction type (static or static and animated) and block availability (no blocks, physical blocks or virtual blocks) during training. See Table 1 for experimental condition types, acronyms and definitions.

	Physical		Virtual
	No blocks	Physical blocks	Virtual blocks
Static instructions	P	PB	V_E
Static and animated instructions	PV_I	PV_IB	V_EA

Fig. 2. Phyiscal lab where the experiment took place (left) and analogous virtual environment (right).

4 METHOD

4.1 Participants

A total of 60 participants (30 female, 30 male; average age 26.51 years, $SD = 6.47$) were recruited from the student and staff population at University College London (UCL). All participants signed a consent form and the study was approved by the UCL Research Ethics Committee (Project ID: 6708/004). Participants were paid £15 for participation. A screener questionnaire was used to filter out potential participants who enjoy solving 3D puzzles or who have any type of colour-blindness. Eligible participants were assigned to the different experimental conditions based on individual results for Purdue's Visualisation of Rotations Test [2] to avoid any possible bias between groups, ensuring a similar mean score for the test in each of the experimental condition groups. Likewise, an equal number of females and males were assigned to each group.

4.2 Materials

The user study was conducted in a lab at UCL. The room consisted of a 3.1 meters long by 2.7 meters wide by 4.0 meters high room. A virtual replica of the laboratory was modeled for the virtual enviornment used in the virtual experimental conditions. Figure 2 contains images of the physical room and analogous virtual environment. An Oculus Rift Consumer Version 1, two Oculus Touch controllers and two Oculus sensors were used for the virtual experimental conditions. The virtual environment was rendered at scale 1:1 in Unity 5.6.0 without VSync at $90FPS$ in each eye on an Intel Core i7-4770K CPU @ $3.50GHz$, with $16GB$ RAM and Nvidia GeForce GTX 1080 GPU running Windows 8.1 Pro. The Oculus Avatar SDK 1.15.0 [9] was used to include hand presence and interaction for the Oculus Touch controllers. The Burr Tools 0.6.3 software was used to digitally create and solve the three versions of the six-piece burr puzzles as well as to generate the paper instructions and assembly process videos [15]. The physical puzzle blocks were 3D printed using a Ultimaker 2+ 3D printer with a 0.4mm nozzle and standard settings, with PLA 3D printing material. Preassembled blocks for the first and second puzzles were glued together. Paper instructions were printed on A3 paper and attached to 5mm A3 foamboards. Assembly videos were presented using VLC 2.2.3 on a 13-inch mid 2014 MacBook Pro laptop running macOS 10.12.2.

4.3 Physical training environment

Participants assigned to the physical experimental conditions (P, PB, PV_I and PV_IB) were seated on a stool in front of the table in the lab on which the blocks had been placed in the correct initial configuration for each puzzle. Participants were seated facing the table and were told that they could adjust the distance to it if they wished to.

Paper instructions were designed to show the initial configuration of the blocks at the top and the assembly process steps at the bottom (see Figure 3). For the first two puzzles, blocks that had been preassembled and the corresponding steps in the assembly process were faded out. The orientation of the images of the blocks in the instructions was randomly selected for each puzzle. For those experimental conditions involving paper instructions, these were placed against the wall on the table in front of the participant. Assembly process videos were generated using Burr Tools [15] and showed a step-by-step animation of the assembly process from the perspective matching the one in the paper instructions. The laptop was placed on the table in front of the participant. Participants could interact with the video (play, pause, stop, rewind, and fast forward) using the VLC user interface.

For those experimental conditions in which the physical blocks were available during training (PB and PV_IB) these were initially placed on the table following the same configuration as the paper instructions. Preassembled puzzles were placed behind the blocks.

4.4 Virtual training environment

Participants assigned to the virtual experimental conditions (V_E and V_EA) were seated on a stool in the center of the lab. They were then asked to put on the Oculus Rift and hold the two Oculus Touch controllers with the experimenter's help. The virtual environment showed the virtual replica of the room and table used in the physical environment in front of them, with the blocks for the corresponding puzzle arranged in the correct configuration. Participants were seated facing the virtual table and were told that they could adjust the distance to it if they wished to. For the first two puzzles (in which two or three of the blocks had been preassembled) participants could see the preassembled puzzle hovering over the table in front of them. Virtual paper instructions were presented against the wall on the table in the same location as the physical paper instructions were presented in the physical training environment.

Using the Oculus Avatar SDK 1.15.0 [9], virtual hands were rendered using the default shader (see Figure 4). Participants could then manipulate the 3D environment by grabbing the virtual puzzle blocks. They could hold the trigger button to grab unassembled puzzle blocks and the grip button to move and rotate assembled blocks as a single unit. Participants could grab any block at any given time, but only the correct block in the assembly process could be attached to the puzzle. No physics constraints were added to the blocks meaning they could be moved through each other and through the virtual hands and table.

Visual feedback was provided to aid participants in learning the assembly process during training. When participants grabbed the correct block in the assembly process, a blue transparent preview block was shown in the puzzle, indicating where the block had to be assembled. Participants had the option to deactivate the block preview. A blue highlight was used to indicate what the next block in the assembly process was. This highlight would then turn to red when the block collided with the preview block, indicating that the piece was near its correct location but in the wrong orientation. The highlight would turn green when the block was within an angle of twenty degrees from the

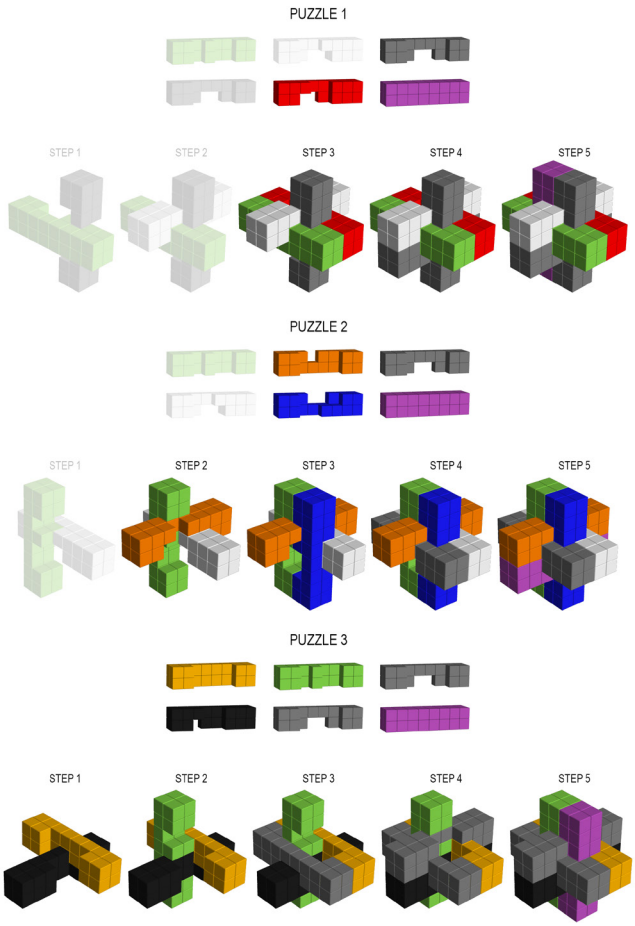

Fig. 3. Assembly instruction sheet for each of the three burr puzzles used in the study. Each instruction sheet contains a diagram of the six pieces and five ordered steps needed to solve the puzzle. Preassembled pieces and steps for Puzzles 1 and 2 were faded out.

correct orientation. If the participant released the trigger when the block showed a green highlight, it would snap into the correct location and the participant could move on to assemble the next piece or reset the puzzle. No audio or vibration feedback was used in the experience.

A user interface with virtual buttons was added on the right-hand side of the virtual table. Buttons were represented by blue spheres which the participant could interact with by touching them, after which they would turn to grey and back to blue to indicate that the interaction was successful. For participants in the V_E and V_EA conditions, two buttons were available: RESET and HELP ON/OFF. Interacting with the RESET button would immediately relocate all blocks in their initial positions so participants could restart the assembly process whenever they wished. The HELP ON/OFF button acted as a toggle to activate and deactivate the blue transparent preview of the block in the puzzle so participants could practice assembling the puzzle with and without the visual aid.

For participants in the V_EA condition, two more buttons were added: NEXT STEP and REPLAY LAST STEP. The NEXT STEP button would trigger the animation of the assembly of the next block in the process. The REPLAY LAST STEP would reposition the last block assembled in its original location on the table and animate its assembly onto the puzzle.

All interactions in the virtual training environment could be equally carried out using either hand and participants could concurrently complete one interaction with each hand. For example, a participant could grab and rotate the assembled pieces with one hand and grab the next block to attach with the other hand.

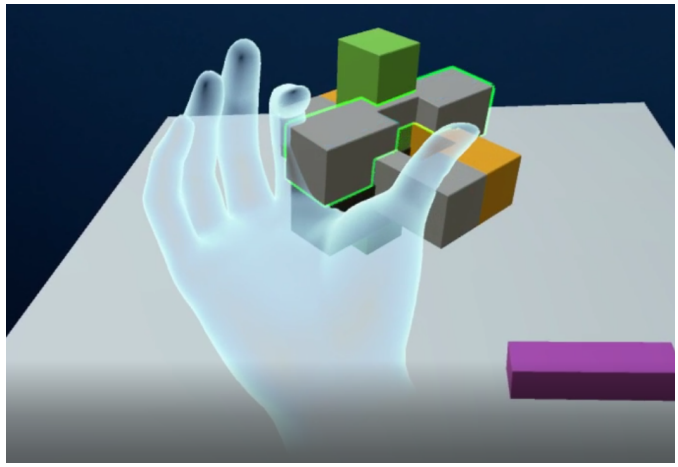

Fig. 4. Screenshot of a participant grabbing a virtual block and assembling it onto the 3D puzzle. The green highlight indicates on the block is colliding with its preview block and within twenty degrees from the correct orientation. By releasing the trigger button of the Oculus Touch controller the virtual block would snap into its correct location.

4.5 Procedure

The experimental task consisted of two lab sessions. The first session comprised training and immediate testing. The second session, two weeks after the first, comprised retention testing. Figure 5 shows an outline of the experimental task. Before the first lab session, participants were asked to read and sign an online informed consent form and answer a digital version of Purdue's Visualisation of Rotations Test [2] used to pre-allocate participants to the experimental conditions. Participants also answered a background questionnaire with a specific focus on prior experience with videogames, 3D modelling software and virtual environments.

During the first lab session, participants were asked to sign a paper copy of the consent form and asked to read an information sheet with written instructions describing the experimental task. In this session, participants completed a familiarisation task and three trials, each with a training and a testing stage. The three trials corresponded with each of the three burr puzzles in increasing order of difficulty. During the familiarisation task participants were introduced to the physical or virtual training environment depending on the experimental condition they had been assigned to. A sample assembly task involving piling up rectangular blocks was used and participants were able to familiarise themselves with the paper instruction format, the video player and the interactive virtual environment, accordingly.

For each of the trials, the training stage involved learning to assemble the corresponding puzzle in one of the six experimental conditions in a maximum time of eight minutes. During the testing stage participants were asked to assemble the physical 3D puzzle in a maximum of three minutes. Time limits for training and testing were defined through piloting of the experimental task. In each trial participants completed the training stage and, after a thirty second break, the testing stage. They then completed a questionnaire at the end of each trial (see Table 3). For both training and testing, participants were told what the limit times were and were advised that they could end the stage before the time expired if they wished to. Participants were also told that the initial configuration of the blocks on the table during training would match the initial configuration of the blocks during testing and the paper instructions.

Participants were asked to try their best if they were in doubt as to how to assemble the puzzles during testing. An experimenter was present at all times during the experimental task to manage cables for those particpiants in the virtual experimental conditions and provide guidance on the different phases of the experimental task. After completing all trials, participants were interviewed regarding the strategies

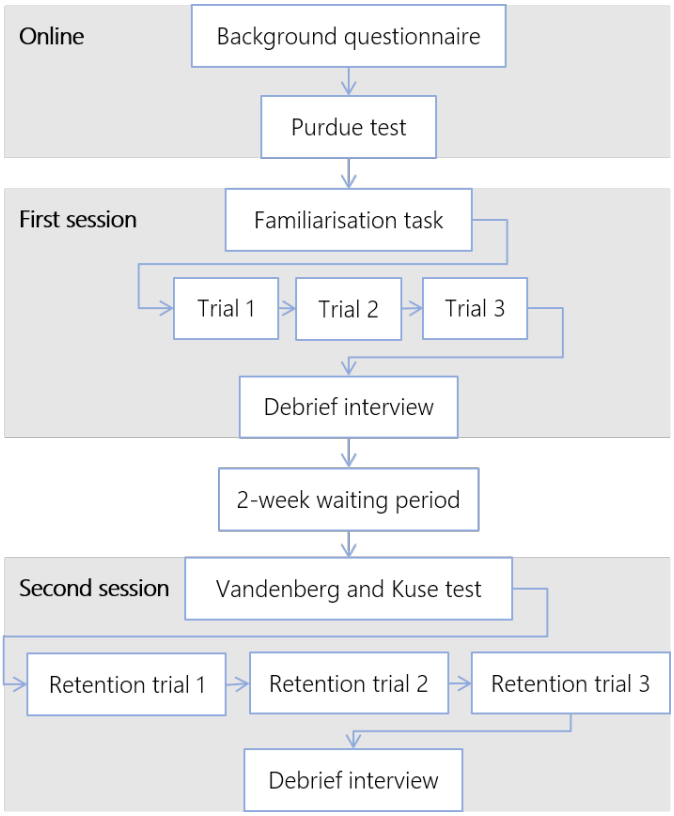

Fig. 5. Overview of the experimental procedure.

used throughout the sessions.

After a waiting period of two weeks, participants returned to the lab for the second session. In this session participants were asked to complete a paper version of the Vandenberg and Kuse Mental Rotations Test [22]. They then completed the retention test for each of the three puzzles, in which they were asked to solve the three burr puzzles from the first session without a training phase, in the same order and in a maximum of three minutes. They completed the same questionnaire from the first session at the end of each retention trial (see Table 3). After completing all retention trials they were interviewed regarding strategies used throughout the session.

5 Results

5.1 Types of errors

Unsuccessful puzzle completions during immediate and retention testing were due to one of two reasons. In most cases, participants did not complete the 3D puzzles within the given maximum time (180s). On the other hand, a low number of participants decided to stop the time before the upper limit thinking that they had successfully solved the puzzle. However, close inspection showed that they had not correctly assembled the pieces. Completion time values for both immediate and retention testing were corrected by assigning the upper time limit (180s) to all unsuccessful attempts.

5.2 First session

5.2.1 Training times

Boxplots with training times for each of the puzzles are shown in Figure 6. Non-parametric statistical analysis was performed for training times because our data was not normally distributed as shown by a Shapiro-Wilk test.

A Kruskal-Wallis H test showed that there was an overall statistically significant difference in training times for the first puzzle between the different experimental conditions, $\chi^2(5) = 25.648, p < 0.001$, with a mean rank score of 15.35 for P, 38.85 for PB, 13.85 for PV_I, 40.15 for PV_IB, 36.25 for V_E and 38.55 for V_EA.

A Kruskal-Wallis H test showed that there was an overall statistically significant difference in training times for the second puzzle between the different experimental conditions, $\chi^2(5) = 22.764, p < 0.001$, with a mean rank score of 22.50 for P, 40.00 for PB, 10.60 for PV_I, 36.70 for PV_IB, 37.45 for V_E and 35.75 for V_EA.

A Kruskal-Wallis H test showed that there was no overall statistically significant difference in training times for the third puzzle between the different experimental conditions, $\chi^2(5) = 10.701, p = 0.058$, with a mean rank score of 21.95 for P, 33.50 for PB, 18.90 for PV_I, 35.70 for PV_IB, 37.15 for V_E and 35.80 for V_EA.

Pairwise comparisons were performed using Dunn's procedure [4] with a Bonferroni correction for multiple comparisons with adjusted p-values. These are displayed in Table 4. Note that pairwise comparisons for puzzles in which the Kruskal-Wallis H test showed no overall statistically significant difference have not been included.

The post hoc analysis revealed statistically significant differences in training times for the first puzzle. There was a statistically significant difference between P (mean rank = 15.35) and PB (mean rank = 38.85) ($p = 0.036$), PV_IB (mean rank = 40.15) ($p = 0.020$) and V_EA (mean rank = 38.55) ($p = 0.041$). There was also a statistically significant difference between PV_I (mean rank = 13.85) and PB (mean rank = 38.85) ($p = 0.018$), PV_IB (mean rank = 40.15) ($p = 0.010$) and V_EA (mean rank = 38.55) ($p = 0.021$).

The post hoc analysis revealed statistically significant differences in training times for the second puzzle. There was a statistically significant difference between PV_I (mean rank = 10.60) and PB (mean rank = 40.00) ($p = 0.002$), PV_IB (mean rank = 36.70) ($p = 0.010$), V_E (mean rank = 37.45) ($p = 0.007$) and V_EA (mean rank = 35.75) ($p = 0.015$).

5.2.2 Immediate testing success rates

A binomial logistic regression was performed to ascertain the effects of experimental condition on the likelihood that participants succeed at assembling each puzzle during the immediate testing phase. Figure 7 shows the number of successful and unsuccessful completions of each puzzle for all experimental conditions. PV_IB was chosen as the reference category as this was the condition that produced the highest number of successful puzzle completions, overall.

The binomial logistic regression model was not statistically significant, $\chi^2(5) = 8.809, p = 0.117$ for the first puzzle. The model explained 18.3% (Nagelkerke R^2) of the variance in success rate and correctly classified 61.7% of cases. The Wald criterion demonstrated that only condition P made a significant contribution to prediction (p = 0.016). The model suggested that participants in this condition were 0.05 times as likely to successfully assemble the first puzzle than participants in the reference category (PV_IB).

The binomial logistic regression model was statistically significant, $\chi^2(5) = 12.016, p = 0.035$ for the second puzzle. The model explained 24.7% (Nagelkerke R^2) of the variance in success rate and correctly classified 71.7% of cases. The Wald criterion demonstrated that P and PV_I made a significant contribution to prediction ($p = 0.016$ and $p = 0.035$, respectively). The model suggested that participants in the P experimental condition were 0.048 times as likely to successfully assemble the second puzzle than participants in the reference category (PV_IB). The model suggested that participants in the PV_I experimental condition were 0.074 times as likely to successfully assemble the second puzzle than participants in the reference category (PV_IB).

The binomial logistic regression model was statistically significant, $\chi^2(5) = 24.255, p < 0.001$ for the third puzzle. The model explained 45.8% (Nagelkerke R^2) of the variance in success rate and correctly classified 78.3% of cases. The Wald criterion demonstrated that P and PV_I made a significant contribution to prediction (p = 0.035 and p = 0.007, respectively). The model suggested that participants in the P experimental condition were 0.074 times as likely to successfully assemble the third puzzle than participants in the reference category (PV_IB). The model suggested that participants in the PV_I experimental condition were 0.028 times as likely to successfully assemble the third puzzle than participants in the reference category (PV_IB). Condition

Table 3. Trial questionnaire.

Dimension	Question	Likert scale extremes
Difficulty	Please rate the difficulty of the task you just completed.	1: Very difficult - 5: Very easy
Ease of use	Please rate the ease of use in assembling parts in the training environment.	1: Very difficult - 5: Very easy
Seriousness	Please rate how seriously you took the task.	1: Very unseriously - 5: Very seriously

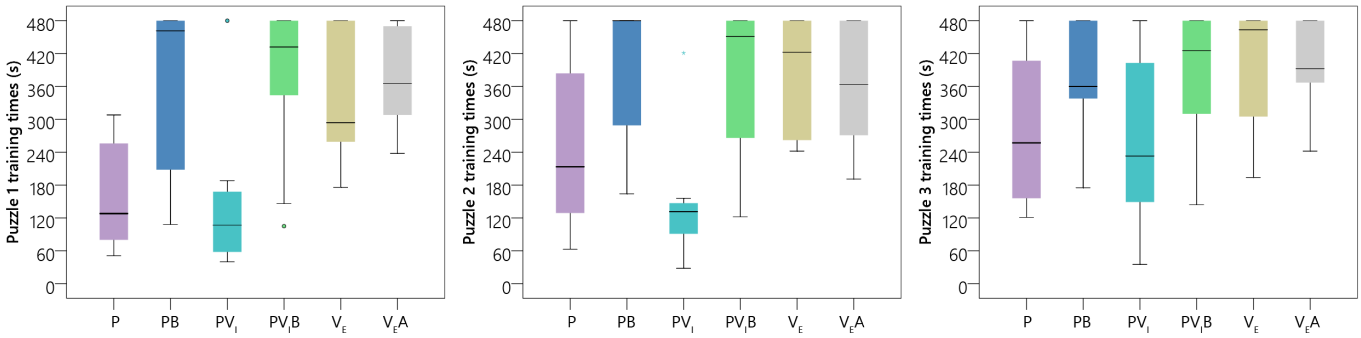

Fig. 6. Boxplot containing training times for each of the puzzles. Medians are shown as dark horizontal lines. Boxes represent the interquartile ranges (IQR). Whiskers represent either the extreme data points or extend to 1.5 x IQR. Outliers (data points outside the whiskers) are shown by circles. A value, X, is an outlier if X $<$ lower quartile $-$ 1.5 x interquartile range or if X $<$ upper quartile $+$ 1.5 x IQR. See Table 4 for pairwise interactions.

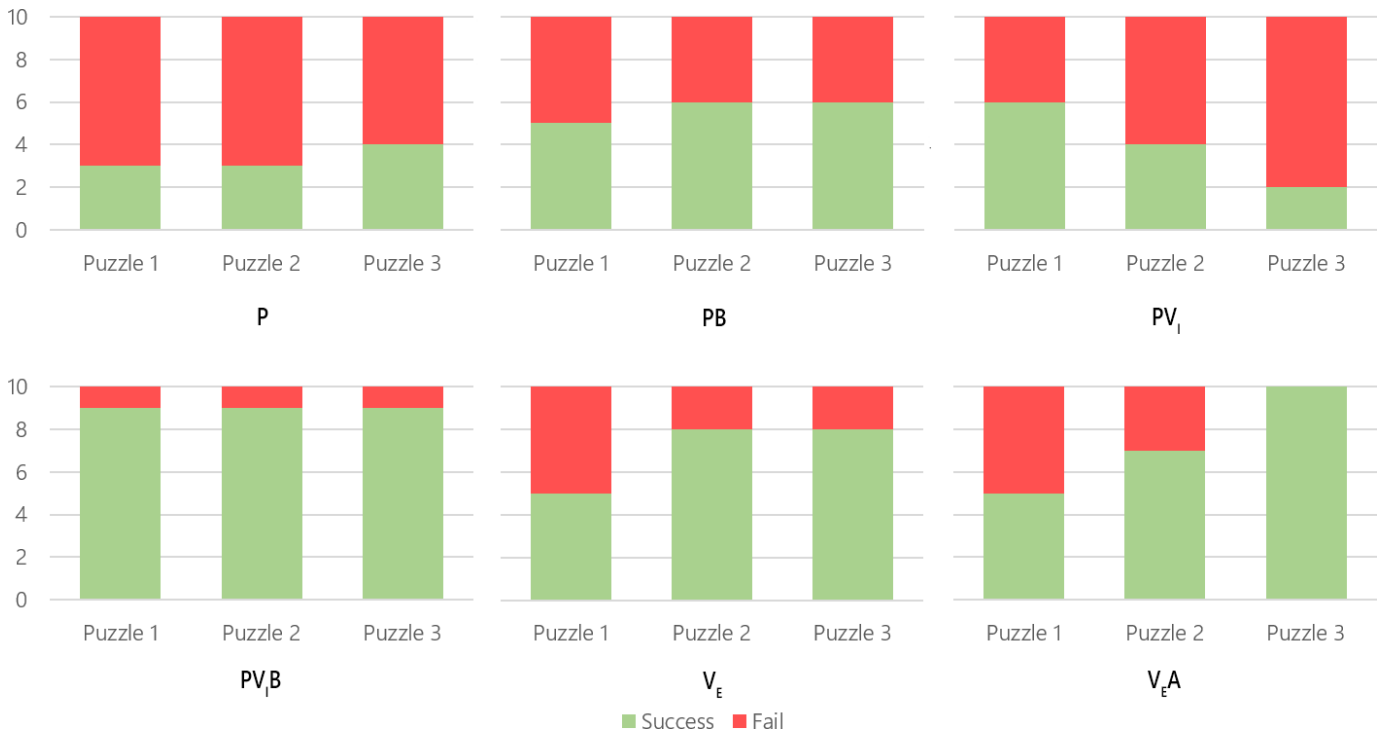

Fig. 7. Number of successful (green) and failed (red) attempts at solving the three puzzles in the immediate testing phase for each of the experimental conditions.

V_EA did not contribute to this model (Wald = .000). However, it is important to note that all participants in this condition successfully completed the third puzzle.

A binomial logistic regression was then performed to ascertain the effects of successful completion of the first puzzle on the likelihood that participants succeed at assembling the second puzzle during the immediate testing phase. The logistic regression model was statistically significant, $\chi^2(1) = 12.993, p < 0.001$. The model explained 26.5% (Nagelkerke R^2) of the variance in success rate and correctly classified 73.3% of cases. The model suggested that participants who succeeded at correctly assempling the first puzzle were 7.65 times as likely to successfully assemble the second puzzle than participants in the reference category (PV_IB).

A binomial logistic regression was also performed to ascertain the effects of successful completion of the second puzzle on the likelihood that participants succeed at assembling the third puzzle during the immediate testing phase. The logistic regression model was statistically significant, $\chi^2(1) = 15.174, p < 0.001$. The model explained 30.8%

Table 4. Test statistics using Dunn's procedure [4] for training times between the different experimental conditions. Significance values have been adjusted by the Bonferroni correction for multiple tests.

		PB	PV_I	PV_IB	V_E	V_EA
Puzzle 1	P	-23.50[a]	1.50[b]	-24.80[a]	-20.90[b]	-23.20[a]
	PB		25.00[a]	-1.30[b]	2.60[b]	0.30[b]
	PV_I			-26.30[a]	-22.40[b]	-24.70[a]
	PV_IB				-3.90[b]	-1.60[b]
	V_E					-2.30[b]
Puzzle 2	P	-17.50[b]	11.90[b]	-14.20[b]	-14.95[b]	-13.25[b]
	PB		29.40[a]	3.30[b]	2.55[b]	4.25[b]
	PV_I			-26.10[a]	-26.85[a]	-25.15[a]
	PV_IB				0.75[b]	-0.95[b]
	V_E					1.70[b]

[a] interaction is significant at the 0.05 level (two-tailed)
[b] interaction is not significant

Table 5. Test statistics using Dunn's procedure [4] for immediate testing times between the different experimental conditions. Significance values have been adjusted by the Bonferroni correction for multiple tests.

		PB	PV_I	PV_IB	V_E	V_EA
Puzzle 1	P	5.00[b]	4.25[b]	27.35[a]	6.80[b]	8.80[b]
	PB		-0.75[b]	22.35[a]	1.80[b]	3.80[b]
	PV_I			23.10[a]	2.55[b]	4.55[b]
	PV_IB				20.55[b]	18.55[b]
	V_E					2.00[b]
Puzzle 2	P	11.30[b]	4.60[b]	29.05[a]	16.90[b]	14.65[b]
	PB		-6.70[b]	17.75[b]	5.60[b]	3.35[b]
	PV_I			24.45[a]	12.30[b]	10.05[b]
	PV_IB				12.15[b]	14.40[b]
	V_E					-2.25[b]
Puzzle 3	P	11.50[b]	-2.15[b]	25.45[a]	15.70[b]	25.40[a]
	PB		-13.65[b]	13.95[b]	4.20[b]	13.90[b]
	PV_I			27.60[a]	17.85[b]	27.55[a]
	PV_IB				9.75[b]	0.50[b]
	V_E					9.70[b]

[a] interaction is significant at the 0.05 level (two-tailed)
[b] interaction is not significant

(Nagelkerke R^2) of the variance in success rate and correctly classified 76.7% of cases. The model suggested that participants who succeeded at correctly assembling the second puzzle were 9.687 times as likely to successfully assemble the third puzzle than participants in the reference category (PV_IB).

As a result, a binomial logistic regression was performed to ascertain the effects of successful completion of the first puzzle and experimental condition on the likelihood that participants succeed at assembling the second puzzle during the immediate testing phase. The logistic regression model was statistically significant, $\chi^2(1) = 22.265, p = 0.001$. The model explained 42.1% (Nagelkerke R^2) of the variance in success rate and correctly classified 75% of cases. The Wald criterion demonstrated that only none of the experimental conditions made a significant contribution to prediction. The Wald criterion also showed that successful completion of the previous puzzle did contribute significantly to prediction ($p = 0.003$). The model suggested that participants who succeeded at correctly assempling the first puzzle were 8.273 times as likely to successfully assemble the second puzzle than participants in the reference category (PV_IB). This model presented with the highest percentage of completely classified observations for the second puzzle.

A binomial logistic regression was also performed to ascertain the effects of successful completion of the second puzzle and experimental condition on the likelihood that participants succeed at assembling the third puzzle during the immediate testing phase. The logistic regression model was statistically significant, $\chi^2(1) = 32.441, p < 0.001$. The model explained 57.5% (Nagelkerke R^2) of the variance in success rate and correctly classified 83.3% of cases. The Wald criterion demonstrated that condition PV_I and successful completion of the previous puzzle made a significant contribution to prediction ($p = 0.030$ and $p = 0.007$, respectively). The model suggested that participants in the PV_I condition were 0.048 times as likely to successfully assemble the third puzzle than participants in the reference category (PV_IB). Participants who successfully completed the second puzzle were 8.475 times as likely to successfully assemble the third puzzle than participants in the reference category (PV_IB). Note that condition V_EA did not contribute to this model (Wald = .000). However, it is important to note that all participants in this condition successfully completed the third puzzle. This model presented with the highest percentage of completely classified observations for the third puzzle.

To summarise, the binomial logistic regression model for the first puzzle was not statistically significant, with only condition P significantly contributing to the model. For the second puzzle, the binomial logistic regression model with the highest percentage of correctly classified observations was the one that ascertained the effect of successful completion of the previous puzzle during immediate testing. For the third puzzle, the binomial logistic regression model with the highest percentage of correctly classified observations was the one that ascertained the effect of both experimental condition and successful completion of the previous puzzle. These results show some support for H1 and H3.

5.2.3 Immediate testing completion times

We compared puzzle completion times between the different experimental conditions during the immediate testing phase. Completion time values were corrected by assigning the upper time limit (180s) to all unsuccessful attempts (see Section 5.1). All the corrected data satisfied the assumption of homogeneity.

Boxplots with immediate testing times for each of the puzzles are shown in Figure 8. Non-parametric statistical analysis was performed for immediate testing times because our data was not normally distributed as shown by a Shapiro-Wilk test.

A Kruskal-Wallis H test showed that there was an overall statistically significant difference in time taken to assemble the first puzzle in the testing phase between the different experimental conditions, $\chi^2(5) = 16.618$, p = 0.005, with a mean rank score of 39.20 for P, 34.20 for PB, 34.95 for PV_I, 11.85 for PV_IB, 32.40 for V_E and 30.40 for V_EA.

A Kruskal-Wallis H test showed that there was an overall statistically significant difference in time taken to assemble the second puzzle in the testing phase between the different experimental conditions, $\chi^2(5) = 17.986, p = 0.003$, with a mean rank score of 43.25 for P, 31.95 for PB, 38.65 for PV_I, 14.20 for PV_IB, 26.35 for V_E and 28.60 for V_EA.

A Kruskal-Wallis H test showed that there was an overall statistically significant difference in time taken to assemble the third puzzle in the testing phase between the different experimental conditions, $\chi^2(5) = 24.536, p < 0.001$, with a mean rank score of 43.15 for P, 31.65 for PB, 45.30 for PV_I, 17.70 for PV_IB, 27.45 for V_E and 17.75 for V_EA.

Pairwise comparisons were performed using Dunn's procedure [4] with a Bonferroni correction for multiple comparisons with adjusted p-values. These are displayed in Table 5. Note that pairwise comparisons for puzzles in which the Kruskal-Wallis H test showed no overall statistically significant difference have not been included.

The post hoc analysis revealed statistically significant differences in immediate testing times for the first puzzle. There was a statistically significant difference between PV_IB (mean rank = 11.85) and P (mean rank = 39.20) ($p = 0.004$), PB (mean rank = 34.20) ($p = 0.003$) and PV_I (mean rank = 34.95) ($p = 0.002$).

The post hoc analysis revealed statistically significant differences in immediate testing times for the second puzzle. There was a statistically significant difference between PV_IB (mean rank = 14.20) and P (mean rank = 43.25) ($p = 0.002$) and PV_I (mean rank = 38.65) ($p = 0.019$).

The post hoc analysis revealed statistically significant differences in immediate testing times for the third puzzle. There was a statistically

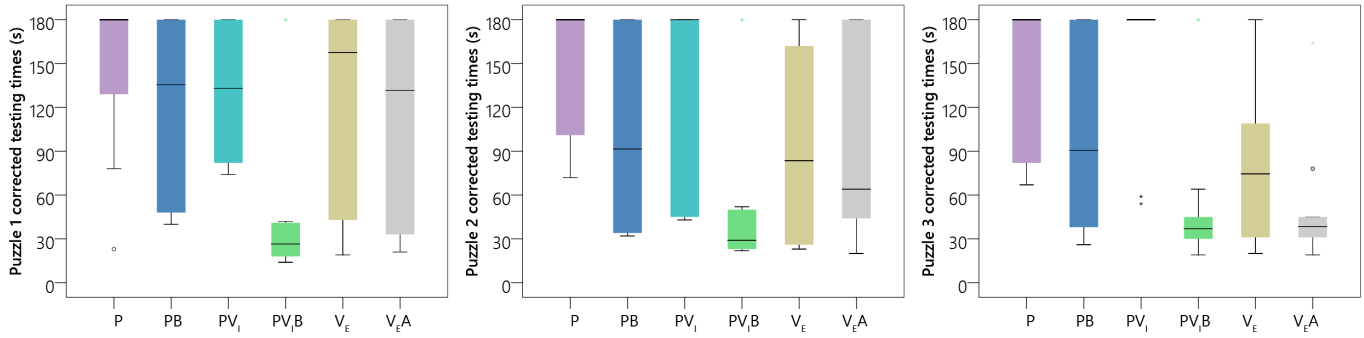

Fig. 8. Boxplot containing corrected immediate testing times for each of the puzzles. Medians are shown as dark horizontal lines. Boxes represent the interquartile ranges (IQR). Whiskers represent either the extreme data points or extend to 1.5 x IQR. Outliers (data points outside the whiskers) are shown by circles. A value, X, is an outlier if X $<$ lower quartile $-$ 1.5 x interquartile range or if X $<$ upper quartile $+$ 1.5 x IQR. See Table 5 for pairwise interactions.

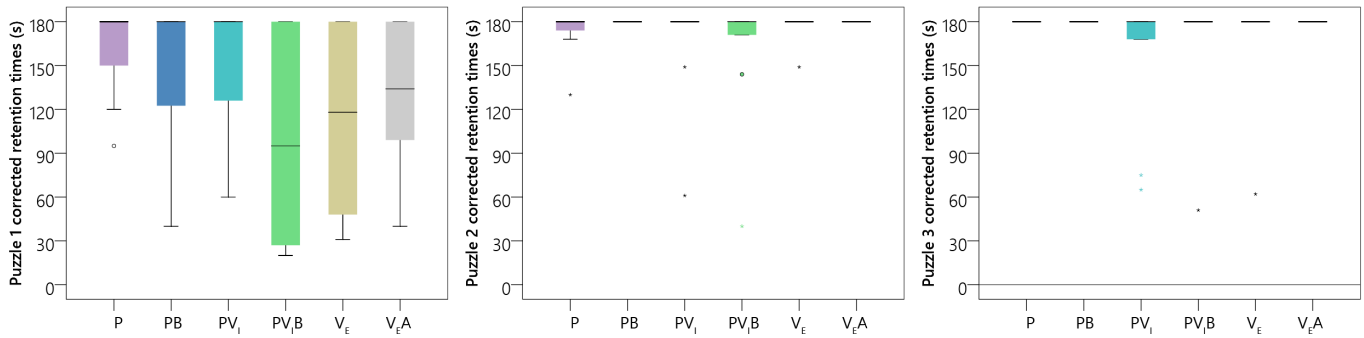

Fig. 9. Boxplot containing corrected retention testing times for each of the puzzles. Medians are shown as dark horizontal lines. Boxes represent the interquartile ranges (IQR). Whiskers represent either the extreme data points or extend to 1.5 x IQR. Outliers (data points outside the whiskers) are shown by circles. A value, X, is an outlier if X $<$ lower quartile $-$ 1.5 x interquartile range or if X $<$ upper quartile $+$ 1.5 x IQR.

significant difference between PV_IB (mean rank = 17.70) and P (mean rank = 43.15) ($p = 0.013$) and PV_I (mean rank = 45.30) ($p = 0.005$). There was a statistically significant difference between V_EA (mean rank = 17.75) and P (mean rank = 43.15) ($p = 0.013$) and PV_I (mean rank = 45.30) ($p = 0.005$).

The analysis of immediate testing completion times shows some support for H2 and H3.

5.2.4 Subjective questionnaire ratings

There was no statistically significant difference in rated difficulty, ease of use and seriousness between groups as determined by one-way ANOVA for the first puzzle.

There was a statistically significant difference in ease of use of the training environment (F(5,54) = 5.006, p = 0.001) between groups as determined by one-way ANOVA for the second puzzle. A Tukey post hoc test revealed that participants in the P condition (M = 2.5, SD = 1.08) rated the ease of use of the training environment as significantly more difficult than participants in the V_E (M = 4.4, SD = 0.70, p = 0.001) and V_EA (M = 4.1, SD = 1.1, p = 0.007) conditions. No other significant interactions were found for the second puzzle.

There was a statistically significant difference in task difficulty (F(5,54) =4.613, p = 0.001) between groups as determined by one-way ANOVA for the third puzzle. A Tukey post hoc test revealed that participants in the P condition (M = 1.9, SD = 1.00) rated the difficulty of the task as significantly more difficult than participants in the V_EA (M = 4.1, SD = 0.88, p = 0.002) condition. Participants in the PB condition (M = 2.7, SD = 1.34) also rated the difficulty of the task as significantly more difficult than participants in the VEA condition (M = 4.1, SD = 1.34, p = 0.003). No other significant interactions were found for the third puzzle.

There was a statistically significant difference in ease of use of the training environment (F(5,54) = 3.044, p = 0.017) between groups as determined by one-way ANOVA for the third puzzle. A Tukey post hoc test revealed that participants in the P condition (M = 2.4, SD = 1.35) rated the ease of use of the training environment as significantly more difficult than participants in the V_EA (M = 4.4, SD = 0.70, p = 0.007) condition. No other significant interactions were found for the third puzzle.

5.3 Second session

5.3.1 Participants

A total of 56 participants that completed the first part session returned to complete the second session two weeks later (average number of days between training session and retention session: 14.16, SD = 0.918). Overall, retention testing performance was lower than expected for all conditions both in terms of success rates and completion times. We believe this is due to the high complexity of the 3D puzzles.

5.3.2 Retention testing success rates

A binomial logistic regression was performed to ascertain the effects of experimental condition on the likelihood that participants succeed at assembling each puzzle during the immediate testing phase. PV_IB was chosen as the reference category (the condition with most successful puzzle completions, overall).

The binomial logistic regression model was not statistically significant, $\chi^2(5) = 6.240, p = 0.284$ for the first puzzle. The model explained 14.3% (Nagelkerke R^2) of the variance in success rate and correctly classified 65.5% of cases. The Wald criterion demonstrated that none of the conditions made a significant contribution to prediction.

Table 6. Number of successful attempts, failed attempts and participants solving the three puzzles in the retention testing phase for each of the experimental conditions.

		P	PB	PV_II	PV_IB	V_E	V_EA
Puzzle 1	Success	2	3	3	6	7	5
	Fail	6	6	7	4	3	4
Puzzle 2	Success	2	0	2	4	1	0
	Fail	6	9	8	6	9	9
Puzzle 3	Success	0	0	3	1	1	0
	Fail	8	9	7	9	9	9
	N	8	9	10	10	10	9

The binomial logistic regression model was not statistically significant, $\chi^2(5) = 10.054, p = 0.074$ for the second puzzle. The model explained 28.3% (Nagelkerke R^2) of the variance in success rate and correctly classified 83.6% of cases. The Wald criterion demonstrated that none of the conditions made a significant contribution to prediction.

The binomial logistic regression model was not statistically significant, $\chi^2(5) = 8.289, p = 0.141$ for the third puzzle. The model explained 30.7% (Nagelkerke R^2) of the variance in success rate and correctly classified 90.9% of cases. The Wald criterion demonstrated that none of the conditions made a significant contribution to prediction.

5.3.3 Retention testing completion times

We compared puzzle retention testing times between the different experimental conditions. Completion time values were corrected by assigning the upper time limit (180s) to all unsuccessful attempts. All the corrected data satisfied the assumption of homogeneity.

Boxplots with training times for each of the puzzles are shown in Figure 9. Non-parametric statistical analysis was performed for retention testing times because our data was not normally distributed as shown by a Shapiro-Wilk test.

A Kruskal-Wallis H test showed that there was no overall statistically significant difference in time taken to assemble the first puzzle in the retention testing phase between the different experimental conditions, $\chi^2(5) = 8.101$, p = 0.151, with a mean rank score of 34.69 for P, 32.88 for PB, 33.45 for PV_I, 20.90 for PV_IB, 21.95 for V_E and 26.28 for V_EA.

A Kruskal-Wallis H test showed that there was no overall statistically significant difference in time taken to assemble the second puzzle in the retention testing phase between the different experimental conditions, $\chi^2(5) = 5.832$, p = 0.323, with a mean rank score of 25.25 for P, 32.00 for PB, 26.35 for PV_I, 23.70 for PV_IB, 29.35 for V_E and 32.00 for V_EA.

A Kruskal-Wallis H test showed that there was no overall statistically significant difference in time taken to assemble the third puzzle in the retention testing phase between the different experimental conditions, $\chi^2(5) = 7.151$, p = 0.210, with a mean rank score of 30.50 for P, 30.50 for PB, 22.55 for PV_I, 27.55 for PV_IB, 27.65 for V_E and 30.50 for V_EA.

5.3.4 Subjective questionnaire ratings

There was no statistically significant difference in rated difficulty and seriousness between groups as determined by one-way ANOVA for any of the three puzzles. Tukey post hoc tests showed no significant interactions.

6 Discussion

In terms of training times, post hoc analysis revealed a significant difference between the physical conditions where no blocks were available during training (P and PV_I) and the rest of physical conditions where blocks were available during training (PB and PV_IB), amongst other significant interactions. For the second puzzle, we observe a significant difference in training times between PV_I and all other conditions except P, amongst other effects. For the third puzzle we found no significant interactions. We believe it is important to note the lack of significant differences in terms of training times between the virtual conditions and condition PV_IB, the overall best performing condition. We also believe that the lower training times for conditions P and PB could be due to the lack of blocks to practice with during training, which meant participants did not have any activities to perform during training and therefore decided to move on to the next stage of the experimental task. This could be related to a high number of unsuccessful puzzle completions in these conditions. An increase in training times for these conditions in puzzles 2 and 3 could be due to participants understanding the complexity of the tasks after the immediate testing for the first puzzle and deciding to spend more time inspecting the paper instructions and video (when available).

Regarding success rates for immediate testing, we observed that condition PV_IB yielded the highest number of successful completions of the three puzzles (see Figure 7). We also observed that condition P yielded the lowest number of successful completions of the puzzles during immediate testing. Condition PB showed a ceiling effect in the second and third puzzle. Successful puzzle completions in condition PV_I decreased with each puzzle. Immediate testing success rates for the virtual conditions, V_E and V_EA, increased with each puzzle. Our analysis showed that the binomial logistic regression model for the first puzzle was not statistically significant, with only condition P significantly contributing to the model. For the second puzzle, the binomial logistic regression model with the highest percentage of correctly classified observations was the one that ascertained the effect of the successful completion of the previous puzzle during immediate testing. For the third puzzle, the binomial logistic regression model with the highest percentage of correctly classified observations was ones that ascertained the effect of experimental condition as well as the successful completion of the previous puzzle during immediate testing.

In terms of immediate testing completion times, we observed how (parallel to an increase in success rate) the immediate testing times for condition V_EA decreased with each puzzle. A statistically significant difference was not found between this condition and condition PV_IB (the condition with overall lowest testing times). This result could indicate that the availability of static and animated instructions in the virtual training environment contributed to effective training.

Anecdotal evidence from the training videos as well as participant feedback during debrief interviews shows that virtually trained participants initially struggled to assemble the pieces during the immediate testing phase. We believe this is due to the lack of experience in handling and joining the physical blocks during training. However, after the first and second tasks, participants refined their strategy during the training stage to include physically plausible movements of the puzzles pieces. This is, participants replicated the movement they would then perform with the physical blocks in the virtual training environment and avoided allowing the pieces to go through each other, as no physics restrictions were assigned to the virtual blocks in the virtual training environment.

Subjective questionnaire ratings answered by participants during the first session showed no statistically significant difference in rated difficulty, ease of use and seriousness between groups as determined by one-way ANOVA for the first puzzle. For the second puzzle, results indicated that participants in the P condition rated the training environment as significantly more difficult to use than participants in the V_E and V_EA conditions. When asked about task difficulty in the third puzzle, participants in the P and PB conditions rated the difficulty of the task as significanly more difficult than participants in the V_EA condition. In terms of ease of use of the training enviornment, participants in the P condition rated the ease of use of the training enviornment as significantly more difficult than participants in the V_EA condition.

One of the limitations in our design was the high complexity of the puzzles. Overall, retention testing resulted in lower performance than we had expected and we believe this is due to the difficulty associated with remembering the process to solve the three puzzles two weeks after the training. This was further validated by verbal feedback from our participants during the second session. Our previous piloting of the task had not shown this effect. Future studies should further evaluate

the suitability of the task for retention. This evaluation should aim to balance the amount of training and complexity of the task to avoid floor and ceiling effects in subsequent retention sessions.

7 Conclusion

In this paper we have presented a study that compares the effectiveness of virtual training and physical training for learning transfer of a bimanual assembly task. Our study extends previous research by comparing two virtual and four physical training formats which represent common training formats for assembly processes. We aimed to verify if exposure to a virtual training environment is sufficient for effective training when haptic devices are not available and when the physical components and tools used in the assembly process are not accessible.

Following a between-subjects experimental design, participants were trained to assemble three versions of a 3D burr puzzle in one of six experimental conditions (see Table 1 for definitions). 3D models of the burr puzzles used in the study are available to download at `https://vr.cs.ucl.ac.uk/research/virtual-training`. All participants completed an immediate testing phase and a retention test two weeks after the training, both with physical versions of the puzzles. Participants were trained and tested in solving the puzzles in increasing order of complexity in that they had to remember a higher number of steps in the assembly process for each of the puzzles.

We analysed performance in terms of success rates as well as immediate testing times and retention testing times. Our results show that the performance of virtually trained participants was promising. A statistically significant difference wasn't found between condition V_EA and the best performing physical condition (PV_IB, in which physical blocks and animated instructions were available during training) for the last and most complex puzzle in terms of success rates and immediate testing times. We believe these results are of great importance given that virtually trained participants did not have the chance to interact with the physical blocks at any point during training. We also observed that participants were more likely to successfully assemble a puzzle during immediate testing if they had successfully assembled the previous one. Retention testing performance was unexpectedly low due to the high complexity of the task. We believe that the results of this study further validate the effectiveness of virtual training for bimanual assembly tasks.

References

[1] R. J. Adams, D. Klowden, and B. Hannaford. Virtual Training for a Manual Assembly Task. *Haptics-e, The Electronic Journal of Haptics Research*, 2(2), Oct. 2001.

[2] G. M. Bodner and R. B. Guay. The Purdue Visualization of Rotations Test. *The Chemical Educator*, 2(4):1–17, Oct. 1997. doi: 10.1007/s00897970138a

[3] P. Carlson, A. Peters, S. B. Gilbert, J. M. Vance, and A. Luse. Virtual Training: Learning Transfer of Assembly Tasks. *IEEE Transactions on Visualization and Computer Graphics*, 21(6):770–782, Jan. 2015. doi: 10.1109/TVCG.2015.2393871

[4] O. J. Dunn. Multiple comparisons using rank sums. *Technometrics*, 6(3):241–252, 1964.

[5] N. Gavish, T. Gutiérrez, S. Webel, J. Rodríguez, M. Peveri, U. Bockholt, and F. Tecchia. Evaluating Virtual Reality and Augmented Reality Training for Industrial Maintenance and Assembly Tasks. *Interactive Learning Environments*, 23(6):778–798, July 2015. doi: 10.1080/10494820.2013.815221

[6] M. Gonzalez-Franco, R. Pizarro, J. Cermeron, K. Li, J. Thorn, W. Hutabarat, A. Tiwari, and P. Bermell-Garcia. Immersive Mixed Reality for Manufacturing Training. *Frontiers in Robotics and AI*, 4(3):1, Feb. 2017. doi: 10.3389/frobt.2017.00003

[7] C. R. Hall and C. D. Horwitz. Virtual Reality for Training: Evaluating Retention of Procedural Knowledge. *International Journal of Virtual Reality*, 5(1):61–70, Nov. 2015.

[8] M. Murcia-López and A. Steed. The Effect of Environmental Features, Self-avatar and Immersion on Object Location Memory in Virtual Environments. *Frontiers in ICT*, 3:24, Nov. 2016. doi: 10.3389/fict.2016.00024

[9] Oculus VR, LLC. Oculus Avatar SDK 1.15.0. `https://developer.oculus.com/downloads/package/oculus-avatar-sdk/1.14.0/`, 2017.

[10] M. Oren, P. Carlson, S. Gilbert, and J. M. Vance. Puzzle Assembly Training: Real World vs. Virtual Environment. In *Proceedings of the IEEE Virtual Reality Conference*, pp. 27–30. IEEE, 2012. doi: 10.1109/VR.2012.6180873

[11] M. J. Piller and M. M. Sebrechts. Spatial Learning in Transparent Virtual Environments. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 47, pp. 2133–2136. SAGE Publications Sage CA: Los Angeles, CA, 2003.

[12] M. Prachyabrued and C. W. Borst. Visual Feedback for Virtual Grasping. In *IEEE Symposium on 3DUI*, pp. 19–26. IEEE, April 2014. doi: 10.1109/3DUI.2014.6798835

[13] F. Ritter, T. Strothotte, O. Deussen, and B. Preim. Virtual 3D Puzzles: A New Method for Exploring Geometric Models in VR. *IEEE Computer Graphics and Applications*, 21(5):11–13, 2001. doi: 10.1109/38.946625

[14] A. A. Rizzo and J. G. Buckwalter. Virtual Reality and Cognitive Assessment and Rehabilitation: The State of The Art. *Virtual Reality in Neuro-Psycho-Physiology: Cognitive, Clinical and Methodological Issues in Assessment and Rehabilitation*, 44:123, 1997.

[15] A. Roever. Burr Tools 0.6.3. `http://burrtools.sourceforge.net/`, 2015.

[16] F. Rose, E. A. Attree, B. Brooks, D. Parslow, and P. Penn. Training in Virtual Environments: Transfer to Real World Tasks and Equivalence to Real Task Training. *Ergonomics*, 43(4):494–511, Nov. 2010. doi: 10.1080/001401300184378

[17] D. Schmorrow, J. Cohn, and D. Nicholson. *The PSI Handbook of Virtual Environments for Training and Education [Three Volumes]: Developments for the Military and Beyond*. Praeger Publishers, 2008.

[18] A. Seth, J. M. Vance, and J. H. Oliver. Combining Dynamic Modeling with Geometric Constraint Management to Support Low Clearance Virtual Manual Assembly. *Journal of Mechanical Design*, 132(8):081002, July 2010. doi: 10.1115/1.4001565

[19] D. Shuralyov and W. Stuerzlinger. A 3d Desktop Puzzle Assembly System. In *IEEE Symposium on 3DUI*, pp. 139–140. IEEE, March 2011. doi: 10.1109/3DUI.2011.5759244

[20] A. Sowndararajan, R. Wang, and D. A. Bowman. Quantifying the Benefits of Immersion for Procedural Training. In *Proceedings of the workshop on Immersive projection technologies/Emerging display technologiges*, p. 2. ACM, Aug. 2008. doi: 10.1145/1394669.1394672

[21] K. M. Stanney, R. R. Mourant, and R. S. Kennedy. Human Factors Issues in Virtual Environments: A Review of The Literature. *Presence: Teleoperators and Virtual Environments*, 7(4):327–351, March 1998. doi: 10.1162/105474698565767

[22] S. G. Vandenberg and A. R. Kuse. Mental Rotations, a Group Test of Three-Dimensional Spatial Visualization. *Perceptual and motor skills*, 47(2):599–604, Dec. 1978. doi: 10.2466/pms.1978.47.2.599

[23] D. Waller, E. Hunt, and D. Knapp. The Transfer of Spatial Knowledge in Virtual Environment Training. *Presence: Teleoperators and Virtual Environments*, 7(2):129–143, March 1998. doi: 10.1162/105474698565631