The Effects of Visual Realism on Search Tasks in Mixed Reality Simulation

Cha Lee, Gustavo A. Rincon, Greg Meyer, Tobias Höllerer, and Doug A. Bowman

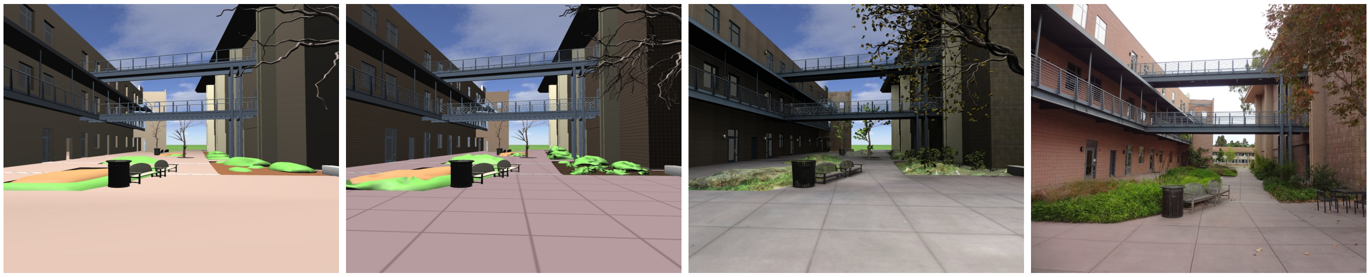


Fig. 1. From left to right are the low, medium, and high levels of visual realism, and a photograph of the real-world location used for the real AR condition.

Abstract—In this paper, we investigate the validity of Mixed Reality (MR) Simulation by conducting an experiment studying the effects of the visual realism of the simulated environment on various search tasks in Augmented Reality (AR). MR Simulation is a practical approach to conducting controlled and repeatable user experiments in MR, including AR. This approach uses a high-fidelity Virtual Reality (VR) display system to simulate a wide range of equal or lower fidelity displays from the MR continuum, for the express purpose of conducting user experiments. For the experiment, we created three virtual models of a real-world location, each with a different perceived level of visual realism. We designed and executed an AR experiment using the real-world location and repeated the experiment within VR using the three virtual models we created. The experiment looked into how fast users could search for both physical and virtual information that was present in the scene. Our experiment demonstrates the usefulness of MR Simulation and provides early evidence for the validity of MR Simulation with respect to AR search tasks performed in immersive VR.

Index Terms—MR Simulation, visual realism, augmented reality.

1 Introduction

User experiments in the domain of AR are particularly difficult to control and repeat for multiple reasons. One major reason is the uniqueness of the display systems used during the experiments. While advances in AR display technologies have increased the availability of hardware and software to choose from in creating AR applications, it is also generally the case that different researchers use inherently different AR display systems. The components of these display systems can range from low-cost off-the-shelf devices to expensive state-of-the art devices, including different visual displays and tracking systems. With so many choices for each component of hardware and software, it is prohibitive in time and cost to attempt to replicate prior experiments using the same display systems. In some cases it is impossible as those components become obsolete or unavailable. While using real display systems produces valid results, these results are often not generalizable as the results are heavily influenced by the display system itself. Experiments conducted this way are limited as it is impossible to isolate and study the influence of particular display properties at fine and incremental levels since we are restricted to what is currently available in the market.

Another major issue is the overall difficulty of conducting AR experiments in practical and useful environments. These types of environments are often crowded outdoor locations and not in controlled laboratory settings. Factors such as natural lighting and the presence of passersby make it very hard to create controlled and repeatable conditions. Natural lighting can change from moment to moment, even in the best of times, and can drastically change the appearance of the environment. The most interesting locations, where we would expect to use our AR systems, usually have real people coming in and out of the scene. These and similar issues can play havoc with any experiment. On the technical side, it is still very difficult to obtain robust, low-latency tracking in outdoor environments. Most sensor-based tracking systems are not practical for outdoors, and computer-vision methods do not yet produce robust enough tracking for extended experiments outdoors.

One method to achieving both control and repeatability is to use MR Simulation [3, 6]. MR Simulation is the concept of using a high-fidelity VR display system to simulate other displays and environments from the MR continuum for the express purpose of conducting controlled user experiments. This approach simulates the different displays by replicating the level of immersion of the target display system. In this paper, we use Mel Slater's definition of immersion as the objective measure of sensory fidelity provided by a display system [14]. By using a high-fidelity VR display system, we are able to simulate a target MR display system possessing equal or less sensory fidelity.

Once we have simulated the AR display system, we can also use a simulated environment where we can conduct our AR experiments in a controlled and repeatable manner. An important question that remains is whether this approach to conducting AR user experiments produces valid results. In other words, are the results obtained from an experiment conducted in VR and using MR Simulation the same as those garnered from the same experiment conducted in the real world

- *Cha Lee, Gustavo A. Rincon, Greg Meyer, and Tobias Höllerer are with the University of California Santa Barbara. Email: chalee21@cs.ucsb.edu, gustavo.a.rincon@gmail.com, meyer.greg.pro@gmail.com, and holl@cs.ucsb.edu*
- *Doug A. Bowman is with the Center for Human-Computer Interaction at Virginia Tech. Email: dbowman@cs.vt.edu*

Manuscript received 13 September 2012; accepted 10 January 2013; posted online 16 March 2013; mailed on 1 May 2013.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

using a real display system? It is safe to assume that this approach is not valid for every possible scenario, but we also predict that there exists a subset of AR displays and AR tasks that, when simulated in VR, can produce valid experimental results. Part of our research is to find these scenarios.

We have classified the major components of the fidelity of an MR system as display, interaction, and simulation fidelity [10]. Display fidelity is defined by how accurately we replicate the sensory fidelity provided by the real-world display. Interaction fidelity is defined by how well we replicate the interactions performed in the real world. And simulation fidelity is how faithfully we are able to replicate the environment and objects as seen in the real world. A big question regarding the validity of MR Simulation concerns simulation fidelity, or more precisely, the appearance of the simulated environments. To discuss this, we define *visual realism* as the degree to which the images of the simulated world are perceived to be real by the user. With the currently available technology, a simulated environment will not appear as realistic as the real world and so such an environment would have a lower level of visual realism. If we conduct an experiment in a simulated environment and compare it to the same experiment conducted in the equivalent real-world environment, would this mismatch in visual realism cause different results?

All other things being equal, we are interested in the question of what level of visual realism is needed to successfully replicate AR experiments involving search tasks. To address this question, we designed an outdoor AR experiment that asked participants find both virtual and physical information that was present in the environment, using a video-see-through AR display system. We created three different virtual models of the same outdoor environment, each with a different perceived level of visual realism. Next we conducted the AR experiment in the real world, outdoor location and then repeated the same experiment indoors, using MR Simulation with the three virtual models we created. Our results shed light on how visual realism affects the validity of MR Simulation and how visual realism affects search tasks performed in immersive VR.

2 Related Work

MR Simulation is our term for an approach that has been used by other researchers when investigating the effects of display factors. Our colleagues at Virginia Tech and our group have been using this approach to study the effects of level of immersion on both VR and AR display systems. Ragan et al. [12] and our group [6, 7, 8] have used MR Simulation to conduct controlled AR experiments. Ragan used a four-sided CAVE to simulate and investigate the effects of registration error on user performance for an AR task involving precise motor control and object manipulation. In our work, we used an HMD-based VR display to study the effects of simulator latency on a 3D path tracing experiment and a visual path following task.

One of the earlier AR works was by Gabbard et al. [3]. The authors used MR Simulation to study the effect of natural lighting and textures on users' ability to read and identify text information in optical see-through AR. Their motivation was very similar to ours, as it is very difficult to conduct controlled experiments with unpredictable environmental conditions. While the authors could not successfully simulate real-world lighting conditions, this approach was well suited for night, dawn, dusk, and indoor AR. In more recent work, Knecht et al. [5] demonstrated a framework for rendering photo-realistic objects in AR. In an early study, their results suggest that illumination did not affect task performance (depth estimation and object placement). Conversely, Sugano et al. [17] studied the effects of realistic shadows on virtual objects in AR. Their experiments report that shadows provide a better sense of presence despite incorrect light cues.

While we have stated our definition of visual realism, it is a difficult concept to define precisely. How do the components of a rendered image increase or decrease a person's perception of realism? Rademacher et al. [11] presented a method for measuring the perception of visual realism in images. In an experiment, users were presented with a series of images and asked to rate whether it was real or not. By controlling and varying the different factors of shadow softness, surface smoothness, lighting, and geometry the authors show that both soft shadows and surface type affect perception of visual realism. Elhelw et al. [2] used an eye-tracking system to systematically determine important features of objects that contribute to higher visual realism in rendered images. In their experiments, users were asked to rank and choose between rendered and real objects. Using eye-tracking, the authors were able to determine that specular reflection was an important feature to render correctly for higher visual realism. In an image processing approach, Wang et al. [19], proposed a realism metric based on the appearance of roughness, color, and shadow of rendered scenes. Although this approach is able to measure certain factors of visual realism, additional work will be needed to determine if the metrics correspond to human responses.

As for the effects of different levels of visual realism, there have been many prior studies in the psychology and VR communities. In the famous pit experiment, Slater et al. [15, 4] compared real-time ray tracing with ray casting to determine if the differences in the rendered images affected users' sense of presence. Users were placed in a virtual room with a large pit or hole in the ground and asked to look around and into the pit while standing on the edge. Using both a presence questionnaire and physiological recordings, their results indicated an increase in presence for the ray tracing condition which lends evidence that higher visual realism does have an effect on presence. Another experiment by Vinayagamoorthy et al. [18] investigated the impact of texture quality and character realism on presence in a street walking experiment. The results from this study demonstrated the "uncanny valley" effect as the condition with the high fidelity characters and low fidelity environment produced the lowest sense of presence. The authors proposed that this effect was due to the inconsistency between the characters and environment which adversely affected presence.

In VR, Mania et al. [9] and Stinson et al. [16] recently looked at the effects of visual realism on training transfer. Mania investigated the effect of different shading techniques (flat-shading or radiosity) on a memory task. Users would be exposed to the virtual environment and afterwards were asked to arrange the objects in the real-world equivalent room to match what they remembered. Their results indicated that users in the flat shaded environment were able to better remember the location of objects. Stinson et al. studied how the complexity of the environment affects training transfer of a scanning task. Users were trained in searching for threats in an urban environment, with different levels of complexity and realism. Their results indicate a slight advantage for users trained in the more realistic scenes.

3 MR Simulation and Visual Realism

We mentioned earlier that simulation fidelity is one of three components of the overall fidelity (along with display and interaction fidelity) of an MR system. This is perhaps the most difficult of the three components to achieve when considering simulated AR. Prior work by Lee et al. [6, 7, 8] has found evidence to support using MR Simulation for AR experiments, concerning display and interaction fidelity, but little work has been done to investigate simulation fidelity as it pertains to the validity of MR Simulation. When simulating AR, the most common approach is to use some type of virtual model to represent the real-world environment. Since virtual models can be created using a large variety of techniques and formats, this can also result in different models with a wide range of visual realism. Model geometry can be be image-based, point-based, or stored as polygonal data. Color information can be represented as simple colors, material properties, or textures. And there are a variety of techniques for lighting virtual environments that can produce significantly different images, such as ray casting, or ray tracing.

The data formats and rendering techniques are all factors of visual realism and their effect on perceived visual realism is difficult to judge, as different techniques work better in different environments and in different combinations. For example, high-resolution panoramas can be much more visually realistic in static environments than even the most complex 3D models. In a dynamic environment, when both the user and the objects are allowed to move, panoramas lose their ad-

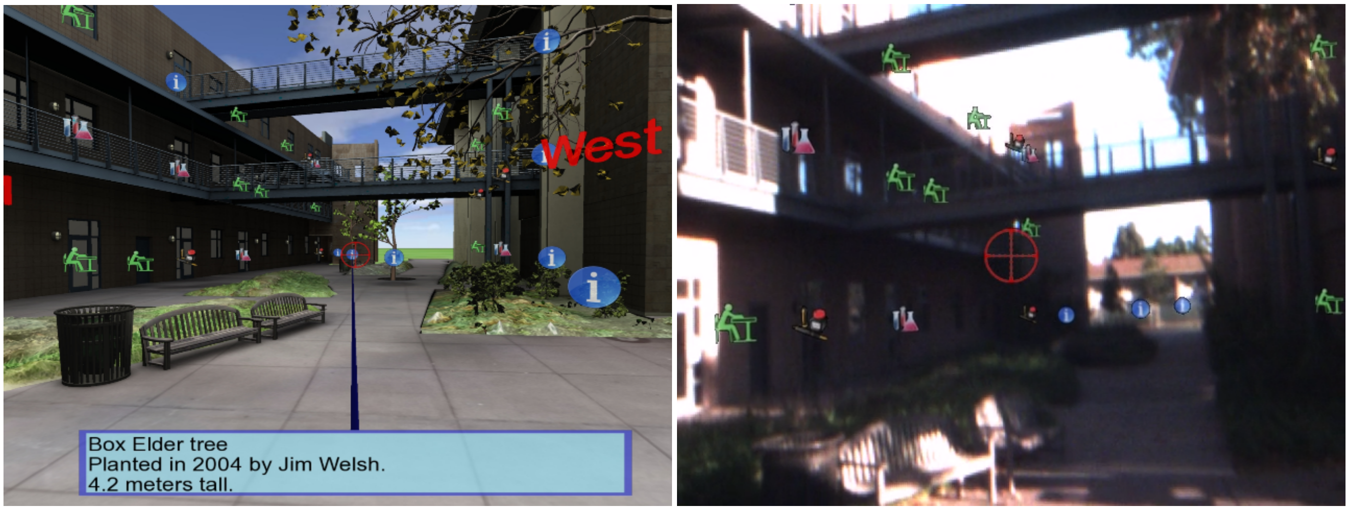


Fig. 2. The high-fidelity model and the real environment, both with virtual annotations. There are four different types of annotations: The blue icon with the "i" represents miscellaneous information. The green icon that looks like a student at a desk represents classroom information. The icon with beakers indicates lab information. And the pen and ink icon indicates information about an office. The real environment here was captured at a lower resolution, but the lighting artifacts are accurately represented.

vantage as depth and image artifacts begin to appear. Another difficulty occurs when different factors of visual realism are intermixed or are changed in different directions. For example, low-polygon models with high-resolution textures are often used in video games due to memory constraints. How does this type of model compare with a high-polygon model with simple or no textures? One is more correct with respect to color information, but the other is more correct with respect to spatial information. While the factor of geometry and spatial information increases when polygon count is increased, the factor of color and material information decreases as texture realism is reduced. Determining the interaction of these different factors on visual realism is beyond the scope of this paper. For this work, we are only interested in the effect of visual realism on an AR search task in simulation. All we require are models with objectively higher levels of perceived visual realism.

We argue we can achieve this by choosing to use the same type of virtual model for all our models, to vary only a few factors of visual realism, and to always vary them in the same direction. If we are careful not to inter-mix different data formats and rendering techniques, or change their levels in opposite directions, there should be no interactions that could cause unanticipated responses in visual realism. If we only increase the factors, we argue that this produces progressively and objectively higher levels of visual realism. In our experiment we created three different models of the same scene with different levels of visual realism. For the data format, we chose a polygon-based model and we chose to vary three factors of visual realism: geometry, texture, and lighting. We increase geometry by increasing polygon count. We increase texture realism by progressing from no textures to simple textures and to high resolution textures. And we increase lighting realism by progressing from simple flat shading to baked-in realistic lighting. Since we never decrease a factor, we avoid the issue of interactions between these factors that may cause a decrease in visual realism and obtain three models with objectively increasing levels of visual realism. These can be seen from a sample vantage point in Figure 1, while Figure 2 shows the high-fidelity simulated backdrop and the real AR case, both with virtual annotations.

4 Experiment

The motivation for our work stems from the past research into the impact of visual realism on task performance in VR and the validation of MR Simulation as reported by [3, 12, 6, 7, 8]. While the general response from the community to MR Simulation has been cautiously positive, simulation fidelity is a potential impact factor that has yet to be tested. That is, is it necessary to create visually realistic environments when conducting AR experiments in VR? Our goals are two-fold. For the validity of MR Simulation, we need to determine if visual realism has an impact on task performance and if so how. These tasks must be practical, useful, and common. As a corollary of this goal, we also want to determine if visual realism has an impact within simulation itself. That is, if we only consider experiments in virtual environments, does visual realism impact results?

Our approach for investigating these goals was as follows. We would design an outdoor AR experiment requiring a common and useful task. We would choose a real-world location with interesting and common features to conduct our outdoor experiment. Then we would carefully measure and create three models of this location with varying levels of visual realism by varying the factors we chose: geometry, textures, and lighting. Finally, we would conduct the AR experiment in the real-world location and execute this experiment indoors, using the three models as simulated AR backdrops.

4.1 Task and Environment

Environment Our choice of the real-world location for our experiment was based on practical constraints and on the features present in the scene. This can be seen in Figure 1. We were constrained to a location within practical walking distance and to a location in which we could safely and legally conduct our outdoor experiment. This limited us to the university campus, but within that constraint we chose a location that offered a variety of interesting physical features. As seen in the figure, this location had many physical objects placed at varying and interesting points throughout the scene in both the vertical and horizontal directions. It had a good balance of vegetation to man made objects and it offered different lighting conditions throughout the scene.

Before the actual modeling work began, we took extensive measurements of the location. With the help of Dr. Bodo Bookhagen of the UCSB Geography Department, we were able to scan the location using a high-end LIDAR system. This produced a very accurate ($<$ 1 cm error) model represented as a colored point cloud. Although this point cloud model was highly accurate, we did not use it for our experiments. We felt it was visually less appealing and the cost of rendering that many points was too expensive for our simulator. Instead we used the point cloud as ground truth data and as a guide for placing our final polygonal models.

To create the models themselves, we hired two talented graduate students from our Media Arts and Technology Department with extensive experience in architectural design and 3D modeling. These individuals obtained detailed CAD plans from the facilities department on campus and combined these measurements with the point cloud models. With this additional information and our own hand measurements, we built the three different virtual models of the courtyard, as seen in Figure 1. This task proved to be a major project in itself, requiring multiple iterations and physical on-site measurements that spanned over six months.

The process of modeling the environment began by creating the low-fidelity model first. This model, as seen in Figure 1, contained polygons for all the major features in the scene. The buildings, windows, trees, benches, and trash bins can be clearly seen and there is no confusion as to what these objects are. In this model, the minor features were not modeled, such as door knobs, individual squares in the courtyard, railings, etc. No textures were used for this model and only color was applied to the surfaces. Lighting was also simple ambient lighting with no shadows. In the medium-fidelity model, we increased the number of polygons for geometry. We modeled the railings, windows, and created more complex versions of the vegetation. For textures we used simple low-resolution textures and the lighting was unchanged. In the high-fidelity model we increased geometry by increasing the polygons for the objects already in the scene and by modeling all the visible features in the scene, including signs and door knobs and such. High resolution images were used to create the textures and we used baked lighting to create static shadows and shading. As seen, each factor of visual realism (geometry, textures, lighting) remains the same or is only increased for the next model.

Task

The task in this experiment takes its cue from the work of Bowman et al. [1] on Information-Rich Virtual Environments (IRVE). Bowman and colleagues define IRVEs as virtual environments that are augmented with additional abstract information such as text or images. This is similar to AR browsers, where virtual content is placed in the real world. Taking our cue from this prior work, we wanted an AR search task that would require our participants to look through the environment surrounding them for both *virtual* and *physical* information. Similar to the task properties described in [1], virtual information refers to any information that can only be obtained via the augmentations in the real-world AR scenario such as text information. Physical information refers to the information that is available naturally in the environment such as spatial relationships and real-world objects.

The tasks users were given asked them to search for both virtual and physical information, based on both virtual and physical information criteria. This type of task is required when using most AR applications in general and AR browsers specifically. We felt it was also arguably the most recognizable type of AR task, requiring very little training or prior knowledge to use. Since the virtual objects would be registered (situated) in 3D space at the location of the real-world object they refer to, it should be an intuitive AR layout once it was explained. Figure 2 is a screen-shot of the environment with the types of information available. Due to the physical constraints (cables, weight), the users were limited in their ability to move freely. The participants could rotate approximately two full turns and move around a small area (2 m by 2 m) before cable length became an issue.

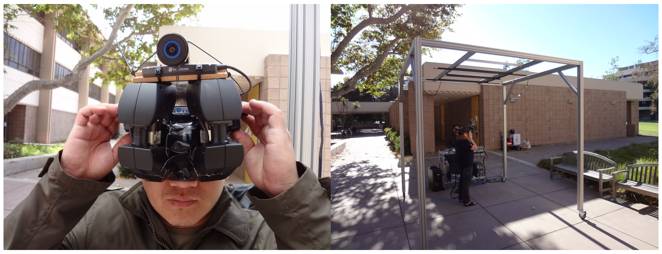


Fig. 3. The display system used for this experiment consisted of: an NVis SX111 HMD, a Pointgrey USB3 Flea camera, and an InterSense IS900 tracking system. The image on the left shows the HMD setup with the camera and the image on the right shows the custom tracking frame built for the outdoor condition. For the simulated conditions with the low, mid, and high models, the experiment was conducted indoors with the same display and tracking systems.

The content of the virtual information we created was based on what a user would want to do in that particular environment. Since we were in a school environment, we created virtual information that would help a new student. To reduce visual clutter, we generated icons that contained the information about classes, professors, labs, offices, and landmarks in the scene and overlaid them onto their object of reference. We created four different icons that represented the four different types of information in the scene: miscellaneous information, class information, office information, and lab information. Using these four icons, we now had a layout that our users could easily understand with minimal training as seen in Figure 2.

Once we had created the information and placed the icons, we generated 16 task questions. The 16 task questions can be seen in Table 1. For each task question the participant was required to find certain pieces of information and verbally report that information. There were a mix of questions that required the user to find both virtual information and physical information based on both virtual and physical information context. Target information refers to the exact nature of the response required by the task question. Physical information is inherent in the real world while virtual information needs to be provided by the virtual icons and the text they contain. Criteria information refers to what information the user is using to search the scene. For instance, Q6 and Q10 ask for trees planted by someone (virtual information) but the participant always begins by searching for trees (a physical piece of information). The task questions also varied in terms of their complexity. Some required the user to search for a single item; some required a search for multiple items; some asked the user to compare two or more items; and a few required deeper analysis and understanding of the content in the scene. As this is an exploratory study, the task questions are varied with respect to their difficulty.

Presence Questionnaire

We also measured the user's sense of presence and so we chose a subset of the questions in the Witmer-Singer (WS) presence questionnaire [20]. Although there are still questions about the validity of using questionnaires to measure the level of presence [13], the WS questionnaire is at least a recognized questionnaire that we can use to qualify our own results. The subset of questions we chose were those that concerned the visuals and visual realism. For each of these questions, users were asked to respond with a value from a rating scale with 7 positions, starting from low to high. Only low and high appeared at the ends of the scale. Questions 7 and 10 asked users for a negative response and so the resulting numerical value response was subtracted from 8, for consistency. The questions were:

- P1 : How much did the visual aspects of the environment involve you?
- P2 : How much did your experiences in the virtual environment seem consistent with your real-world experiences?
- P3 : How completely were you able to actively survey or search the environment using vision?
- P4 : How compelling was your sense of moving around inside the virtual environment?
- P5 : How involved were you in the virtual environment experience?
- P6 : How completely were all of your senses engaged?
- P7 : How inconsistent or disconnected was the information coming from your various senses?
- P8 : How closely were you able to examine objects?
- P9 : How well could you examine objects from multiple viewpoints?
- P10 : To what degree did you feel confused or disoriented at the beginning of breaks or at the end of the experimental session?

Task ID	Task Question	Search Target Information Type	Search Criteria Information Type	Task Complexity
Q1	On what floor and which building is the course, Research Methods in Social Psychology, taught?	physical	virtual	single item
Q2	What course is taught in the room behind the third window from the left, on the second floor of East Hall?	virtual	physical	single item
Q3	Who runs the lab located by the first floor entrance to West Hall, under the two bridges?	virtual	physical	single item
Q4	What kind of tree is the large tree in the courtyard?	virtual	physical	single item
Q5	What course is taught in the room directly below the undergraduate computer lab?	virtual	virtual	single item
Q6	How many trees were planted by John Deere?	physical	physical	multiple item
Q7	How many benches are located in the courtyard?	physical	physical	multiple item
Q8	How many trash bins are located in the courtyard?	physical	physical	multiple item
Q9	How many entrances into West Hall can you see from this point?	physical	physical	multiple item
Q10	Who planted more trees in the courtyard?	virtual	physical	multiple item
Q11	Which building has the most classrooms?	physical	virtual	multiple item
Q12	Which professor has the office which is located the furthest from their lab?	virtual	virtual	comprehension
Q13	What type of courses are generally taught in North Hall?	virtual	virtual	comprehension
Q14	What kind of research is the only Distinguished professor in Psychology involved in?	virtual	virtual	comprehension
Q15	What is the common theme for the location of upper division Psychology courses?	physical	virtual	comprehension
Q16	Which building would an undergrad student go to for answers to non-course related questions?	physical	virtual	comprehension

Table 1. Table of the task questions. Task questions can be categorized by their search target information type and their search criteria information type. Task complexity is a rough description of the goals of the task. Single item refers to a single search target. Multiple item refers to tasks that require counting or searching for multiple targets. Comprehension refers to tasks that ask the participant to understand and reason about the relationships between the virtual information and the environment.

4.2 Display System

The display system used for this experiment can be seen in Figure 3. The display was an NVis SX111 head-mounted display. This wide field-of-view (FOV) display allowed for a combined 102 degrees horizontal FOV and 64 degrees vertical FOV. It also provided relatively high resolutions at 1280x1024 pixels per eye, with a spatial resolution of 3.6 arc-min/pixel. The camera image was provided by a PointGrey USB3 Flea camera, with a Theia SY110M ultra-wide/no distortion lens. This camera setup was configured to provide approximately 100 degrees horizontal FOV, at 1600x1024 pixels, and a frame rate of approximately 60 frames per second. The tracking system used was InterSense's IS900, with a wired head tracker. The display system was run on a Windows 7 PC with a Quadro 5600, an Intel Core2 2.4 GHz Duo-Core CPU, and 2.0 GB of memory. The software used to run the simulation was based on WorldViz's VR toolkit, Vizard 4.0. The camera was not used for the simulated conditions since the camera feed was simulated.

Although the HMD was a stereo display, we decided to use it as a monocular display due to the inherent difficulty of using and correctly aligning and calibrating two cameras in correct stereoscopic video-see-through AR. With our camera setup, it was not possible to stream the high resolution images we desired at 60 fps for two cameras. We felt that a stereo AR setup was not as important in our experiment as realistic update rates, and that a monocular setup with wide FOV was more important to this type of task. Since the SX111 is not a full-overlap (only 66 %) display, creating monocular images is not straight forward. We used an off-screen rendering technique to create a monocular view in the stereo display, which mapped the smaller FOV of the camera onto the FOV of the HMD in a correct manner. This left small black areas at the edges, but did not distort the camera image.

4.3 Experimental Design

We used a between-subjects study design with 16 different tasks. The between-subjects variable for each task question was the level of realism of the environment. The three virtual models and the real-world location created four different levels: low, medium, high, and real that correspond to the low-fidelity model, medium-fidelity model, high-fidelity model, and real-world location. The dependent variable for each task was search time, and this was measured from the moment the user was told to start searching, until the correct answer was verbally reported to the study administrator.

One important decision we made was to treat all task questions as independent tasks. We felt that although some task questions were similar at the high level, they were fundamentally different when carefully considered. Different questions asked the participants to search for different number of objects or annotations, and they also varied in that they asked for virtual or physical pieces of information in the scene. Some questions required the participants to simply look for virtual icons, or look for simple real-world objects, or look for objects or annotations in a more roundabout way (involving comprehension). These differences made each task question fundamentally different, which is reflected in our analysis.

The experiment consisted of 16 task questions. We did not random-

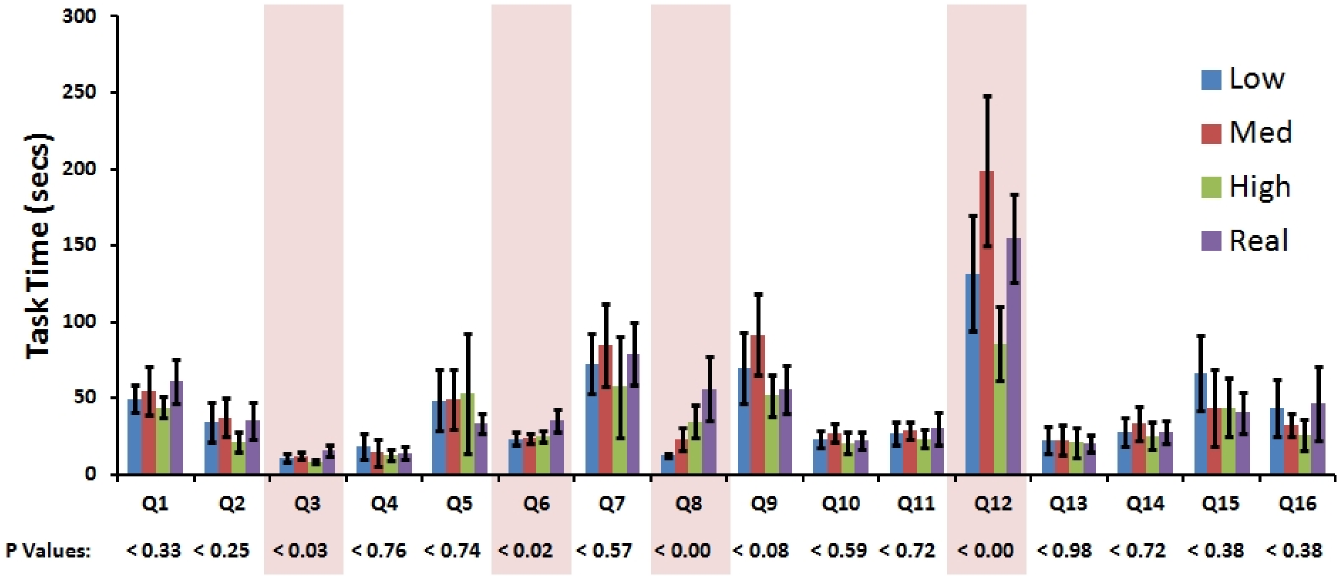


Fig. 4. A plot of the mean task times, 95 % confidence intervals, and significance P values for each level of realism for all 16 task questions. The P value for each task question was calculated using a between-subjects ANOVA with level of realism as the independent variable. Each task question was treated as a different test so that 16 independent ANOVAs were conducted. Significant differences were found between conditions in task questions Q3, Q6, Q8, and Q12 that are highlighted in red.

ize the order of the questions, and instead the questions were ordered based on their complexity. The questions that asked users to find a single item in the scene were presented first (Q1-Q5), followed by the questions that asked for multiple items (Q6-Q11), and finally the questions that asked for comprehension answers (Q12-Q16). There was no particular order to the task questions within each of the three groups of questions. Although we can assume there are learning effects, we minimized this by asking for different search targets throughout the questions. After some pilot trials, we broke up the 16 questions to create three sessions, to allow for breaks. This was designed to make the three sessions take approximately the same amount of time, and to last approximately 45 minutes.

4.4 Procedure

When each participant arrived, they were given a brief explanation of the experiment, both verbally and in text. Then a color vision test was administered to confirm that all users were able to perceive colors normally using Ishihara color palettes. Any participants who failed the test or reported any uncorrected vision problems were excused at this point. Those who passed were then given the demographic questionnaire.

After this questionnaire, the participants were then fitted with the HMD. A brief training session was then administered with some sample task questions. These sample questions were designed to force the participants to look for all four different icons in the scene. During this time, the study administrator would explain what each icon represented and would also train the participant on how to access the information contained in each icon. The administrator would also explain the rules of the experiment:

- Users must face the same direction before each task question.
- Users must verbally read out the question to the study administrator.
- Users must wait for the study administrator to repeat the question verbally to the user.
- Users must verbally confirm their understanding of the objective of the question to the study administrator.
- Users must then wait for the start countdown to finish before searching, upon which the timer for the task will begin.
- Users must be sure of the answer to the question before reporting it, but must also attempt to be as fast as possible. The timer for the task will be stopped by the study administrator once the correct answer is verbally reported.
- The study administrator will inform the users of any incorrect answers and the users must successfully complete a task question before the next one can begin.
- Users must verbally ask the study administrator to repeat the question if needed during the trials.

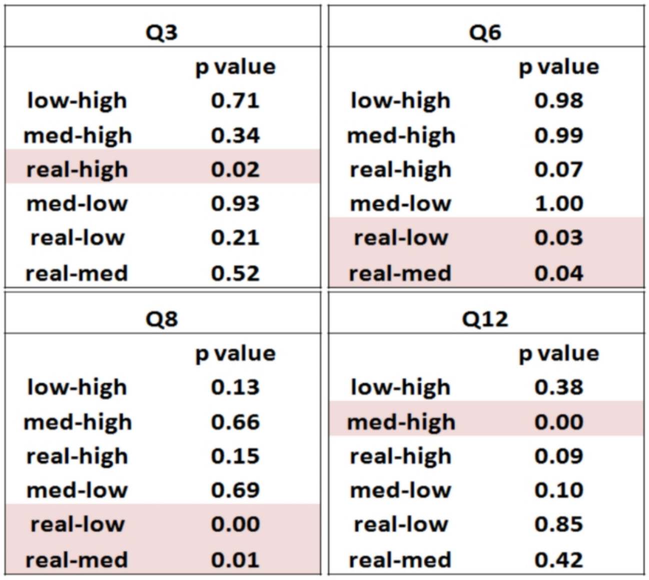
After this training session, a forced break was introduced where the user had to take off the HMD for two minutes at minimum. Once the break was over, the three timed sessions began. The 16 task questions were given during the course of all three timed sessions. Between each of these sessions the participants were forced to take a two minute break. At the start of each session, the participants were also briefly reminded of the objectives of the task, which was accurate and fast completion. Participants were also reminded to verbally ask for clarification or repeat of questions since we had observed in our pilot trials that some users forgot the questions in the more difficult tasks.

A large difference between the real and the low, medium, high conditions was that the user could see the environment in-between the breaks. We did not blindfold the participants in the real condition since we were in a public area and felt that the users would not be comfortable. This may have given a slight advantage to the participants in the real condition, but we felt it would not be substantial as we could not control their prior familiarity with the location.

Once the timed sessions were complete, the participants were then given a post-questionnaire. This questionnaire asked users to respond with their physical comfort levels during and after the experiment. And it also contained a sub-set of the WS presence-questionnaire. We took the 10 questions that we felt were pertinent. After completing this questionnaire the users were then thanked, paid, and excused from the experiment.

4.5 Participants

Using a paid-subject pool through our university, we obtained 54 paid participants. Of these participants, 48 were used in the final data, 12 for each condition. Six of the 54 participants were not counted due to errors in the system during the experiment or from not passing the



Q3	p value
low-high	0.71
med-high	0.34
real-high	0.02
med-low	0.93
real-low	0.21
real-med	0.52

Q6	p value
low-high	0.98
med-high	0.99
real-high	0.07
med-low	1.00
real-low	0.03
real-med	0.04

Q8	p value
low-high	0.13
med-high	0.66
real-high	0.15
med-low	0.69
real-low	0.00
real-med	0.01

Q12	p value
low-high	0.38
med-high	0.00
real-high	0.09
med-low	0.10
real-low	0.85
real-med	0.42

Fig. 5. A Tukey post-hoc analysis of Q3, Q6, Q8, and Q12. Pair-wise significant differences between each level of visual realism are highlighted in red.

color vision test. There were 20 males and 28 females, ranging from 18 to 40 years old (21.18 average). The genders were evenly spread out among the conditions, with five males and seven females for each condition. We used a pre-questionnaire to obtain demographic information regarding our test subjects. Collectively the users did not have much experience with VR or AR and their experience with video games was relatively high, with most having previous experience with the media. This was expected, as the participants came from a pool of mainly undergraduate students. All of the 48 recorded participants also reported normal or corrected to normal vision and were tested for color blindness via color palettes.

5 Issues with Replicating an Outdoor Environment

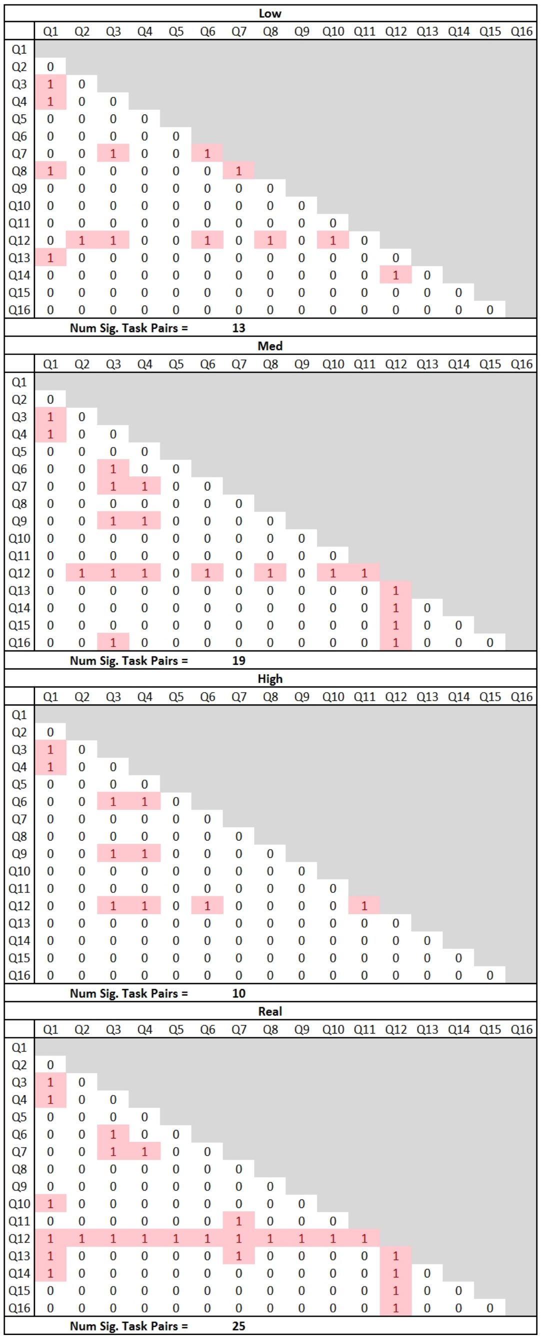
During the course of this experiment, there were two major difficulties we faced when we conducted the real AR part of the experiment. The first of these issues is with regards to the accuracy of our models to the real world. Even with the extensive amounts of data and measurements we obtained, we were not able to reproduce the real scene to complete accuracy. Within the six months that it took to complete the model, the physical location had changed. Chairs and tables had been added to the location and the vegetation had noticeably grown. One issue we did not foresee was that unlike the CAD plans, the physical buildings were not constructed on a flat plane, but were intentionally raised for irrigation purposes. This resulted in buildings that were at slightly different base heights compared to the CAD plans. In the end, what we produced were three virtual models of a real-world scene that were subjectively similar to the real location but contained geometric differences.

These differences in turn forced us to modify the placement of the virtual icons for the real conditions. Our models in the simulated conditions had assumed a flat ground, and the icons had been placed using these models. The ground in the real world was not flat but was curved slightly. Thus the modeled elevation of buildings and ground slopes varied slightly from reality. Due to time constraints, we decided that it was better to have accurate registration of the icons and so we manually shifted the icons for only the real condition.

Conditions	Task ID	dF	Epsilon	P Value
Low vs. Medium	Q1	17.54	28.76	0.0158
Low vs. High	Q1	20.64	28.76	0.0008
Low vs. Real	Q1	18.58	28.76	0.0446
Medium vs. High	Q1	15.13	28.76	0.0387
Medium vs. Real	Q1	21.80	28.76	0.0393
Low vs. Medium	Q2	21.99	23.89	0.0248
Low vs. Real	Q2	21.84	23.89	0.0141
Medium vs. Real	Q2	21.91	23.89	0.0176
Low vs. Medium	Q3	21.07	7.67	0.0046
Low vs. High	Q3	14.96	7.67	0.0063
Medium vs. High	Q3	16.80	7.67	0.0115
Medium vs. Real	Q3	18.84	7.67	0.0435
High vs. Real	Q4	21.91	8.89	0.0179
Low vs. Medium	Q6	20.67	14.96	0.0001
Low vs. High	Q6	21.51	14.96	0.0002
Medium vs. High	Q6	21.76	14.96	0.0000
Low vs. Real	Q7	21.99	40.44	0.0219
Medium vs. Real	Q7	20.37	40.44	0.0396
Low vs. Medium	Q8	12.15	42.52	0.0000
Low vs. High	Q8	11.50	42.52	0.0037
Medium vs. High	Q8	19.06	42.52	0.0002
High vs. Real	Q9	21.20	32.26	0.0120
Low vs. Real	Q10	22.00	10.85	0.0157
High vs. Real	Q10	21.00	10.85	0.0339
Low vs. Medium	Q11	20.86	21.66	0.0005
Low vs. High	Q11	21.46	21.66	0.0011
Low vs. Real	Q11	19.18	21.66	0.0105
Medium vs. High	Q11	21.87	21.66	0.0009
Medium vs. Real	Q11	16.65	21.66	0.0045
High vs. Real	Q11	17.44	21.66	0.0226
Low vs. Real	Q14	20.66	14.73	0.0214
High vs. Real	Q14	21.25	14.73	0.0366
High vs. Real	Q15	19.99	27.25	0.0397
Low vs. Medium	Q16	13.96	48.50	0.0027
Low vs. High	Q16	16.75	48.50	0.0100
Low vs. Real	Q16	20.85	48.50	0.0070
Medium vs. High	Q16	19.65	48.50	0.0000
Medium vs. Real	Q16	12.86	48.50	0.0141
Highvs. Real	Q16	14.73	48.50	0.0353

Table 2. Results from TOST analysis for equivalent realism conditions across all 16 task questions. Alpha is 0.05 and Epsilon values are set as the magnitude of the confidence interval of the Real condition.

The other major difficulty we faced was with illumination in the real condition. During the real AR part of the experiment, the weather was sunny with some cloud coverage. We were using a PointGrey USB3 Flea camera, and wanted to reduce the latency as much as possible. With the resolution (1600x1200 pixels) and frame rate (60 fps) we could not adjust the brightness of the camera using software without affecting performance. We could only change the aperture of the camera. This resulted in a very different experience between the participants in the real condition, depending on the time of the day. The image quality was generally better in the morning when there was more cloud and fog coverage, and worse during the afternoon where the sun was brightest. We attempted to manage this by adjusting the aperture



Low																
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16
Q1																
Q2	0															
Q3	1	0														
Q4	1	0	0													
Q5	0	0	0	0												
Q6	0	0	0	0	0											
Q7	0	0	1	0	0	1										
Q8	1	0	0	0	0	0	1									
Q9	0	0	0	0	0	0	0	0								
Q10	0	0	0	0	0	0	0	0	0							
Q11	0	0	0	0	0	0	0	0	0	0						
Q12	0	1	1	0	0	1	0	1	0	1	0					
Q13	1	0	0	0	0	0	0	0	0	0	0	0				
Q14	0	0	0	0	0	0	0	0	0	0	0	1	0			
Q15	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Q16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Num Sig. Task Pairs =							13									

Med																
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16
Q1																
Q2	0															
Q3	1	0														
Q4	1	0	0													
Q5	0	0	0	0												
Q6	0	0	1	0	0											
Q7	0	0	1	1	0	0										
Q8	0	0	0	0	0	0	0									
Q9	0	0	1	1	0	0	0	0								
Q10	0	0	0	0	0	0	0	0	0							
Q11	0	0	0	0	0	0	0	0	0	0						
Q12	0	1	1	1	0	1	0	1	0	1	1					
Q13	0	0	0	0	0	0	0	0	0	0	0	1				
Q14	0	0	0	0	0	0	0	0	0	0	0	1	0			
Q15	0	0	0	0	0	0	0	0	0	0	0	1	0	0		
Q16	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	
Num Sig. Task Pairs =							19									

High																
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16
Q1																
Q2	0															
Q3	1	0														
Q4	1	0	0													
Q5	0	0	0	0												
Q6	0	0	1	1	0											
Q7	0	0	0	0	0	0										
Q8	0	0	0	0	0	0	0									
Q9	0	0	1	1	0	0	0	0								
Q10	0	0	0	0	0	0	0	0	0							
Q11	0	0	0	0	0	0	0	0	0	0						
Q12	0	0	1	1	0	1	0	0	0	0	1					
Q13	0	0	0	0	0	0	0	0	0	0	0	0				
Q14	0	0	0	0	0	0	0	0	0	0	0	0	0			
Q15	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Q16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Num Sig. Task Pairs =							10									

Real																
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16
Q1																
Q2	0															
Q3	1	0														
Q4	1	0	0													
Q5	0	0	0	0												
Q6	0	0	1	0	0											
Q7	0	0	1	1	0	0										
Q8	0	0	0	0	0	0	0									
Q9	0	0	0	0	0	0	0	0								
Q10	1	0	0	0	0	0	0	0	0							
Q11	0	0	0	0	0	0	1	0	0	0						
Q12	1	1	1	1	1	1	1	1	1	1	1					
Q13	1	0	0	0	0	0	1	0	0	0	0	1				
Q14	1	0	0	0	0	0	0	0	0	0	0	1	0			
Q15	0	0	0	0	0	0	0	0	0	0	0	1	0	0		
Q16	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	
Num Sig. Task Pairs =							25									

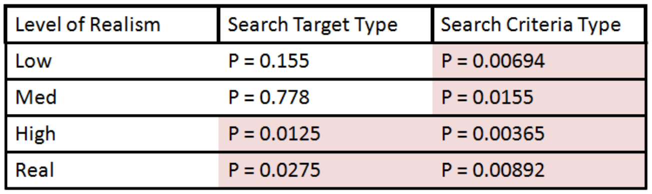
Fig. 6. A post-hoc analysis of the differences between task pairs within each level of realism. A paired t-test using Bonferroni correction was used to determine significance at the 95% level. Task pairs that showed a difference are marked by a 1 and shaded in light red. As seen, Q12 is most different from all other task questions.

after every session, but it was difficult in some cases. The right image in Figure 2 gives a reasonably typical impression of some of the lighting challenges in the real AR case.

6 Results

Differences in Task Times

For the analysis of the time results from our experiment, we used an ANOVA test to determine if the level of realism had a significant effect on task time. The main results of this analysis for each task question can be seen in Figure 4 and a post-hoc analysis of all task pairs can be seen in Figure 6. As can be seen in Figure 4, only four of the 16 task questions revealed a significant difference between the realism conditions. As we stated earlier, we treated each task question as a separate test and so 16 ANOVAs were conducted. The resulting P values are shown at the bottom of each bar plot. The four task questions for that a significant effect of level of realism on task was demonstrated were Q3(F (3,44) = 3.271, p = 0.0299), Q6(F(3,44) = 3.837, p = 0.0159), Q8(F(3,44) = 7.066, p = 0.0006), and Q12(F(3,44) = 5.455, p = 0.0028).



Level of Realism	Search Target Type	Search Criteria Type
Low	P = 0.155	P = 0.00694
Med	P = 0.778	P = 0.0155
High	P = 0.0125	P = 0.00365
Real	P = 0.0275	P = 0.00892

Fig. 7. A repeated measures ANOVA of search target type and level of realism, and search criteria type and level of realism. With search target type, the high and real conditions reveal similar significant effects. With search criteria, all four levels of realism reveal similar significant values.

For the four task questions that revealed a significant difference, we used a Tukey posthoc analysis to determine the pairwise differences. These results are shown in Figure 5. As these results show, only the real condition was found to be different in Q3, Q6, and Q8. Only Q12 reveals a difference between any two simulated conditions (low, medium, high). The results of Q3, Q6, and Q8 may reflect the difficulties we had in conducting the real-world AR portion of the experiment. Each of these questions also gave a piece of physical information as the search criteria. It was the door under the bridge for Q3, a tree for Q6, and a trash bin for Q8. In the low, medium, and high conditions, these objects were easily seen but in the real conditions this was more difficult. Camera artifacts, vegetation, and lighting conditions, as seen in Figure 2, affected the visibility of the real-world objects much greater than the virtual objects. It is reasonable to expect that this would have increased task time when the search criteria was something physical and relatively small.

Equivalence in Task Times

All other tasks did not reveal a significant difference for level of realism. To investigate this further, we used the two-one-sided t-test (TOST) analysis to look for equivalent task time performance within our groups. For each task question, we ran the analysis with an alpha value of 0.05 and used the magnitude of the confidence interval of the high condition task time as the epsilon or magnitude of equivalence region. The conditions that show statistical equivalence under these parameters are shown in Table 2. Using the confidence interval may cause the epsilon value to be overly generous (as seen for Q7 and Q16), but without any prior experience with these tasks it is not possible to determine a region of equivalence correctly. Since our sample size of 12 per group is too small to make any conclusive claims regarding similarity, the confidence interval suffices as a first attempt. We used the confidence interval of the high condition since the high condition represents the ground truth data; even when considering the difficulties we had in the real condition.

Under this premise, only Q12 and Q13 did not show a case where at least two realism conditions were statistically equivalent. All other task questions produced at least one instance where two of the realism conditions were equivalent. Counting the number of times each realism condition appears in Table 2, all realism levels are represented equally across all the equivalent pairs (20 low, 19 medium, 19 high, 20 real).

Notable Task Questions

The results for Q8 are interesting; it is the only task with results indicating any obvious performance trend with respect to task time. The plot in Figure 4 shows task time increasing as we progressively increase the level of realism. We believe this result is indicative of visual clutter and the visibility of the trash bins as the level of realism was increased. As seen in Figure 1, the environment became more cluttered as the level of realism increased. There were three trash bins in the scene, and they were placed next to

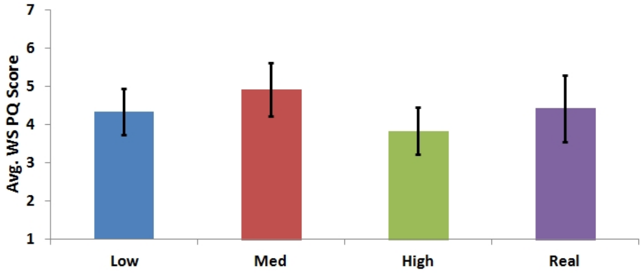


Fig. 8. The mean scores from the Witmer-Singer presence questionnaire with standard error. A score of 1 represents a low sense of presence and a score of 7 represents a high sense of presence.

vegetation. As more detail was added to the environments, it became more difficult to find these small objects. This may explain why the task time increased gradually for each successive level of realism.

Q12 is also interesting. This question was by far the most difficult and the results shown in Figure 4 and Figure 6 demonstrate this. Figure 6 plots all the significant differences between each task question pair within each realism condition. Using Figure 6, it is easy to see that Q12 was the most different from other task questions, regardless of level of realism. Q12 asked the participants to identify the professor with the office that was located the furthest from his lab. Participants needed to find all the professors and their respective labs, and spatially sort out the correct professor. Although the answer was obvious, since the correct professor's office was located in a far corner, most participants had a difficult time. We observed that a large number of participants (across all the conditions) became stuck in a loop, and kept searching in the same area (seen in Figure 1). This particular area contained a large group of the professors' offices and labs. Those who used a systematic strategy or happened to see the office and lab early were able to complete the task quicker. It is for this reason, that we believe the results here reflect the difficulty of the task and the effects from strategy more than from level of realism.

Task Type and Visual Realism We categorized our task questions into two types, as seen in Table 1, depending on their search target and search criteria type: physical information or virtual information. Physical information refers to information that would be present in the natural world, regardless of the level of realism. Examples are physical objects such as trees and spatial relationships such as the floor number of a building. On the other hand, virtual information has to be obtained by looking at the virtual icons or the text contained in those icons. Examples are 'course name', 'type of tree', 'lab owner', etc. In summary, tasks vary in two major ways, by search target type and search criteria type. Each of these types can be virtual or physical.

Using search target type as the within subjects variable and level of visual realism as the between subjects variable, we find an overall significant effect on task time (F(1,3) = 7.450, p = 0.009) with no interaction between search target type and level of realism. Generally, participants were quicker to respond when the search target type was virtual information. Using search criteria as the within subjects variable and level of realism again as the between subjects variable, we also find a significant effect (F(1,3) = 37.760, p = 0.001). Here, the participants were quicker to respond when the search criteria was a physical piece of information.

An interesting result appears when we look at the effect of search target type and search criteria type within each level of realism. These results can be seen in Figure 7. Search criteria type is shown to be significant in all four conditions while search target type has a similar effect on the high and the real conditions. This suggests that the high level of realism performs similarly to the real conditions. The similarity of these effects support the validity of MR Simulation. Search criteria does indeed have an effect on task time and this effect is the same across all four levels of visual realism. The effects of search target type are more difficult to explain. It does not have an effect on the results from the low and medium conditions but has a clear effect on the high and real conditions.

By WS Presence Questionnaire Results The overall results of the WS questionnaire can be seen in Figure 8. The average score for all four conditions were relatively high. Using the Kruskal-Wallis rank sum test, we did not find an effect of visual realism on presence (chi-squared = 3.80, dF = 3, p = 0.28). This result was not unexpected as it reflects findings in previous work [15]. In that experiment, Slater hypothesizes that using questionnaires to measure presence between different environments may not be valid. As explained, users who have never experienced such environments will invariably interpret these questions differently.

For the most part, all of our users had very little experience with VR and AR. Most expressed delight and amazement when they first put on the HMD and walked around, regardless of the condition they were in. We received many verbal comments with respect to this. Our system was able to provide very reliable tracking for the users and this was the first time most of the users had experienced an immersive display system such as ours. This may have been why the conditions were rated similarly in terms of presence.

7 Discussion

We stated two goals, when we discussed the motivation for our experiments. The first goal was to determine whether visual realism had an effect on the validity of MR Simulation. The second corollary goal was to determine whether visual realism had an effect on tasks performed in VR. With respect to the task time results, we found a significant difference between the conditions in four instances: Q3, Q6, Q8, and Q12. For Q12, we argued that the differences were more due to strategy and the difficulty of the task. In the other cases we argued that these differences were due to camera artifacts, vegetation, and lighting conditions on the physical objects the participant needed to find. In particular, the lighting conditions made these objects significantly harder to find in high and low contrast lighting areas. The results from the equivalence tests did not reveal any trends pertaining to the equivalence of one condition to another with respect to level of realism in general. A more precise task needs to be tested to further study of this.

It is also interesting that differences in task performance were limited to questions that involved search for physical objects, which leads us to believe that differences in the clarity with which such objects could be perceived were at the root of the differences (i.e., because of lighting and camera artifacts it was actually more difficult to perceive and work with real objects in the "real AR" case). Our observations and the fact that we found differences between the real and virtual conditions in task performance on a small subset of questions indicates there were significant differences between our real AR experiment and its MR Simulation incarnations. We would like to determine the exact nature of these differences in follow-up work.

For our second goal, which was to determine the effect of visual realism on search tasks performed in VR, the results are consistent. Among the 16 task questions, it was only in Q12 that a significant difference was found between the simulated conditions. We argued that this was mainly due to the difficulty of the task. We observed that strategy and luck allowed some users to complete the task quicker, while some became stuck in a search pattern in an area with multiple virtual and physical objects. As for the rest of the tasks, we did not find any significant differences. Although we did not find any conclusive evidence to suggest that visual realism does not have an effect on this task in VR, this at least opens the discussion.

As for presence, the WS presence questionnaire did not reveal any significant differences between the levels of visual realism. Our results are similar to the results from Slater et al. [15]. This highlights the difficulty in using questionnaires to compare level of presence between different environments. Our observations of the users indicated a similar response across all levels of visual realism. Repeated comments from various users indicated how "cool" the experiment was. Since a user only saw one condition, this is an expected result from participants who have very little experience with virtual environments.

Perhaps the most important lesson we learned in this experiment was the difficulty we had in conducting the real AR portion of our experiment. Part of the motivation for MR Simulation is to enable researchers to conduct controlled experiments in AR. Our difficulties highlight this motivation again. Even with the care and preparations we made, we were confounded at times by the real world. The results from the real AR must be interpreted with that understanding. In one sense, the difficulties we had with the real-world condition actually support MR Simulation. By any measurement, we did our due diligence to obtain the very best model we could within our time and money constraints. We spent over six months collecting data, building plans, and physical measurements to construct our models. And yet, the natural changes in the environment and the dynamic lighting in the real-world condition proved very difficult to manage. In the end, we were forced to modify the placement of the icon annotations to fit the real world. We were also forced to make constant corrections to our camera to retain decent overall lighting. This points to the fact that even with current technology, it is very difficult to perform controlled AR experiments in outdoor environments, and the results from such experiments must take this into consideration.

8 Conclusions and Future Work

We have conducted a user experiment investigating the effects of a scene's visual realism on task performance in simulated AR environments and compared the results with a true AR experiment in the corresponding real-world environment. We considered 16 different search tasks, representative for information browsing and comprehension in AR. Among the simulation conditions, we only found one task question (Q12) that resulted in significant timing differences, which involved complicated reasoning and was by far the most difficult task. The real AR portion revealed differences in task performance in three task questions (Q3, Q6, Q8), which could either be due to the lighting artifacts in the high-dynamic range scene or actual differences in scene realism. Overall, these minimal differences are promising for the validity of MR Simulation, and the difficulties of conducting outdoor real-world experiments demonstrate the usefulness of MR Simulation.

In future work, we would like to clarify through follow-up studies what the exact cause of the differences between the AR case and the simulated AR cases was. Since it became very clear to us that high-dynamic range lighting coupled with outdoor HMD usage had a detrimental effect on distinguishing small physical objects in the real scene, we will look at better ways to present AR imagery to users, i.e. at improving the actual AR experience. Maybe an interesting intermediary simulated AR condition could also be using a (well-lit and adjusted) panoramic image backdrop with the same annotations as in this study, sacrificing motion parallax and real sensory input (being situated in an actual outdoor environment) for better image visibility and experimental control. This would give evidence if the differences in task performance were indeed due to lighting and resolution differences or due to the modeling abstractions.

More generally, we are continuing with exploring Mixed Reality Simulation for different task types (not just search, but also browsing and annotation and interaction tasks) and for different training environments to simulate.

Acknowledgments

This work was supported in part by Office of Naval Research (ONR) grant N00014-09-1-1113 and US National Science Foundation (NSF) CAREER grant IIS-0747520. We would also like to thank Dr. Bodo Bookhagen for his assistance in acquiring the initial LIDAR data.

References

[1] D. A. Bowman, C. North, J. Chen, N. F. Polys, P. S. Pyla, and U. Yilmaz. Information-rich virtual environments: theory, tools, and research agenda. In *Proceedings of the ACM symposium on Virtual reality software and technology*, VRST '03, pages 81–90, New York, NY, USA, 2003. ACM.

[2] M. Elhelw, M. Nicolaou, A. Chung, G.-Z. Yang, and M. S. Atkins. A gaze-based study for investigating the perception of visual realism in simulated scenes. *ACM Trans. Appl. Percept.*, 5(1):3:1–3:20, Jan. 2008.

[3] J. L. Gabbard, J. E. Swan, II, and D. Hix. The effects of text drawing styles, background textures, and natural lighting on text legibility in outdoor augmented reality. *Presence: Teleoper. Virtual Environ.*, 15(1):16–32, 2006.

[4] P. Khanna, I. Yu, J. Mortensen, and M. Slater. Presence in response to dynamic visual realism: a preliminary report of an experiment study. In *Proceedings of the ACM symposium on Virtual reality software and technology*, VRST '06, pages 364–367, New York, NY, USA, 2006. ACM.

[5] M. Knecht, A. Dünser, C. Traxler, M. Wimmer, and R. Grasset. A framework for perceptual studies in photorealistic augmented reality. In P. W. Frank Steinicke, editor, *Proceedings of the 3rd IEEE VR 2011 Workshop on Perceptual Illusions in Virtual Environments*, pages 27–32, Mar. 2011.

[6] C. Lee, S. Bonebrake, T. Hollerer, and D. A. Bowman. A replication study testing the validity of ar simulation in vr for controlled experiments. In *Proceedings of the 2009 8th IEEE International Symposium on Mixed and Augmented Reality*, ISMAR '09, pages 203–204, Washington, DC, USA, 2009. IEEE Computer Society.

[7] C. Lee, S. Bonebrake, T. Hollerer, and D. A. Bowman. The role of latency in the validity of ar simulation. In *Proceedings of the 2010 IEEE Virtual Reality Conference*, VR '10, pages 11–18, Washington, DC, USA, 2010. IEEE Computer Society.

[8] C. Lee, S. Gauglitz, T. Hollerer, and D. A. Bowman. Examining the equivalence of simulated and real ar on a visual following and identification task. In *Proceedings of the 2012 IEEE Virtual Reality*, VR '12, pages 77–78, Washington, DC, USA, 2012. IEEE Computer Society.

[9] K. Mania, S. Badariah, M. Coxon, and P. Watten. Cognitive transfer of spatial awareness states from immersive virtual environments to reality. *ACM Trans. Appl. Percept.*, 7(2):9:1–9:14, Feb. 2010.

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

[11] P. Rademacher, J. Lengyel, E. Cutrell, and T. Whitted. Measuring the perception of visual realism in images. In *Proceedings of the 12th Eurographics Workshop on Rendering Techniques*, pages 235–248, London, UK, UK, 2001. Springer-Verlag.

[12] E. Ragan, C. Wilkes, D. A. Bowman, and T. Höllerer. Simulation of augmented reality systems in purely virtual environments. *Virtual Reality Conference, IEEE*, 0:287–288, 2009.

[13] M. Slater. Measuring presence: A response to the witmer and singer presence questionnaire. *Presence: Teleoper. Virtual Environ.*, 8(5):560–565, Oct. 1999.

[14] M. Slater. A note on presence terminology. In *In Presence-Connect*, 2003.

[15] M. Slater, P. Khanna, J. Mortensen, and I. Yu. Visual realism enhances realistic response in an immersive virtual environment. *IEEE Comput. Graph. Appl.*, 29(3):76–84, May 2009.

[16] C. Stinson, R. Kopper, B. Scerbo, E. Ragan, and D. Bowman. The effects of visual realism on training transfer in immersive virtual environments. *Human Systems Integration Symposium*, 2011.

[17] N. Sugano, H. Kato, and K. Tachibana. The effects of shadow representation of virtual objects in augmented reality. In *Proceedings of the 2nd IEEE/ACM International Symposium on Mixed and Augmented Reality*, ISMAR '03, pages 76–, Washington, DC, USA, 2003. IEEE Computer Society.

[18] V. Vinayagamoorthy, A. Brogni, M. Gillies, M. Slater, and A. Steed. An investigation of presence response across variations in visual realism. In *In Proceedings of Presence 2004: The 7th Annual International Workshop on Presence. Available online at: http://www.cs.ucl.ac.uk/research/equator/papers/VisualRealism.pdf*, pages 119–126, 2004.

[19] N. Wang and W. Doube. How real is really? a perceptually motivated system for quantifying visual realism in digital images. In *Proceedings of the 2011 International Conference on Multimedia and Signal Processing - Volume 02*, CMSP '11, pages 141–149, Washington, DC, USA, 2011. IEEE Computer Society.

[20] B. G. Witmer and M. J. Singer. Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoper. Virtual Environ.*, 7(3):225–240, June 1998.