Evaluating Multiple Levels of an Interaction Fidelity Continuum on Performance and Learning in Near-Field Training Simulations

Ayush Bhargava, Jeffrey W. Bertrand, Anand K. Gramopadhye, Kapil C. Madathil and Sabarish V. Babu

Abstract—With costs of head-mounted displays (HMDs) and tracking technology decreasing rapidly, various virtual reality applications are being widely adopted for education and training. Hardware advancements have enabled replication of real-world interactions in virtual environments to a large extent, paving the way for commercial grade applications that provide a safe and risk-free training environment at a fraction of the cost. But this also mandates the need to develop more intrinsic interaction techniques and to empirically evaluate them in a more comprehensive manner. Although there exists a body of previous research that examines the benefits of selected levels of interaction fidelity on performance, few studies have investigated the constituent components of fidelity in a Interaction Fidelity Continuum (IFC) with several system instances and their respective effects on performance and learning in the context of a real-world skills training application. Our work describes a large between-subjects investigation conducted over several years that utilizes bimanual interaction metaphors at six discrete levels of interaction fidelity to teach basic precision metrology concepts in a near-field spatial interaction task in VR. A combined analysis performed on the data compares and contrasts the six different conditions and their overall effects on performance and learning outcomes, eliciting patterns in the results between the discrete application points on the IFC. With respect to some performance variables, results indicate that simpler restrictive interaction metaphors and highest fidelity metaphors perform better than medium fidelity interaction metaphors. In light of these results, a set of general guidelines are created for developers of spatial interaction metaphors in immersive virtual environments for precise fine-motor skills training simulations.

Index Terms—Bimanual Interaction, Interaction Fidelity, Empirical Evaluation, Educational Virtual Reality

1 Introduction

One of the key considerations when designing a VR training application is to determine to what extent interaction fidelity facilitates the objectives of the simulation. Interaction fidelity refers to the objective degree of exactness with which real-world interactions are reproduced in the system [22]. It has been assumed by researchers that high-fidelity interactions would be the most beneficial for virtual training due to resemblance to the real world [32]. Developing high-fidelity simulations often requires more development time and higher costs. However, low-fidelity simulations have been shown to be just as effective, if not more so, in several medical training simulations [28]. Thus, there exists a need for further research into the effects of various levels of interaction fidelity on learning outcomes and user performance.

In order to accurately compare interaction metaphors, researchers require an objective method for classifying levels of fidelity. McMahan developed the Framework for Interaction Fidelity Analysis (FIFA) [22] that classifies characteristics of interaction fidelity into a standardized taxonomy. This allows researchers to systematically compare interaction metaphors without making subjective assumptions. After reexamining prior studies in the context of FIFA, McMahan et al. found that in some cases, the mid-fidelity interaction metaphors performed worse than the low and high fidelity metaphors. Therefore, it cannot be assumed that as interaction fidelity increases, user performance will necessarily increase. This makes it difficult for developers to determine what level of fidelity their simulation should attain.

Near-field, bimanual or two-handed virtual training are becoming economically viable for the first time. Commercial off-the-shelf devices such as the Razer Hydra and the Oculus Rift head-mounted display (HMD) enable high-fidelity interaction that can be leveraged for training. With widespread adoption comes the need for evaluating the benefits and drawbacks for using such devices. Previous research into bimanual interaction metaphors have revealed several advantages, potentially due to the fact that people are naturally accustomed to performing actions using both hands [17]. Bimanual devices afford new opportunities for fine-motor dexterous training tasks that traditional input devices such as the mouse and keyboard cannot replicate.

While much research has been conducted on comparing low versus high fidelity simulations [28], few studies have examined the constituent components of interaction fidelity and their effects on learning outcomes and user performance [27]. Even fewer studies have examined a range of interaction fidelity levels on virtual training simulations. In this work, we investigate six distinct levels of interaction fidelity to teach technical college students basic metrology concepts. We examine measures including cognitive and psychomotor learning outcomes, accuracy, efficiency, and subjective response. Each level was categorized into the biomechanical, dimensional, and transfer function symmetry aspects of the FIFA framework. Additionally, physics fidelity with regard to the presence of gravity was evaluated for determining usability enhancements. Interaction levels were ranked onto an Interaction Fidelity Continuum (IFC) from lowest to highest fidelity based on the FIFA framework aspects to investigate relative differences. Results indicate that the mid-fidelity metaphors performed worse than the high and low fidelity metaphors, but not in all cases. However, cognitive learning outcomes were not significantly different between conditions. Based on the results, general guidelines for VR developers, researchers, and educators with respect to near-field training tasks are presented.

• *Ayush Bhargava, Clemson University E-mail: ayushb@clemson.edu*
• *Jeffrey W. Bertrand, Clemson University E-mail: jbertra@clemson.edu*
• *Anand K. Gramopadhye, Clemson University E-mail: agramop@clemson.edu*
• *Kapil C. Madathil, Clemson University E-mail: kmadath@clemson.edu*
• *Sabarish V. Babu, Clemson University E-mail: sbabu@clemson.edu*

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.2018.2794639

2 Related Work

2.1 Interaction Fidelity Analysis

Interaction fidelity is one of the main factors for determining the effectiveness of virtual simulations. McMahan developed the FIFA framework to systematically categorize aspects of interaction fidelity into logical categories [22]. Three main components of interaction fidelity include biomechanical symmetry, control symmetry, and system appropriateness. Control symmetry is the amount correspondence between the control provided by an interaction technique and control possible in the real world. After reexamining prior studies, McMahan et al. found that control symmetry was the most significant contributor to task performance [22]. A sub-factor of control symmetry is transfer function symmetry or the degree of exactness with which a real-world transfer function is reproduced through interaction. The researchers evaluated three levels of interaction fidelity in a virtual pointing task. A high-fidelity wand technique was compared to a low-fidelity mouse technique in a CAVE. However, participants were losing track of their mouse crosshair, so they developed a mid-fidelity technique called the e-mouse that attached a carrier to the participant so that they could rotate their body and move the mouse relative to their physical pointing direction. Results indicated that the mid-fidelity e-mouse technique performed worse than both the low and high fidelity techniques. The transfer function symmetry was determined to be the most significant contributor to the performance results, suggesting an important area for further research.

Nabiyouni et al. also evaluated various levels of interaction fidelity using the FIFA framework [27]. The researchers compared a low-fidelity gamepad for a locomotion task to a mid-fidelity Virtusphere, a large, hollow apparatus that allows users in an HMD to walk in any direction. Both techniques were compared to the high-fidelity natural walking technique. Results indicate that performance was worse in terms of speed, precision, control, and fatigue levels in the mid-fidelity technique as compared to the low and high end. The results were consistent with prior work by McMahan [22]. Therefore, increasing interaction fidelity may not always result in improved performance outcomes [23].

2.2 Interaction Metaphors

Widespread availability of 6-DOF 3D interaction devices allows for reproduction of real-world fine-motor tasks in virtual environments, resulting in increased fidelity. The recent influx of more intuitive interaction devices like the Razer Hydra and the Oculus Touch have made it easier to design 3D interaction metaphors, but finding the right balance between usability, force expelled, and performance is challenging. To better understand and evaluate these metrics, researchers have segregated and categorized substages of an interaction metaphor based on the workspace, types of manipulation methods, and selection mechanisms [7, 22].

The identification of different components of the interaction metaphor and its evaluation is heavily influenced by the functionality implemented and the ability of the device to be used in distinct virtual spaces. The interactive virtual space, as defined in the literature, can be categorized into motor space and visual space [2]. Motor space refers to the physical volume the user can interact in and virtual space refers to the visual representation of the environment. It has been demonstrated that superimposing the two spaces, often referred to as 1:1 Control/Display or C/D ratio [2, 9], results in more natural interactions and users exert less mental load as they do not have to perform mental transformations. The control space, defined as the virtual area the user can control objects in, is often impacted by this mapping. This may require special attention as control space may fall outside the user's motor space depending on task requirements and the end-effector location. Methods like the Go-Go technique can be used to expand the control space beyond a user's physical reach, but at the expense of visual/motor mismatch [29]. Acknowledging the significance of C/D ratio and visuo-motor synchrony, our work examines how fine-motor performance metrics may be affected by it.

Virtual simulations utilizing bimanual interaction have shown positive advantages in terms of user performance and preference. Schultheis et al. evaluated a bimanual and mouse-based interaction technique for a 3D docking task and found that time to complete and user preference was significantly better in the bimanual condition [30]. All human interactions involve direct or indirect use of both hands. Writing was once considered a unimanual task, but the non-dominant hand is often used to handle the paper which actually makes the procedure 20% faster [12]. It has been observed that using both hands to perform tasks in parallel improves accuracy and efficiency along with grounding the user in the interaction space to reduce disorientation effects [8, 15]. One of the principles in Guiard's Kinematic Chain (KC) model states that the non-dominant hand sets a dynamic frame of reference for the dominant hand [12] and is often cited by VR researchers to aid the design of bimanual interaction metaphors in virtual environments. The benefits of using both hands for interaction have often been highlighted in the literature. For example, a two-handed metaphor outperforming one-handed metaphor after minimal training [11], two-handed interaction technique outperforming mouse and wand-based interactions for construction tasks [30], and bimanual task improvements becoming more pronounced over unimanual tasks as cognitive demand increases [19]. Therefore, research shows that bimanual interaction metaphors can have significant advantages over unimanual interaction.

2.3 Virtual Training and Skill Transfer

Transfer of skills from training simulations to the real world is one of the most essential aspects justifying the development and use of complex VR-based educational simulations. Past work has shown that VR can help users learn higher level cognitive skills which are difficult to acquire through didactic or passive lecture-style delivery of facts [31]. Other advantages include instantaneous personalized feedback [10], improved motivation [25], and active learning. Researchers have also shown success in teaching cognitive and psychomotor skills related to assembly tasks [13, 26]. Psychomotor skills transfer has been studied in the context of surgery simulations and seen positive results [20].

Metrology is the scientific study of measurement and is a foundation for many engineering domains. It often involves skilled use of tools like the Vernier calipers and micrometers. The demand for knowledge and skills for such instruments exceeds the education supply in the engineering fields as of 2013 [18]. Al-Zahrani has determined that online and virtual training simulations are valuable assets for metrology training [1]. Their group developed JAVA applets that teach students how to use and read such equipment with minimal interactivity. Since metrology has enough cognitive and dexterous fine-motor complexity to allow studying learning and performance outcomes, we further build on this concept by incorporating real-time physics, bimanual interaction, and a scaffolded learning environment. The VR training simulation developed is highly innovative and helps offload elementary concepts such as learning how to take precise measurements using calipers and micrometers.

3 System Description

The virtual metrology training simulation has followed an evolutionary design over the course of several years and two user studies to examine the effects of various levels of interaction fidelity on task performance and learning outcomes [4, 5]. Prior to the studies, the initial application was browser-based and users interacted with virtual instruments including calipers and micrometers by clicking and dragging them to the correct position to take a measurement. Interacting using a mouse does not simulate how one would physically take measurements in the real world and exploring new interaction metaphors was of interest. VR devices such as the Razer Hydra allow for 6-DOF interaction in near-field space that enable spatial interaction analogous to taking measurements in the real world. The Hydra afforded a wide range of potential interaction techniques and several were developed and studied to determine the advantages of various levels of fidelity on task performance and learning outcomes.

The simulation for both the studies was designed in a scaffolded learning approach where guidance was reduced as the student became more proficient in the skill. It began with an introduction phase that showed the user the different parts of each instrument and how to interpret the Vernier scale. A guided practice phase took the user step-by-step through the process of taking a measurement and would not allow them to continue until they performed the correct action. For instance, one instruction asked the participant to clamp the jaws around the measurement object and would not allow them to proceed until both jaws were fully in contact with the object. Users clamped the jaws by tilting the thumbstick on the Hydra to the left or right after they initiated a grasping action by pulling the trigger on the back of the Hydra. Their task was to take measurements of simplified abstract objects such as cylinders and cubes. Additional functionality such as locking the Vernier scale in place was implemented via button presses. Users input their measurements via a numberpad in the first experiment and a radial

dial in the second experiment. The radial dial was implemented in order to improve interaction in the HMD since physically moving the Hydra controller to the numbers was cumbersome.

Next, users would repeat what they learned in an open exercise phase with multiple trials. No guidance was provided to the user and they had to demonstrate that they learned the task from the guided practice phase by taking measurements of real-world objects including bearings, gears, and pistons. The scale of the measurement objects were varied between trials. The system accepted their measurement if it was within a range of ± 1 mm from the ground truth. The user was asked to try again if the answer was incorrect. Tolerance ranges are commonly used in industrial settings and were adopted here to account for the precision of the physics engine. The metrology simulation was divided into 6 modules comprising of the inside, outside, and depth functions of the Vernier caliper and the inside, outside, and depth micrometers.

Presentation method also affects how one would interact with the system. In conditions using a large-screen immersive display (LSID), we incorporated head-tracking using an Ascension Flock of Birds tracker, perspective correction, and stereoscopy via Nvidia's 3D Vision active stereo glasses in order to enhance depth perception. The camera view in the first study was a top-down view of the workbench looking into the screen similar to how the web-based version was implemented. Utilizing an HMD affords the ability to have the virtual end-effectors co-located with their physical hands, which could improve performance and preference. However, confounds between the two presentation methods exist including different resolutions, FOV, and orientation tracking which could affect performance and would need to be studied independently to determine to what extent performance is affected. In this contribution, in the interest of ecological validity, we have examined how native properties of viewing and perceptual-motor affordances of the different interaction fidelity conditions have an effect on performance and learning of fine motor tasks in a near-field virtual reality simulation.

In this work, six levels of interaction fidelity were examined across two user study interventions and each level represented an application point on the interaction fidelity continuum using McMahan's FIFA framework [22]. The conditions sorted in order from lowest to highest interaction fidelity include: L-3DF, L-6DF, M-LSID-NG, M-LSID-G, H-HMD-NG, H-HMD-G explained below (see Table 1). The first study compared the L-3DF and L-6DF conditions and the second study compared the conditions M-LSID-NG, M-LSID-G, H-HMD-NG and H-HMD-G. The studies were conducted in a between-subjects manner in which participants were randomly assigned to one of the interaction fidelity conditions.

3.1 L-3DF Condition

Our first user study varied interaction fidelity based on the dimensional symmetry aspect of McMahan's FIFA framework. Dimensional symmetry is the amount of degrees-of-freedom afforded in the simulation compared to the same actions in the real world. For instance, the user could reach out and grab a virtual instrument and move them along 3 positional and 3 rotational axes. After some pilot testing, it was determined that simplifying the interaction by reducing the number of degrees-of-freedom could potentially aid user interaction. The drawback was that reducing the degrees-of-freedom would not simulate how users would interact with the objects in the real world. The user study evaluated a simplified 3-DOF interaction metaphor compared to a full 6-DOF metaphor [5]. The 3-DOF condition allowed for two positional axes about X and Y and one rotational axis about the Z-axis. There is an inherent proprioceptive mismatch in the 3-DOF condition due to the reduced number of degrees-of-freedom. The camera view was directly over the workbench looking down at the instruments similar to the browser-based version (see Figure 1). Compared to each of the conditions in both studies, the 3-DOF condition was regarded as having the lowest interaction fidelity out of all six levels and is labeled L-3DF.

3.2 L-6DF Condition

The advantage of enabling 6-DOF interaction was that actions performed would simulate how one would perform the actions in the real world. However, one of the issues with the layout of the simulation was that participants in the 6-DOF condition seemed unwilling to maneuver the objects away from the workbench after grasping the objects. They would hover closely over the workbench and did not take full advantage of the interaction space. When they rotated an instrument, it would collide with the workbench. The 3-DOF condition did not have this issue as it restricted movement in the XY plane and allowed rotation only about the axis pointing into the screen. In order to mitigate this problem, the next iteration of the metrology simulation had a forward-facing camera with the gravity vector pointing down relative to the user.

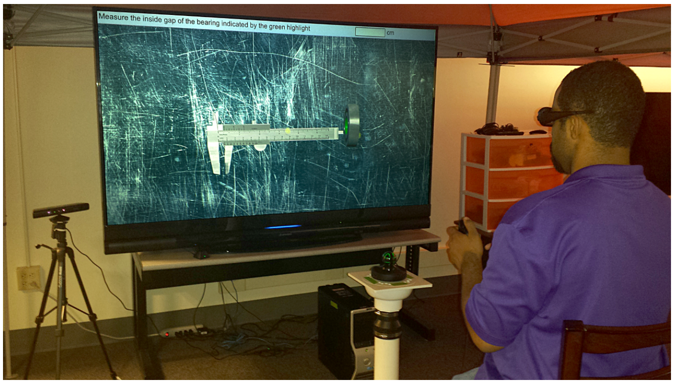

Fig. 1. Experimental apparatus for the first study featuring the L-3DF and L-6DF conditions

Another issue that participants in the 6-DOF condition had was that after they initiated a grasping action, they realized it would be difficult to take the measurement from the angle that they chose to grasp the instrument from. Therefore, they had to drop the object and pick it back up off the workbench from a different angle. This had the potential to increase frustration and was mentioned by several participants as one of the usability issues with the simulation. One way to mitigate the problem was to disable gravity such that when the user released the trigger to drop the object, it would lock in place instead of falling to the workbench. That way, they could quickly reorient their hands without reinitiating the grasping action from the workbench.

3.3 M-LSID-NG Condition

Based on results from our previous study and usability issues observed, the goals for the next version were to determine if higher interaction fidelity could improve learning outcomes or performance scores [4]. In a between-subjects 2x2 experiment design, an LSID was compared to an HMD and gravity was either enabled or disabled to try and mitigate some of the usability issues. The Medium Fidelity, Large-Screen Immersive Display with No Gravity condition (M-LSID-NG) was similar to the 6-DOF condition from the previous study, but changed the camera view such that the workbench was out in front of the user, instead of the camera view being directly above the workbench looking down. This corresponded more to how one would interact in the real world and was regarded as having higher interaction fidelity as compared to the 6-DOF condition, since one does not usually take a measurement directly over the measurement object. When users picked the objects up off the table, there was a reduced chance that rotating the object would collide with the workbench since the user could tell how far away the objects were from the workbench. Disabling gravity affords easier hand repositioning and potentially mitigates the issue where users had to drop the object and pick it back up from a different angle. When users released the trigger on the Hydra, the object was locked in place. Several participants adopted a unimanual strategy where they locked the measurement object in place and used their other hand to take the measurement. The 6-DOF interaction requires the participant to use both hands simultaneously in order to stabilize the objects to take a precise measurement. If they adopted a unimanual approach, they would lose the dynamic frame of reference that the non-dominant

Table 1. FIFA Analysis of all experimental conditions. Green represents higher relative fidelity, yellow is medium and orange is the lowest fidelity.

	L-3DF	L-6DF	M-LSID-NG	M-LSID-G	H-HMD-NG	H-HMD-G
Dimensional Symmetry	x + y + rZ	x + y + z + rX + rY + rZ	x + y + z + rX + rY + rZ	x + y + z + rX + rY + rZ	x + y + z + rX + rY + rZ	x + y + z + rX + rY + rZ
Transfer Function Symmetry	Hand to screen	Hand to screen	Hand to screen	Hand to screen	Hand to hand	Hand to hand
Biomechanical Symmetry	Looking at screen	Looking at screen	Looking at screen	Looking at screen	Looking at hands	Looking at hands
Gravity	No	Yes	No	Yes	No	Yes
Camera View	Overhead	Overhead	Forward	Forward	Egocentric	Egocentric

hand provides according to Guiard's bimanual principles [12], which has been shown to result in usability issues [16]. Disabling gravity does not simulate how a person would take a real-world measurement of a hand-held object such as a gear or bearing, thus would have lower interaction fidelity compared to the conditions with gravity. For hand-held objects such as bearings and gears, the user would need to use both hands concurrently. Therefore, if the user adopted a unimanual approach, it could affect motor skills transference.

3.4 M-LSID-G Condition

The LSID condition with gravity enabled was considered to have higher physical correspondence to the real world compared to the previous condition with gravity was disabled (see Figure 2). The advantage was that it mimicked real-world interaction more accurately than the condition with gravity disabled and thus would be more familiar to users. The disadvantage of not being able to quickly reorient the object after initiating the grasping action was similar to the 6-DOF condition in the previous study, however the amount of confusion was reduced due to the gravity vector pointing down relative to the user instead of directly into the screen as in the previous study.

Fig. 2. Experimental apparatus for the second study featuring the M-LSID-G condition

3.5 H-HMD-NG Condition

When using a large-screen display for simulation, there is an inherent offset or transfer function between the actions performed by the hands and the visual output of the display. The user has to make a mental transformation between their visual and motor systems. However, employing an HMD for near-field interaction reduces the visuo-proprioceptive mismatch and the virtual end-effectors can be co-located with their physical hands. A construct of McMahan's interaction fidelity framework describes the transfer function symmetry as the degree of exactness with which a real-world transfer is reproduced through interaction [22]. The HMD conditions were regarded as having higher transfer function symmetry compared to the large-screen display conditions due to the reduced visual-motor offset between the end-effectors and the display. Most of the participants had never used an HMD before and needed to adapt to the way the virtual content was displayed. There could also be visual clarity issues due to the LSID being approximately 2 meters from the user and the lens of the HMD being approximately 10mm from the user's eyes.

Biomechanical symmetry refers to the amount of correspondence between the body movements made during an interaction and the body movements that would be made for the same task in the real world. The HMD has higher biomechanical symmetry compared to the large-screen display because the user is looking down at their hands, instead of looking at the screen with their hands in their peripheral vision. The HMD conditions were considered to have the highest interaction fidelity and the gravity disabled condition was relabeled to H-HMD-NG (see Figure 3).

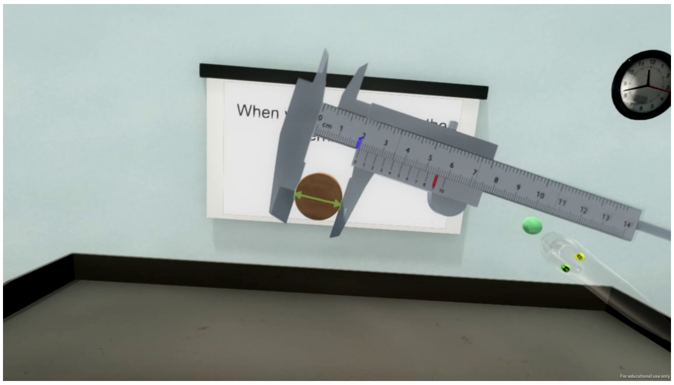

Fig. 3. A view of the metrology task from the HMD with gravity disabled

3.6 H-HMD-G Condition

Out of all six conditions, the HMD with gravity enabled condition was considered to have the highest amount of interaction fidelity. The physical fidelity was higher than the H-HMD-NG condition since gravity was enabled and the biomechanical and transfer function symmetries were higher than the LSID conditions. Users in the HMD conditions could move around like they would in the real world and move up close to the objects to get a better view.

3.7 Research Questions

In order to obtain a broader understanding of interaction fidelity, we have developed a set of new research questions.

1. How do the different levels of interaction fidelity affect task performance in a near-field, fine-motor training task?
2. What effect do the different levels of interaction fidelity have on learning outcomes and motor skills training transference?
3. What are the patterns of performance that emerge across conditions on an interaction fidelity continuum?

3.8 Hypotheses

- Hypothesis 1: Cognition scores will significantly improve from the pre to the post-test over all conditions.
- Hypothesis 2: Performance and usability will improve as the level of interaction fidelity rises.

3.9 Methodology

As we gradually designed and developed the interaction fidelity conditions from low to high fidelity, we conducted our comparative evaluations in two user studies that occurred one after the other. In the first study, a between-subjects experiment compared L-3DF to L-6DF, and a second between-subjects study conducted later compared M-LSID-NG, M-LSID-G, H-HMD-NG, and H-HMD-G conditions. In each of the studies, participants were randomly assigned to one of the interaction fidelity conditions, where the following protocol was administered to each participant. The procedure is as follows:

1. Participants read the informed consent form, filled out a demographics questionnaire, and either completed the Guilford-Zimmerman spatial abilities test or the Cube Comparison mental rotation test to determine innate spatial abilities.
2. Participants were administered the pre-cognitive assessment to determine prior knowledge of metrology concepts.
3. The inter-pupillary distance (IPD) was measured to accurately set the eye separation for stereoscopic viewing.
4. Participants completed a training tutorial to become acclimatized to selection and manipulation.
5. Participants experienced seven modules in this order: Ruler, Outside Calipers, Inside Calipers, Depth Calipers, Outside Micrometers, Inside Micrometers, and Depth Micrometers. The Ruler module was considered training and was left out of the analysis. In each of the modules, participants experienced instruction, multiple iterations of guided practice, and several trials of open exercises in bimanual interaction for precision metrology psychomotor skills learning.
6. Upon completing the simulation based training, the participant was administered the post-cognition questionnaire.
7. The participant completed the Post-Study System Usability Questionnaire (PSSUQ), Presence Questionnaire, and NASA-TLX workload assessment survey.
8. Lastly, the participant completed a psychomotor assessment in which they had to use metrology instruments to take real-world measurements at a physical workbench.

3.10 Participants

Participants were recruited from computing and engineering classes and were monetarily compensated. The full experiment lasted approximately two hours and the virtual simulation lasted approximately 45 minutes. A total of 65 participants (18 female, 47 male) completed the study with an age range from 18 to 38. The low fidelity conditions had 12 participants each. The mid and high fidelity conditions had 10 participants except for M-LSID-G, which had 11 participants. Subjects were asked to participate only if they had little to no experience with calipers and micrometers and was verified by questions in the demographics questionnaire.

4 Measures

Overall, participants in the between-subjects conditions, taken together from both studies, were subjected to the same performance and learning outcome measures except where noted. Results of the participants' cognitive and performance measures between the six interaction fidelity conditions that were designed and implemented using McMahan's FIFA framework were then analyzed as described in the following sections.

4.1 Learning Outcome Measures

The pre and post cognition questionnaires were administered directly before and after experiencing the virtual training simulation to determine basic understanding of metrology concepts. Subject matter experts categorized the questions into five levels of Bloom's Taxonomy [6] including Knowledge, Comprehension, Application, Synthesis, and Evaluation. Questions in the pre-test were similar, but not the same as the post-test. For instance, a Comprehension level question asked if given the measurement picture, does the part meet the specifications? The measurement in the picture was slightly different for the pre- and post-test.

To determine if skills were transferred to the real world, a psychomotor assessment was administered at the end of the experiment. Participants were tasked with taking real-world measurements using the same instruments from the simulation. Their measurements were considered correct if they supplied an answer that was within $\pm$1mm from the ground truth. They were also asked if the measurements were within a tolerance range and if the part could be used safely.

4.2 Quantitative Subjective Responses

In order to determine subjective responses to the simulation, three questionnaires were administered after the experiment including the PSSUQ [21], NASA-TLX [14], and Witmer et al. Presence Questionnaire Ver. 3.0 [33]. The NASA-TLX survey measured subjective workload on factors including Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. Participants first rated each category on a 20-point scale and then did pairwise comparisons between each factor. The weighted score was the sum of the pair-wise counts multiplied by the rating for each factor. The Presence Questionnaire measured how participants felt about certain aspects of the simulation and had questions categorized into factors including Involvement, Sensory Fidelity, Adaptation/Immersion, and Interface Quality. The PSSUQ questions focused on how usable the simulation felt to the user and questions were categorized into System Usability, Information Quality, and Interaction Quality factors.

4.3 Quantitative Performance Variables

Time to complete was measured for each trial in the Exercise phases and was recorded from the time the instrument and object was instantiated to when the participant supplied a correct answer. The first experiment had two trials per instrument module and the second experiment had five trials per module during the Exercise phase. The simulation asked the participant to take their measurements as quickly and accurately as possible. Time to complete is an important measure for determining efficiency and is important in a workplace setting. For instance, quality assurance technicians will oftentimes measure parts on an assembly line to ensure they meet specifications. The technicians need to complete the measurement quickly in order to evaluate the parts in a timely manner. During the Exercise phase, the user needed to supply an answer that was within a $\pm$1 mm range of the ground truth in order to move on to the next trial. If a successful measurement was supplied, the simulation recorded the user's answer as well as the reading on the instrument. Otherwise, the number of attempts was incremented and the user was asked to try again. The absolute value of the difference between the reading on the instrument minus the ground truth was defined as the physical difference. It was a measure of the participant's fine-motor skills. A physical difference of $\pm$0 cm would indicate that the user perfectly clamped the instrument to the measurement object. The reading difference was defined as the absolute value of the difference between what the user supplied as an answer minus what the instrument displayed at the time of submission. It was a measure of how well the user could read and interpret the Vernier scale regardless of how well they positioned the virtual objects. The user answer difference was a combination of both the cognitive and motor skills of the user and was defined as the absolute value of the ground truth minus the answer supplied by the user.

5 Results

5.1 Quantitative Subjective Results

5.1.1 Cognition Questionnaire Results

To analyze the quantitative cognitive questionnaire results based on Bloom's taxonomy of learning, we analyzed the data gathered on the mean percent scores in each measure first using a one-way Multivariate Analysis of Variance (MANOVA) to examine overall the effects of the 6 fidelity conditions as a between-subjects variable, on the pre and post scores of the 5 levels of cognition (Knowledge, Comprehension, Application, Synthesis and Evaluation). Then, a 2 x 6 mixed model Analysis of Variance (ANOVA) was conducted on each level of cognition with appropriate post-hocs as follow-up to examine the significant main effects of session (pre vs. post), condition, and the interaction between session and condition. Parametric analyses were chosen on the data after carefully verifying that the underlying assumptions were met - namely the data in the samples were normally distributed and error variance between samples were equivalent. Thus, it was ensured that Box's test of equality of covariance matrix was not significant, Levene's test of sphericity was conducted to ensure error variance in groups of samples were equivalent. Pairwise post-hoc tests for levels of the between-subjects variables (i.e. conditions) were conducted using Tukey's HSD method, and between levels of the within subjects (i.e. session) was conducted using Bonferroni adjusted alpha method. The MANOVA analysis revealed a main effect of condition $F(50, 275) = 2.46$, $p < .001$, Wilks' $\Lambda = .15$, $\eta^2 = .32$. See Figure 4 for overall cognition score trends.

Knowledge Level The ANOVA analysis revealed a significant main effect of time $F(1, 60) = 73.903$, $p < .001$, $\eta^2 = .56$. Overall, participants scored significantly higher in the post test session (M=73.11%, SD=23.64) as compared to the pre-test session (M=48.11%, SD=24.63), $p < .001$.

Comprehension Level The ANOVA analysis revealed a significant main effect of time $F(1, 60) = 108.3$, $p < .001$, $\eta^2 = .64$. Overall, participants scored significantly higher in the post test session (M=72.47%, SD=21.18) as compared to the pre-test session (M=29.54%, SD=28.46), $p < .001$.

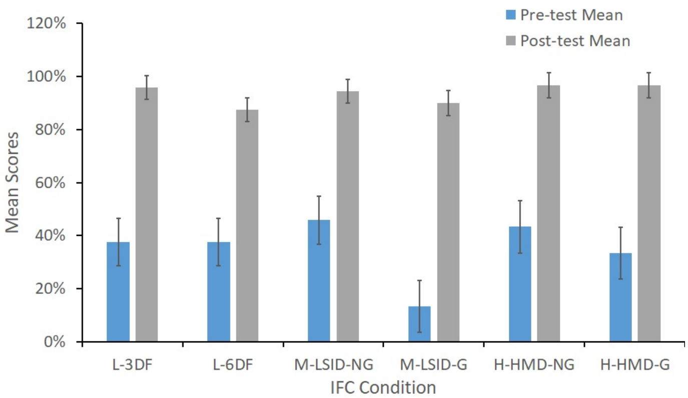

Fig. 4. Overall mean cognition scores for pre-test and post-test (error bars represent standard error)

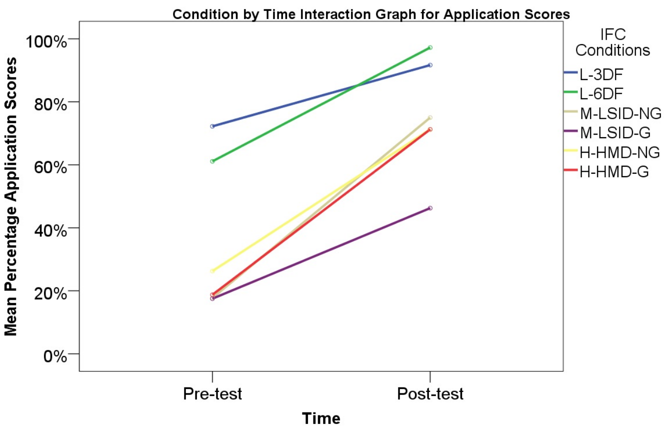

Fig. 5. Time by Condition interaction graph on Application cognition scores

Application Level The ANOVA analysis revealed a significant main effect of time $F(1, 60) = 131.56$, $p < .001$, $\eta^2 = .68$, a significant main effect of conditions $F(5, 60) = 15.29$, $p < .001$, $\eta^2 = .56$, and a significant time x condition interaction $F(5, 60) = 3.02$, $p = .017$, $\eta^2 = .20$ (see Figure 5). Overall, participants scored significantly higher in the post-test session (M=72.47%, SD=21.18) as compared to the pre-test session (M=29.54%, SD=28.46), $p < .001$. Post-hoc pairwise Tukey's HSD comparisons revealed that the pre-test

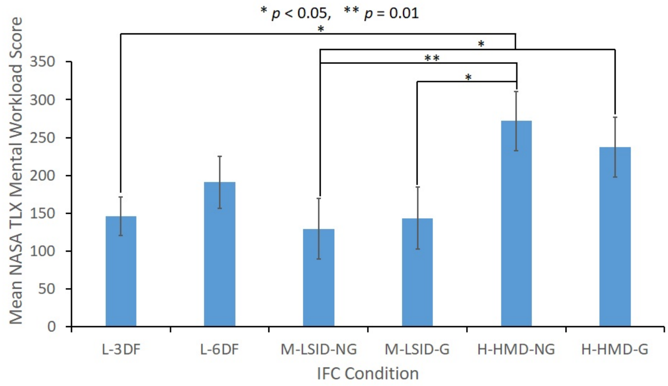

Fig. 6. Mean NASA TLX Mental Workload scores (error bars represent standard error)

scores were significant higher for L-3DF (M=72.22%, SD=23.9) and L-6DF (M=61.11%, SD=27.8) conditions, as compared to M-LSID-NG (M=17.71%, SD=14.56), M-LSID-G (M=17.5%, SD=15.8), H-HMD-NG (M=26.25%, SD=23.2) and H-HMD-G (M=36.9%, SD=32.26) conditions. Post-hoc pairwise comparisons also revealed that post-test scores were highest in condition L-6DF (M=97.22%, SD=9.6) and it was significantly higher than mid fidelity condition M-LSID-G (M=46.25%, SD=20.45) that was also the lowest overall post-test score, $p < .001$. M-LSID-G post-test scores was also significantly lower than L-3DF (M=91.66%, SD=15) $p < .001$, and M-LSID-NG (M=75%, SD=22.6) $p = .028$. Overall, in every condition participants scored significantly higher in the post-test session as compared to the pre-test session in the Application scores.

Synthesis Level The ANOVA analysis revealed a significant main effect of time $F(1, 60) = 68.2$, $p < .001$, $\eta^2 = .53$. Overall, participants scored significantly higher in the post-test session (M=74.24%, SD=25.9) as compared to the pre-test session (M=35.41%, SD=31.1), $p < .001$. Finally, in the Evaluation level, the analysis revealed a significant main effect of time as well $F(1, 60) = 177.47$, $p < .001$, $\eta^2 = .75$. Here too, overall participants scored significantly higher in the post-test session (M=93.43%, SD=15.1) as compared to the pre-test session (M=35.60%, SD=31.6), $p < .001$.

5.1.2 NASA-TLX Workload Results

In order to analyze the effects of conditions in the interaction fidelity continuum on perceived workload, we conducted a one-way ANOVA on the participants' scores in the dimensions of mental, physical, temporal, performance, effort, and frustration workload levels. Assumptions of normality of distribution and equality of variance were verified before parametric ANOVA analysis were performed. Post-hoc pairwise comparisons were conducted using Tukey's HSD method.

The ANOVA analysis revealed a significant effect of condition in the mental workload scores $F(5, 64) = 2.38$, $p = .049$. Post-hoc pairwise comparisons revealed that participants in H-HMD-NG (M=272, SD=123.45) scored significantly higher than L-3DF (M=145.8, SD=87.4) $p = .017$, M-LSID-NG (M=129.5, SD=126) $p = .010$, and M-LSID-G (M=143.6, SD=136) $p = .017$. Post-hoc pairwise comparisons also revealed that scores of participants in condition H-HMD-G (M=237, SD=125.2) was significantly higher than M-LSID-NG $p = .049$, see Figure 6.

The ANOVA analysis revealed a significant effect of condition in the frustration workload scores $F(5, 64) = 2.49$, $p = .041$. Post-hoc pairwise comparisons revealed that contrary to mental workload scores, participants in H-HMD-NG (M=25, SD=35.1) scored significantly lower than L-3DF (M=133.3, SD=158.8) $p = .035$, M-LSID-NG (M=155.5, SD=129.8) $p = .015$, and M-LSID-G (M=171.4, SD=137.4) $p = .006$. Post-hoc pairwise comparisons also revealed that frustration scores of participants in condition H-HMD-G (M=61, SD=114.5) was significantly lower than M-LSID-G (M=171.3, SD=137.4) $p = .035$, see Figure 7.

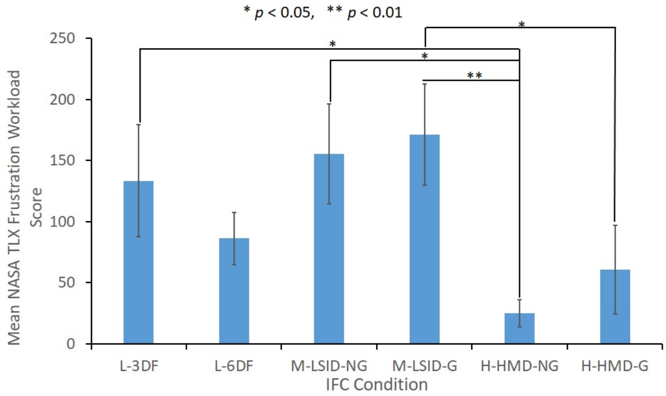

Fig. 7. Mean NASA TLX Frustration Workload scores (error bars represent standard error). Smaller values are better

5.1.3 System Usability Results

In order to analyze the effects of condition in the interaction fidelity continuum on system usability scores using the PSSUQ questionnaire, we conducted a one-way ANOVA analysis on all the participants' scores in the dimensions of System Usability, Information Quality and Interaction Quality. The ANOVA analysis revealed a significant effect of condition on the system usability scores of the PSSUQ questionnaire $F(5, 64) = 2.63$, $p = .032$. Post-hoc pairwise comparisons revealed that system usability scores were significantly higher in H-HMD-NG (M=4.25, SD=.38) as compared conditions L-6DF (M=3.52, SD=.76) $p = .008$, M-LSID-NG (M=3.69, SD=.67) $p = .044$, and M-LSID-G (M=3.65, SD=.66) $p = .035$. Similarly, post-hoc pairwise comparisons also revealed that system usability scores were significantly higher in H-HMD-G (M=4.24, SD=.58) as compared to conditions L-6DF $p = .009$, M-LSID-NG $p = .049$, and M-LSID-G $p = .039$.

Statistical analysis did not reveal any significant difference in Presence or Psychomotor Skills assessment questionnaires.

5.2 Quantitative Objective Results

5.2.1 Time to Complete

Time to complete was measured from the beginning to the end of each trial in the open exercise phase in each experimental condition. Time to complete is both a measure of how efficiently and accurately the user was able to take a measurement. If a user supplied incorrect measurements, their time to complete necessarily increased, as they had to repeat the trial. A one-way ANOVA was performed on the mean time to complete scores between IFC conditions. The ANOVA analysis revealed a significant effect of condition $F(5, 1516) = 34.37$, $p < .001$. Post-hoc pairwise Tukey HSD analysis revealed that mean time to complete scores for L-6DF was the highest and was significantly higher than M-LSID-NG $p = .001$, M-LSID-G $p < .001$, H-HMD-NG $p < .001$, and H-HMD-G $p < .001$. Mean time to complete in L-3DF was second highest and was significantly higher than M-LSID-G $p < .001$, H-HMD-NG $p < .001$, and H-HMD-G $p < .001$. Mean time to complete in M-LSID-NG was the third highest and was significantly higher than M-LSID-G $p = .002$, H-HMD-NG $p < .001$, and H-HMD-G $p < .001$. Mean time to complete in M-LSID-G was the fourth highest and was significantly higher than H-HMD-NG $p = .010$. Mean time to complete in H-HMD-G was not significantly different from H-HMD-NG. Generally, as interaction fidelity increase we found that mean time to complete the task gradually decreased. Please see Figure 8.

5.2.2 Physical Difference

The physical difference was calculated by taking the absolute value of the difference between the ground truth defined by the bounding volumes of the measurement object (e.g. bearing, gear, etc.) minus the reading on the instrument at the time of submitting the answer. With regard to performance, the physical difference measured how accurately users were able to maneuver the virtual objects to take the measurement

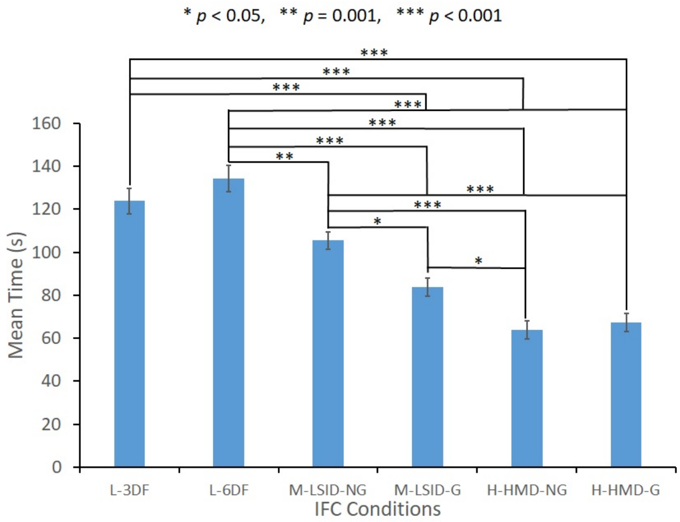

Fig. 8. Overall mean time to complete by module (error bars represent standard error). Smaller values are better

and was a measure of their motor abilities. A value closer to 0cm means a more accurate physical measurement. The one-way ANOVA on the physical difference scores revealed a significant main effect of condition $F(5, 1516) = 5.33$, $p < .001$. Post-hoc Tukey's HSD results revealed that mean physical difference was the lowest in condition L-3DF, and it was significantly lower than M-LSID-NG $p = .001$, M-LSID-G $p < .001$, and H-HMD-NG $p = .020$. Post-hoc tests revealed that H-HMD-G was the second lowest, and it was significantly lower than M-LSID-NG $p = .05$, and M-LSID-G $p = .019$. Conditions L-6DF and H-HMD-NG were the third highest pairs of physical difference scores and were not significantly different from the others except L-3DF as reported previously. Finally, conditions M-LSID-G and M-LSID-NG had the highest mean physical difference error scores, and were not significantly different from other conditions except L-3DF and H-HMD-G as reported previously. As shown in Figure 9, the mean physical difference error scores were highest in the medium fidelity conditions followed by the L-6DF and H-HMD-NG, and were lowest in the H-HMD-G followed by L-3DF respectively.

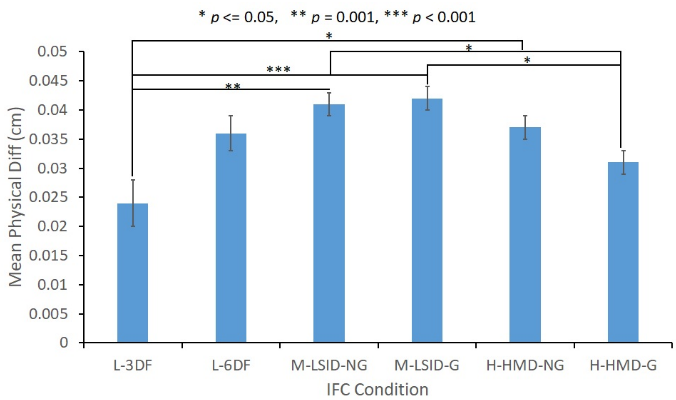

Fig. 9. Physical difference by condition (error bars represent standard error). Smaller values are better

5.2.3 Reading Difference

Similar to the physical difference, the reading difference was the absolute value of the differences between what the user supplied as an answer and the reading on the instrument at the time of submitting the measurement. In terms of performance, this was independent from how well the user physically took the measurement as compared to how well they were able to read and interpret the Vernier scale of the virtual precision metrology instruments that they were interacting with. Higher reading differences resulted in higher error in reading and interpreting the precision metrology instruments during 3D interaction, and a value closer to 0cm means more accurate reading and interpretation. A one-way ANOVA was performed on the mean reading difference error revealed a significant effect of condition $F(5, 1516) = 5.33$, $p < .001$. Post-hoc pairwise comparisons using Tukey's HSD revealed that M-LSID-G had the highest mean difference error, and was significantly higher than condition L-3DF $p = .001$, L-6DF $p < .001$, H-HMD-NG $p < .001$, and H-HMD-G $p = .013$. Mean reading difference error was the second highest in the M-LSID-NG condition, but was not significantly different than the other conditions. Overall, as shown in Figure 10, the trend in the scores were that participants in the medium interaction fidelity conditions had the highest reading error as compared to the lowest and highest fidelity conditions, which were generally more accurate.

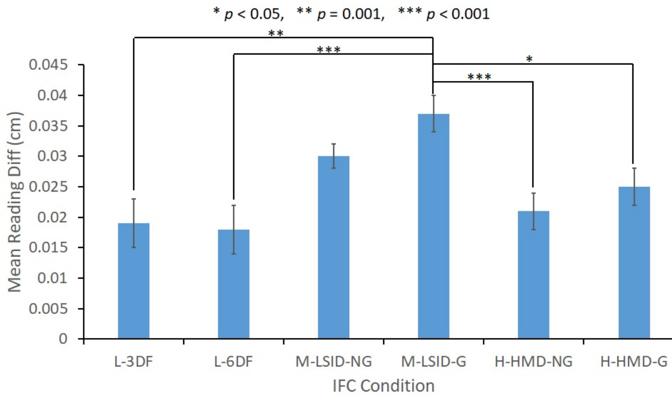

Fig. 10. Reading difference by condition (error bars represent standard error). Smaller values are better

5.2.4 User Answer Difference

The user answer difference was the absolute difference between what the user supplied as an answer as compared to the ground truth in the open exercise sessions of the precision metrology modules. It is a combined measure of both how well the user was able to maneuver the virtual objects for taking an accurate measurement as well as how well the user was able to read and interpret the Vernier scale. Mean user answer difference error scores of closer to 0cm indicates accurate task performance. A one-way ANOVA analysis of the user answer difference scores revealed a significant effect of condition $F(5, 1516) = 2.35$, $p = .039$. Post-hoc pairwise Tukey's HSD comparison revealed that mean user answer difference scores were significantly lower in L-3DF as compared to L-6DF $p = .036$, M-LSID-NG $p = .007$, M-LSID-G $p = .008$, H-HMD-NG $p = .001$, and H-HMD-G $p = .024$. Overall, as shown in Figure 11, user answer difference was the second lowest in H-HMD-G. However, there was no significant difference between M-LSID-NG, M-LSID-G, H-HMD-NG and H-HMD-G conditions.

5.2.5 Attempts

In the virtual reality bimanual precision metrology training simulation, we logged the number of attempts it took for participants to arrive at the correct answer in the open exercise session of the precision metrology modules. If the user supplied an answer that was $\pm$1mm from the ground truth, the number of attempts was incremented and the participant repeated the open exercise trial. From a performance perspective, the mean number of attempts denotes how well the participants learned the task in the guided practice sessions of the precision metrology training simulation modules to be able to successfully complete the open exercise with the least number of attempts. A one-way ANOVA analysis of the mean number of attempts revealed a significant effect of interaction fidelity conditions $F(5, 1516) = 6.78$, $p < .001$. Overall, as shown in Figure 12, the mean number of attempts is highest in condition M-LSID-G and is lowest in H-HMD-G. Post-hoc pairwise

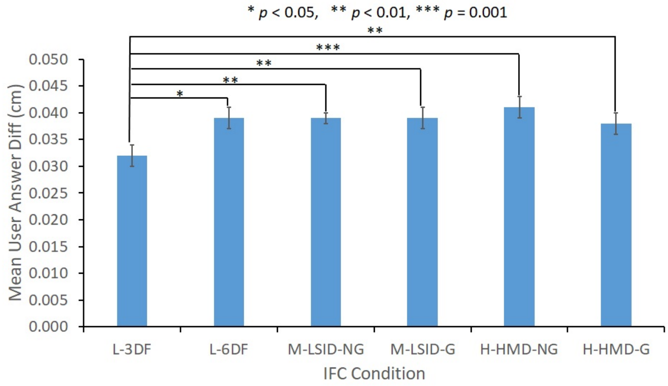

Fig. 11. User answer difference by condition (error bars represent standard error). Smaller values are better

comparisons using Tukey's HSD revealed that condition M-LSID-G was significantly higher than L-3DF (which was the third lowest) $p = .007$, H-HMD-NG $p < .001$, and H-HMD-G $p < .001$. Post-hoc pairwise comparisons also revealed that H-HMD-G was significantly lower than L-6DF $p = .024$ and M-LSID-NG $p < .001$.

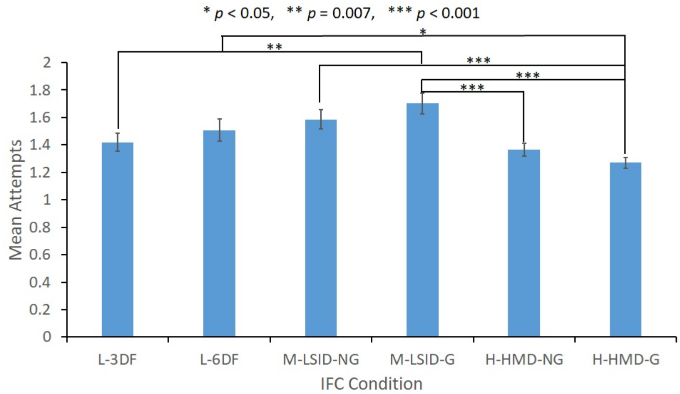

Fig. 12. Overall number of attempts by condition (error bars represent standard error). Smaller values are better

5.3 Qualitative Results

Comments were verbally collected from participants during breaks and after the experiment in each condition. With respect to L-3DF and L-6DF, some participants mentioned concerns about the gravity vector pointing into the screen. They said that it made it difficult to understand how to pick-up and drop objects at first but they got accustomed to it as they progressed. A few participants were more bothered by this and felt that it interfered with their performance to some extent. In the middle and high level conditions, M-LSID and H-HMD respectively, participants did not mention anything in particular that prevented successful completion of the task. Some participants mentioned that having no gravity made it easier to take measurements as they could leave objects in midair and measure from a better angle. A few of the participants who experienced the M-LSID conditions had problems with eye-strain and took longer to complete the experiment.

Overall, participants thought that the task was fun and engaging and helped them retain focus. Participants mentioned that the repetition of instructions in every phase aided them in the post-survey and expedited the measurement and reading procedure.

We used an after-action review tool to analyze and identify behavior patterns for each condition. In the lower fidelity conditions (L-3DF and L-6DF), due to the gravity vector pointing into the screen, participants often dropped objects to grasp them from a more advantageous angle. While grabbing objects in the L-6DF condition, participants had problems determining how high the objects were suspended and kept hitting the workbench while trying to rotate them. This often obstructed smooth maneuvering and fine-motor skills affecting performance and increasing the time to complete.

In the middle and high fidelity conditions when gravity was absent, participants left objects in midair and measured them using just one hand [4]. For comparatively bigger or complex objects, participants tended to position the objects such that it would be easier to use the measuring instruments. This was often observed in the high fidelity (H-HMD) conditions as it was possible to have a top down view of the workbench and still have the gravity vector align with the real world. Participants also found it easier to read the scale in the high fidelity condition since they could bring it closer to their face for a closer view as opposed to leaning in to make the image bigger in the medium fidelity conditions.

6 Discussion

The goal of this work was to examine several levels of interaction fidelity on a continuum to determine overall trends using a virtual metrology training simulation as a testbed. The interaction fidelity levels were determined using McMahan's FIFA framework and organized on a continuum from lowest to highest. The results reveal that across all levels of Bloom's Taxonomy, users were able learn the task regardless of the condition. Lower levels of Bloom's Taxonomy are generally easier to teach since there is normally a definite right or wrong answer. Higher levels like Synthesis and Evaluation are often overlooked because they require the learner to make value judgments based on knowledge gained in all lower levels of the taxonomy and therefore can be more ambiguous [3]. VR training applications have the advantage of actively engaging the user and it enables them to experience the learning material instead of passively listening to a lecture. Users can make mistakes and learn from them in real time, promoting creative thinking and problem solving. Participants showed significant improvement in the higher levels of the taxonomy in addition to the lower levels. Therefore, our first hypothesis was supported based on the pre and post cognition scores. Training transference to the real world was demonstrated in the psychomotor assessment, however there were no significant differences between the interaction fidelity levels. This could be due to the similarity of the task across conditions. Irrespective of the condition, the participants were asked to perform the exact same task in the real-world using the same equipment. Since the knowledge imparted was similar in each condition and there were no significant differences in the post-test scores between conditions, finding significant psychomotor differences will require finer grain analysis on other performance variables.

Several metrics in virtual task performance revealed significant differences. The overall trend for time to complete decreased as the interaction fidelity level increased. The high-fidelity HMD conditions revealed significantly faster time to complete as compared to the lower levels. Time to complete is both a measure of how efficient and accurate the participant was in completing the task successfully. The user answer difference metric was a combination of both how well they read the Vernier scale and their ability to physically manipulate the virtual objects. Due to the fewer degrees-of-freedom in the L-3DF condition, participants had an easier time physically taking measurements and as a result, their user answer difference scores showed significant improvements as compared to the other conditions. All other conditions, however, did not reveal significant differences between each other. Trends observed in the analyses above support our 2nd hypothesis which states that performance increases with the level of interaction fidelity.

The physical difference metric revealed significantly more accurate measurements for the low and high fidelity conditions. Participants performed worse in physically manipulating the virtual instruments to the correct position in the mid-fidelity conditions compared to the other conditions. The L-3DF condition had the best physical accuracy compared to all the other conditions potentially due to the fewer degrees-of-freedom that participants needed to manipulate to take the measurement. The cognitive task of interpreting the Vernier scale by way of the reading difference metric revealed that the mid-fidelity conditions performed worse than the low and high fidelity conditions, even though the reading difference was not dependent on how well they physically manipulated the instrument. As a consequence of the poor mid-fidelity physical and reading differences, the number of attempts necessarily increased. The number of attempts follows the trend set by the physical and reading difference scores where the mid-fidelity conditions revealed worse scores than the low and high end of the interaction fidelity continuum. As a result, these variables are unsupportive of our 2nd hypothesis.

Participants in the high-fidelity HMD conditions rated the system usability questions from the PSSUQ significantly higher than the mid-fidelity and L-6DF conditions. Examples of the System Usability items include "It was simple to use the system" and "I feel comfortable using the system". The trend with system usability revealed that higher-fidelity and the lowest-fidelity conditions outperform the mid-fidelity conditions. The high-fidelity HMD conditions were rated to have a higher mental demand compared to the lower-fidelity conditions in the NASA-TLX. This is unsupportive of the 2nd hypothesis similar to the physical and reading difference measures. This could be due to the HMD requiring the full attention of the user especially since many participants had not used an HMD before and had to go through an adjustment phase. Frustration scores were significantly lower in the high-fidelity conditions as compared to the lower levels despite this being the first time many participants experienced an HMD in a virtual training scenario. The frustration scores reinforce the results of the usability questionnaire, supporting the 2nd hypothesis, indicating that higher interaction fidelity could be more effective and usable than mid and low fidelity interaction. Qualitative feedback from the participants also reinforce the results from the subjective questionnaires.

An interesting trend emerges when evaluating some of the task performance metrics and subjective responses. In general, the mid-fidelity conditions had worse performance scores as compared to the low and high fidelity conditions. A natural assumption would be that as interaction fidelity linearly increases, task performance would also linearly increase. The drop in mid-fidelity performance scores slightly resembles the uncanny valley effect first hypothesized by Mori [24]. In it, as robotic likeness to humans increases, affinity towards the robot increases until there is a point at which affinity drops sharply and recuperates after some increase in human likeness. Perhaps there is a similar dip in performance in the context of interaction fidelity. Mid-fidelity interaction metaphors may perform worse because they may not be able to cater to the higher expectations of participants in those conditions. For lower fidelity interactions, users may expect an abstracted task and respond accordingly. Alternatively, higher fidelity interaction metaphors may be successful in meeting users' expectations and needs. While not all results in this work support this trend, other work by McMahan [22, 23] and Nabiyouni [27] have found similar trends where their mid-fidelity interaction techniques performed worse than the low and high fidelity techniques. This work adds to the growing body of research suggesting that if high-fidelity cannot be obtained for a virtual training simulation, it may be more beneficial to employ low-fidelity techniques instead of mid-fidelity techniques. However, further research is required to determine if this trend holds in other training domains.

7 Conclusion and Future Work

As adoption rates of VR technology continue to rise, it is important to empirically evaluate relative levels of interaction fidelity in order to maximize the benefits of VR training. The goals of this work were to determine the benefits and drawbacks of levels of interaction fidelity on a linear continuum on learning and performance in a near-field, fine-motor metrology training simulation. Results from the pre and post cognition scores revealed that users were able to learn the content across all levels of Bloom's Taxonomy overall, and interaction fidelity levels did not have a significant effect on learning outcomes in both cognitive and psychomotor scores. Performance scores in terms of accuracy of physically taking the measurements, reading the Vernier scale, and number of attempts showed that the mid-fidelity conditions performed worse than the low and high fidelity conditions. Furthermore, time to complete decreased as interaction fidelity increased. The high-fidelity conditions also had higher subjective user responses in terms of system usability and frustration scores. The physical fidelity of the simulation in terms of enabling or disabling gravity seemed to have little effect on performance outcomes, however, the incorporation of the HMD provided several significant performance advantages. The overall trend revealed that the lowest and highest fidelity interaction metaphors had better user performance as compared to the mid-fidelity metaphors, but learning outcomes were not affected by the interaction fidelity levels.

General guidelines for VR researchers, developers, and educators can be derived from the results of this work. If the primary goal of a virtual training simulation is based on learning outcomes, then a simplified low-fidelity interaction metaphor may be sufficient. If the determinative factor is efficiency, accuracy, or usability, such as a laparoscopic training simulation, then higher fidelity may be more favorable. Developers will need to take into account the possibility that employing a mid-fidelity interaction technique may result in worse performance as compared to simplified low-fidelity or high-fidelity interaction. When using a mid-fidelity metaphor, users' expectations may not be met if the interaction metaphor does not operate exactly as the real-world analogue.

Given the trend associated with medium fidelity conditions performing worse than the lower and higher fidelity conditions, this area of research mandates further investigation. An extension of our work would be to compare different fidelity conditions in an HMD only. Future work would involve analyzing multiple levels of interaction fidelity within the medium and high fidelity ranges using an HMD viewing method only, and also to examine if interaction fidelity affects learning and task performance in medium and far field VR settings.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under Grant No. DUE-1501695.

REFERENCES

[1] F. Al-Zahrani. Web-Based Learning and Training for Virtual Metrology Lab. *Journal of Telecommunications*, 1(2):42–54, 2010.

[2] F. Argelaguet and C. Andujar. A survey of 3D object selection techniques for virtual environments. *Computers & Graphics*, 37(3):121–136, 2013.

[3] J. T. Bell and H. S. Fogler. Ten steps to developing virtual reality applications for engineering education. *American Society for Engineering Education*, pp. 1–8, 1997.

[4] J. Bertrand, A. Bhargava, K. C. Madathil, A. Gramopadhye, and S. V. Babu. The effects of presentation method and simulation fidelity on psychomotor education in a bimanual metrology training simulation. In *IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 59–68. IEEE, 2017.

[5] J. Bertrand, D. Brickler, S. Babu, K. Madathil, M. Zelaya, T. Wang, J. Wagner, A. Gramopadhye, and J. Luo. The role of dimensional symmetry on bimanual psychomotor skills education in immersive virtual environments. In *2015 IEEE Virtual Reality (VR)*, pp. 3–10. IEEE, 2015.

[6] B. Bloom. *Taxonomy of educational objectives - the classification of educational goals - Handbook I: cognitive domain*. David McKay Company, Inc, 1956.

[7] D. A. Bowman and L. F. Hodges. Formalizing the Design, Evaluation, and Application of Interaction Techniques for Immersive Virtual Environments. *Journal of Visual Languages & Computing*, 10(1):37–53, 1999.

[8] W. Buxton and B. Myers. A study in two-handed input. *ACM SIGCHI Bulletin*, 17(4):321–326, 1986.

[9] G. Casiez, D. Vogel, R. Balakrishnan, and A. Cockburn. The Impact of Control-Display Gain on User Performance in Pointing Tasks. *Human-Computer Interaction*, 23(3):215–250, 2008.

[10] D. R. Flatla, C. Gutwin, L. E. Nacke, S. Bateman, and R. L. Mandryk. Calibration games. In *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*, p. 403. ACM Press, New York, New York, USA, 2011.

[11] M. W. Gribnau and J. M. Hennessey. Comparing single- and two-handed 3D input for a 3D object assembly task. In *CHI 98 conference summary on Human factors in computing systems - CHI '98*, number April, pp. 233–234. ACM Press, New York, New York, USA, 1998.

[12] Y. Guiard. Asymmetric Division of Labor in Human Skilled Bimanual Action. *Journal of Motor Behavior*, 19(4):486–517, 1987.

[13] T. Gutierrez, J. Rodriguez, Y. Velaz, S. Casado, A. Suescun, and E. J. Sanchez. IMA-VR: A multimodal virtual training system for skills transfer in Industrial Maintenance and Assembly tasks. In *19th International Symposium in Robot and Human Interactive Communication*, pp. 428–433. IEEE, 2010.

[14] S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in psychology*, pp. 139–183. 1988.

[15] K. Hinckley, R. Pausch, J. C. Goble, and N. F. Kassell. A survey of design issues in spatial input. In *Proceedings of the 7th annual ACM symposium on User interface software and technology - UIST '94*, pp. 213–222. ACM Press, New York, New York, USA, 1994.

[16] K. Hinckley, R. Pausch, and D. Proffitt. Attention and visual feedback. In *Proceedings of the 1997 symposium on Interactive 3D graphics - SI3D '97*, pp. 121–ff. ACM Press, New York, New York, USA, 1997.

[17] K. Hinckley, R. Pausch, D. Proffitt, and N. F. Kassell. Two-handed virtual manipulation. *ACM Transactions on Computer-Human Interaction*, 5(3):260–302, 1998.

[18] H. Jack. 2013 State of Manufacturing Education. In *120th ASEE Annual Conference & Exposition*, 2013.

[19] A. Leganchuk, S. Zhai, and W. Buxton. Manual and cognitive benefits of two-handed input: an experimental study. *ACM Transactions on Computer-Human Interaction*, 5(4):326–359, 1998.

[20] K. S. Lehmann, J. P. Ritz, H. Maass, H. K. Cakmak, U. G. Kuehnapfel, C. T. Germer, G. Bretthauer, and H. J. Buhr. 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, 2005.

[21] J. R. Lewis. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1):57–78, 1995.

[22] R. McMahan. *Exploring the Effects of Higher-Fidelity Display and Interaction for Virtual Reality Games*. Doctoral dissertation, Virginia Polytechnic Institute and State University, 2011.

[23] R. P. McMahan, C. Lai, and S. K. Pal. Interaction fidelity: The uncanny valley of virtual reality interactions. In *International Conference on Virtual, Augmented and Mixed Reality*, pp. 59–70. Springer, 2016.

[24] M. Mori, K. MacDorman, and N. Kageki. The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine*, 19(2):98–100, 2012.

[25] C. I. Muntean. Raising engagement in e-learning through gamification. In *The 6th International Conference on Virtual Learning ICVL 2011*, number 1, pp. 323–329, 2011.

[26] N. Murray and T. Fernando. An Immersive Assembly and Maintenance Simulation Environment. In *Eighth IEEE International Symposium on Distributed Simulation and Real-Time Applications*, pp. 159–166. IEEE, 2004.

[27] M. Nabiyouni, A. Saktheeswaran, D. A. Bowman, and A. Karanth. Comparing the performance of natural, semi-natural, and non-natural locomotion techniques in virtual reality. In *2015 IEEE Virtual Reality (VR)*, pp. 243–244. IEEE, 2015.

[28] G. Norman, K. Dore, and L. Grierson. The minimal relationship between simulation fidelity and transfer of learning. *Medical Education*, 46(7):636–647, 2012.

[29] I. Poupyrev, M. Billinghurst, S. Weghorst, and T. Ichikawa. The go-go interaction technique. In *Proceedings of the 9th annual ACM symposium on User interface software and technology - UIST '96*, pp. 79–80. ACM Press, New York, New York, USA, 1996.

[30] U. Schultheis, J. Jerald, F. Toledo, A. Yoganandan, and P. Mlyniec. Comparison of a two-handed interface to a wand interface and a mouse interface for fundamental 3D tasks. In *2012 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 117–124. IEEE, 2012.

[31] V. J. Shute and J. Psotka. Intelligent Tutoring Systems: Past, Present, and Future. *Armstrong Lab Brooks AFB TX Human Resources Directorate*, 1994.

[32] 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, 1998.

[33] B. G. Witmer and M. J. Singer. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoperators and Virtual Environments*, 7(3):225–240, 1998.