Fusing Self-Reported and Sensor Data from Mixed-Reality Training

Trevor Richardson, Stephen Gilbert, Joseph Holub, Frederick Thompson, Anastacia MacAllister, Rafael Radkowski, Eliot Winer
Iowa State University
Ames, Iowa
trevorr@iastate.edu, gilbert@iastate.edu, jholub@iastate.edu, fvt@iastate.edu, anastac@iastate.edu, rafael@iastate.edu, ewiner@iastate.edu

Paul Davies, Scott Terry
The Boeing Company
St. Louis, MO
paul.r.davies@boeing.com, scott.a.terry@boeing.com

ABSTRACT

Military and industrial use of smaller, more accurate sensors are allowing increasing amounts of data to be acquired at diminishing costs during training. Traditional human subject testing often collects qualitative data from participants through self-reported questionnaires. This qualitative information is valuable but often incomplete to assess training outcomes. Quantitative information such as motion tracking data, communication frequency, and heart rate can offer the missing pieces in training outcome assessment. The successful fusion and analysis of qualitative and quantitative information sources is necessary for collaborative, mixed-reality, and augmented-reality training to reach its full potential. The challenge is determining a reliable framework combining these multiple types of data.

Methods were developed to analyze data acquired during a formal user study assessing the use of augmented reality as a delivery mechanism for digital work instructions. A between-subjects experiment was conducted to analyze the use of a desktop computer, mobile tablet, or mobile tablet with augmented reality as a delivery method of these instructions. Study participants were asked to complete a multi-step technical assembly. Participants' head position and orientation were tracked using an infrared tracking system. User interaction in the form of interface button presses was recorded and time stamped on each step of the assembly. A trained observer took notes on task performance during the study through a set of camera views that recorded the work area. Finally, participants each completed pre and post-surveys involving self-reported evaluation.

The combination of quantitative and qualitative data revealed trends in the data such as the most difficult tasks across each device, which would have been impossible to determine from self-reporting alone. This paper describes the methods developed to fuse the qualitative data with quantified measurements recorded during the study.

ABOUT THE AUTHORS

Trevor Richardson, is a graduate student in Computer Engineering and Human Computer Interaction at Iowa State University and the Virtual Reality Applications Center.

Stephen Gilbert, Ph.D., is an associate director of the Virtual Reality Applications Center and assistant professor of Industrial and Manufacturing Systems Engineering at Iowa State University. His background includes cognitive science and engineering, and he supervises Iowa State's graduate program in Human Computer Interaction. He is PI on a Live, Virtual and Constructive training contract for the U.S. Army Research Laboratory STTC.

Joseph Holub, is a graduate student in Computer Engineering and Human Computer Interaction at Iowa State University and the Virtual Reality Applications Center.

Frederick Thompson, is a graduate student in Mechanical Engineering and Human Computer Interaction at Iowa State University and the Virtual Reality Applications Center.

Anastacia MacAllister, is a graduate student in Mechanical Engineering and Human Computer Interaction at Iowa State University and the Virtual Reality Applications Center.

Rafael Radkowski, is an Assistant Professor of Mechanical Engineering. He specializes in the design and deployment of Augmented Reality systems both in hardware and software.

Eliot Winer, Ph.D., is an associate director of the Virtual Reality Applications Center and associate professor of Mechanical Engineering at Iowa State University. He is currently co-leading an effort to develop a next-generation mixed-reality virtual and constructive training environment for the U.S. Army. Dr. Winer has over 15 years of experience working in virtual reality and 3D computer graphics technologies.

Paul Davies, is an electrical engineer specializing in digital signal processing, and works in the Production Systems Technology group in Boeing Research & Technology. Since joining Boeing in 2003 he has supported the Advanced Tactical Laser, Homeland Security & Services, Delta II and B1B programs in addition to multiple IRAD and CRAD projects in Signal Processing, Augmented Reality and Machine Vision. He currently develops technology for Augmented Reality in manufacturing and investigates new methods of person-machine interaction for technician support. Paul received a BS degree in Electrical Engineering from Rochester Institute of Technology in May 2004, and a MS degree in Electrical Engineering from California State University Long Beach in May 2008

Scott Terry, PMP, is an Industrial Engineer working as the lead technology insertion focal for the Boeing Military Aircraft, Quality and Manufacturing Home Office. Scott joined Boeing in 2006 supporting the F-15 Program and transitioning to Weapons programs before moving in to his current role. Prior to Boeing, Scott worked as a Design Engineer and Process Engineer for Newell/Rubbermaid.

Fusing Self-Reported and Sensor Data from Mixed-Reality Training

Trevor Richardson, Stephen Gilbert, Joseph Holub, Frederick Thompson, Anastacia MacAllister, Rafael Radkowski, Eliot Winer
Iowa State University
Ames, Iowa
trevorr@iastate.edu, gilbert@iastate.edu, jholub@iastate.edu, fvt@iastate.edu, anastac@iastate.edu, rafael@iastate.edu, ewiner@iastate.edu

Paul Davies, Scott Terry
The Boeing Company
St. Louis, MO
paul.r.davies@boeing.com, scott.a.terry@boeing.com

INTRODUCTION

As resources continue to diminish in today's military, methods to effectively train the warfighter must be constantly evaluated for efficacy, retention, cost avoidance and other critical facets. The combined costs of conducting live training including the time necessary to complete the exercises, travel, reviews, evaluations, and other various activities can constitute a substantial expense. The ability to improve training by shortening training times represents an important key to the future success of training and simulation. Current methods for training evaluation such as pre and post training questionnaires, demographic data, and measures of performance, often do not provide trainers with all the necessary information or insight to improve a trainee's performance. In addition, current data points are difficult to integrate into a coherent profile of the trainee due in part to the variety of formats (e.g., paper and digital) with different fidelities (e.g., tracking data can be every second of training while post training questionnaires happen once). The goal is to fuse all of this data into a single coherent profile of the trainee to improve future training accuracy and retention. With this information, training can be completed more effectively using fewer resources.

Sensors for collecting data continue to become smaller and less expensive every year and it is now possible to outfit trainees with relatively low-cost sensors including GPS devices, accelerometers, radios, heart-rate monitors, and electrodermal sensors. These sensors can be combined with other systems-based measures (Orvis, Duchon, & DeCostanza, 2013) like text messages, emails, phone calls, etc., to provide a trainer with an immense amount of training data.

The combination of these low cost sensor and systems-based measures provide several benefits over traditional methods of data collection. For example, these methods are efficient due to their unobtrusive nature in recording the trainee with little effort from the observer. This dynamic eliminates many of the limitations of traditional data recording by a human observer who is subject to fatigue and distractions, resulting in missed observations. These quantitative systems-based collection methods are generally well received by the research community because they are perceived as "objective" measures of performance. The preference for quantitative measures stems from the dangers of rating errors such as "halo effect" which is a cognitive bias in which the overall impression of a person impacts the evaluation of that person's skills (Murphy & Balzer, 1989).

To avoid the use of qualitative evaluation entirely, however, would be to potentially lose out on valuable training insights. An expert can often record observations that are difficult if not impossible to capture with today's technology. It is more a question of how to best leverage both qualitative and quantitative data than whether to use one or the other. This raises important questions about how the military might approach training evaluation. This paper addresses the question of how to leverage new forms of training data by describing an Augmented Reality (AR) training study performed at Iowa State University in cooperation with The Boeing Company. The study evaluated trainees learning to assemble a "wing" comparing traditional model-based work instructions (MBI) with augmented reality work instructions.

Benefits using tracking sensors during training were realized from the data analysis and will be described. The tracking data provided redundancy for some measures as well as prompted new questions such as "How much time did each participant spend in each particular area of the work cell?" Answering these additional questions would have been impossible with traditional data collection techniques. The answers would have relied on what the

participants thought had happened (i.e. their perception of how long they spent in each area). The problem with this is that their perception and the actual timings were very different.

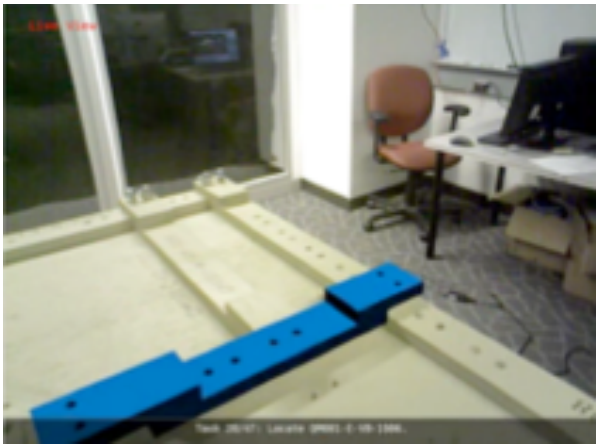

Figure 1: Augmented Reality blue part in assembly task.

BACKGROUND

Augmented Reality

Augmented Reality (AR) is the augmentation of the real world with digitally generated sensory inputs like visuals or sound. When applied to visuals, as was done for the current work, digital objects are registered spatially and rendered within the physical world often using a display device like a tablet or cell phone. Figure 1 shows an example of augmented reality as used in this work. In the figure, a blue digital object in the form of a 3D model is being rendered on screen in proper position as if it were in the physical world. In the early 1990s, Boeing had already begun looking into the use of augmented reality for manufacturing using a head-mounted display to superimpose design diagrams on real parts (Caudell et al., 1992). Numerous advancements in technology have further enhanced the feasibility of using AR. The use of AR for work instruction delivery has been an area showing possible benefit to industry (Nee et al., 2012). Research here has shown that reductions in both time and errors can be gained by using AR (Nakanishi et al., 2007). The study described in this paper was conducted to analyze augmented reality as a delivery method for work instructions in a manufacturing work space or cell.

Fusing Qualitative and Quantitative Data

Mixed methods research attempts to leverage the respective benefits of both qualitative and quantitative evaluation (Durham et al., 2011; Fielding, 2012; Hesse-Biber & Johnson, 2013). Kaplan and Duchon realized years ago the usefulness of leveraging both qualitative and quantitative evaluation techniques applied to systems evaluation (Kaplan & Duchon 1988). The authors were interested in the effects of new computer systems on such things as computer acceptance and the effect of technology on job characteristics and job satisfaction. Qualitative notes taken during the study were used to help resolve disagreements about the quantitative results. This led to a new direction for the study's quantitative analysis and subsequent results. An important fact this study brought to light was that one method might not always be enough to provide clear and accurate results of what actually occurred. Creswell suggests (Creswell, 2003) much research today lies on a continuum (Newman & Benz, 1998) between qualitative and quantitative methods, but often focuses more toward one method alone. A review of mobile HCI research methods in 2003 (Kjeldskov & Graham, 2003) showed the limited use of dual-method research and emphasized a need for more work.

Often there are a number of factors in a study that are not specifically measured, but would provide additional situational insight. Schneiderman proposes that one solution for finding this extra information can be found through what was named Multi-dimensional In-depth Long-term Case studies (MILCs) (Schneiderman & Plaisant, 2006). While long-term studies can certainly add to understanding, sometimes it is not an option due to limited resources. Fortunately, there are often copious amounts of untapped data in a single training study that are not currently being measured. For example, in computational systems research, measures of task time and task success are traditionally recorded (Lazar et al., 2010). Actions like mouse clicks, corresponding positions and related time offer data that is not often used but can present valuable insight in post-study analysis.

An area of growing use and a potential candidate for additional training data relates to wearable devices. Lukowicz et al. describe a large effort by the European Union to leverage wearables in the industrial workplace (Lukowicz et al, 2007). As these and other wearable computing devices become more widespread, and as measurement hardware becomes less expensive and more accurate, the important information lost by not leveraging these data will continue to grow. Orvis et al. discuss this developing issue and suggest a framework modeled after the biodata approach

(Mumford & Owens, 1987) by which to conduct research to better leverage untapped data they refer to as systems data (Orvis et al., 2013).

In an era of tightening fiscal resources with a goal of reducing resource constraints, anywhere that technology can aid in training and practice remains critical. However, simply applying technology is not enough as demonstrated by previous research highlighted in this section. This technology and related data must be properly evaluated as to its effectiveness. What this brief literature search shows is that traditional human subject qualitative data may not be enough to adequately evaluate training effectiveness. Fusing this with quantitative systems data can provide a much more comprehensive evaluation.

METHODOLOGY

This study was designed to evaluate three different methods of presenting work instructions. The three methods were referred to as: 1) Desktop MBI; 2) Tablet MBI; and 3) Tablet AR. The first mode was designed to mimic current work instructions using Model-Based Instructions (MBI) on a stationary display located in one corner of a work cell and not visible from the work area. Figure 2a shows the Desktop MBI mode showing one step in the assembly work instructions. The Tablet MBI mode utilized the exact same Model-Based Instructions as the Desktop MBI, but showed them to the trainee on a tablet PC mounted on a mobile arm device (tablet can be seen in Figure 2b). The third mode was the Tablet AR mode using the same tablet as used in the Tablet MBI mode, but presented the work instructions to the trainee using Augmented Reality. Figure 2b shows an image of a single step in the Tablet AR mode.

Figure 2a: Desktop MBI

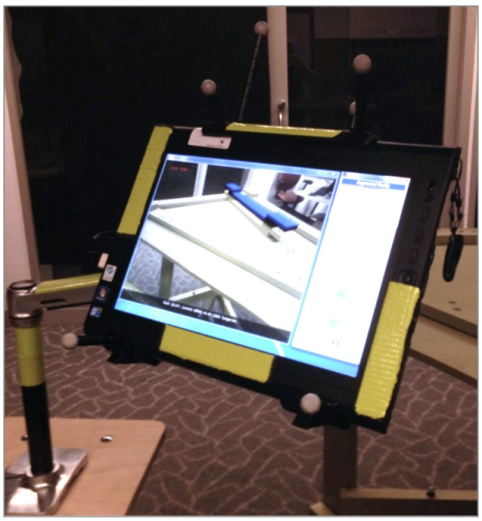

Figure 2b: Tablet AR

Study Hardware/Software Setup

The setup of the study was designed to mimic a traditional work cell for manufacturing where there is a designated assembly area along with specific areas where workers can find large parts and another for smaller parts or hardware items such as nuts and bolts. Figure 3 shows the layout of the training area. All assembly tasks were performed in the Wing location. The Wing was positioned at approximately four feet high for ergonomic purposes. All of the larger parts for the assembly were located on the Parts Table shown near the upper edge of the image. All of the smaller parts like nuts and bolts were located in labeled plastic bins in the Parts Bins location.

The Desktop MBI mode used a desktop computer facing toward the right in Figure 3. The study observer sat behind a desk in the area labeled "Observer" in the Figure 3. The observer recorded participant errors and times by hand on a paper chart throughout each study. All of the parts for the practice task were located on a separated table labeled "Practice Parts" in the figure.

The specific hardware selected for instructional delivery to each participant was dependent upon the mode each participant was assigned. The Desktop MBI mode used a commercially available Dell Precision Desktop computer paired with a twenty inch 3M LCD touch screen monitor. The touch screen monitor was used to mirror the touch screen behavior of the tablet interactions.

The Tablet MBI and Tablet AR modes both used a 12.1-inch Motion Computing tablet running an Intel Core i7 processor. The tablet was mounted on an Ergotron moveable arm designed for LCD monitors. The arm was then mounted on a small rolling base. This combination allowed participants to roll the tablet around the work cell and adjust the arm to achieve their ideal view.

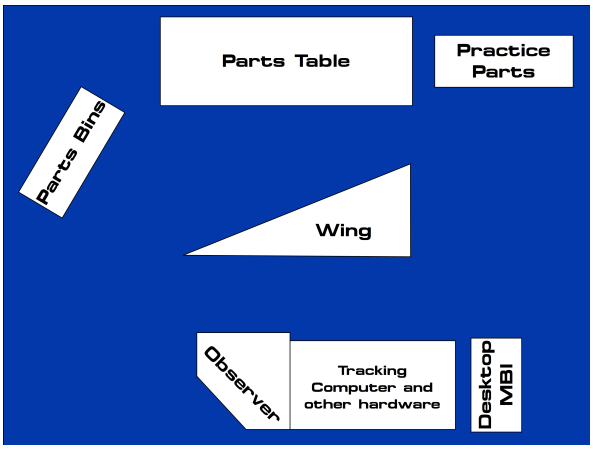

Figure 3: Top down view of the training area

The Desktop MBI and Tablet MBI modes used a custom graphical user interface application built using the Qt interface library. Screen mirroring software was run in the background to allow the observer to see a live view of what the trainee was seeing. This screen recording was additionally saved as data for potential use later in the evaluation.

The Tablet AR mode used a custom Augmented Reality application and user interface built by Boeing Research and Technology. The interface and AR elements were selected through collaboration with Iowa State University. The screen of the Tablet AR mode was observed and recorded throughout using the same screen-mirroring software noted for the Desktop MBI and Tablet MBI modes.

A high precision tracking system was required to accurately align the AR models with the real world. A four camera infrared Vicon tracking system was used for this purpose. The four-camera system allowed tracking of the entire work cell.

The entire study area was recorded on video using four webcams positioned around the work cell. All four webcam feeds were live streamed to the observers desktop computer which, when combined with the live screen capture from the tablet/desktop, provided the observer with different views of the work cell in the event that the view from the observation desk was blocked. This also provided the ability to review what each trainee did in later evaluation, in case of observer recording error.

Study Procedure

Each participant began with an Informed Consent Document and a pre-survey asking for demographic and experience data. After the initial background data was collected, the participant was introduced to his or her instructional mode. Each participant only experienced a single mode throughout his or her entire participation, a between-subjects experimental design.

The introduction to the instruction delivery mode was accomplished with a five-step assembly task similar to the experimental task the participant would be asked to complete. This practice task used all aspects of the full task, including all data collection tools. The only differences being that participants were able to ask the observer questions if they had any and the participants were informed they were not being graded on performance.

After successfully completing the practice assembly, the participant performed the full assembly task twice, with a Paper Folding test in between. The full task consisted of 46 different steps. The MBI instructions combined sets of these steps into a single page of instructions to better represent what is used on a typical factory floor. This resulted in 14 total steps for the MBI instructions. The steps ranged in complexity from selecting the correct parts, to properly aligning and fastening bolts through multiple parts. The most difficult task involved placing a part through

a complex rotation to achieve proper alignment. After completing the task a second time, participants were asked to fill out a short post-survey rating their performance on the tasks with several Lickert scale questions.

On average the entire study took less than two hours to perform all tasks. Each assembly task was capped at 45 minutes. This time was determined from pilot participants used to gauge the various aspects of the task. Only one participant reached the 45-minute time limit on a task.

Data Collection

The data collection methods were critical to this study given the two-hour length and large number of participants desired. Data redundancy permitted the study team to triangulate qualitative analysis results with that of the measured quantitative data. The tracking system and array of webcams played a large part in allowing this opportunity. While these were originally put into place to capture task accuracy and completion time, the resultant data was used for additional analysis including: 1) determining the amount of time spent looking at the instructions on the tablet and 2) the amount of time traveling between different work areas.

The study observer used traditional pen and paper data collection throughout each study. This was comprised of a pre and post-survey, a paper folding test, and all observational notes taken during the task. Task activities were recorded on a piece of paper gridded by five-second intervals with categories of participant activity to be indicated for each time block: reading instructions, gathering parts from bins, gathering parts from the parts table, assembling at the work, or fixing a mistake. This method allowed notes to be taken about a particular point in time for later comparison with the video recording if necessary. Notes were additionally made regarding assembly errors and instances when participants went back and fixed such an error before the end of the task.

The instructional interfaces also recorded data during the assembly task and every button click to move to the next step or back to the previous step was time stamped and recorded. This time stamp capture allowed post processing of the data to determine how much time was spent on individual steps. This also made it possible to determine how many times individual participants went backwards to check a previous step (each of which would have been difficult for an observer to make reliable notes on during the task in combination with the numerous other responsibilities).

Each participant was asked to wear a plastic helmet with reflective tracker balls that would allow an optical tracking system to accurately obtain his or her head position and orientation continually during the assembly tasks. The same reflective tracker balls were applied to the tablet case to allow its tracking (Figure 2). This system allowed the tracking of the position and orientation of both the tablet and the participant's head. A custom application was written to query the tracking system every 0.5 seconds for the raw tracking data to record.

After a participant completed the study, the recorded raw tracking data was fed into a custom post-processing application that converted the raw data into a usable form. It was very important to make the collection and processing steps independent so that the post-processing application could be updated to look for new trends in the data as the results were being processed. The selected method to handle the raw tracking data was to create a visual representation that was easily understandable. Figure 4 shows the post-processing software application displaying the results of a Tablet AR mode session. Each blue dot represents the participant at one time point and the black dots represent the tablet at a single time point. The black outlined boxes

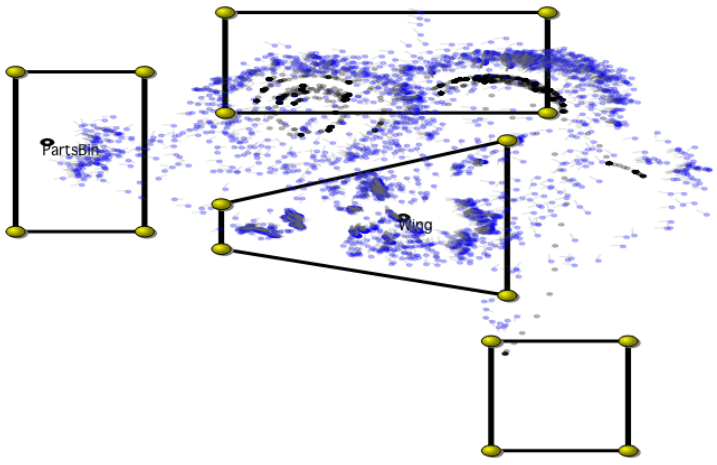

Figure 4: Post-processing of tracking data for Tablet AR mode.

represent areas of interest within the work cell. In this case the areas of interest are the Parts Bins, the Parts Table, the Wing, and the Desktop MBI location.

Tracking data is time-based, allowing the participants' movements to be played back in a real-time animation, fast-forwarded or rewound. Converting the raw tracking data into these visual representations helped reveal trends in the data, such as how the movements around the work cell change based on the mode. Visualizing the data raised new questions about how participants moved around the space, such as how many trips a participant made between the Parts Bins and the Wing. The post-processing app was able to help answer those questions by looking at the tracking data and the areas of interest. The processing of the tracking data allowed discrete values to be placed on these types of questions for use in statistical analysis with other traditional forms of data collection.

RESULTS

The study was performed with a total of 48 participants distributed among the three different modes with 15 Desktop MBI, 15 Tablet MBI, and 16 Tablet AR. Two participants' data were removed. One was due to failure of the data collection software and the other was an outlier in several measured metrics indicating that the participant did not understand the instructions. Among the remaining 46 participants, the age range for participants varied from 18 to 44 years of age. 80% of the participants were between 18 and 20 years of age, 18% were between 23 and 30 years of age, and 2% were between 30 and 44 years of age. All participants were students with 78% of them majoring in engineering. There was an uneven gender split with 78% of the participants being male to 22% female. The results of the study showed no significant differences in errors or time based on any of the demographic data. When comparing the traditional model based instructions with augmented reality instructions, there were three areas of interest: 1) First time quality (lowest errors); 2) Fastest time; and 3) Worker efficiency. Each of these areas show how the approach of fusing system data with human subject data can further enrich training outcomes and measures.

First Time Quality

First time quality is the ability for a novice trainee with little or no experience to perform an operation the first time with no errors. This could be anything from assembling an unfamiliar firearm to correctly implementing a list of safety procedures the first day on the job. First time quality was evaluated based on the number of errors made during the assembly task. These were measured in a traditional method with the observer making detailed notes on errors using a paper evaluation form. The types of errors observer were divided into four categories: Incorrect Part; Incorrect Location; Incorrect Orientation; Extra Part.

The breakdown of errors by instruction delivery mode is shown in Figure 5. The median of each mode was used instead of average errors to eliminate the influence of outliers. The blue bar represents the first assembly attempt and the green bar represents the second attempt. Both tablet modes performed significantly better than the Desktop MBI mode, with Tablet MBI $p < .038$ and Tablet AR $p < .0001$. When looking at errors by participant, the Tablet AR mode had more participants with zero errors. This indicates AR has the potential to improve first time quality on assembly tasks.

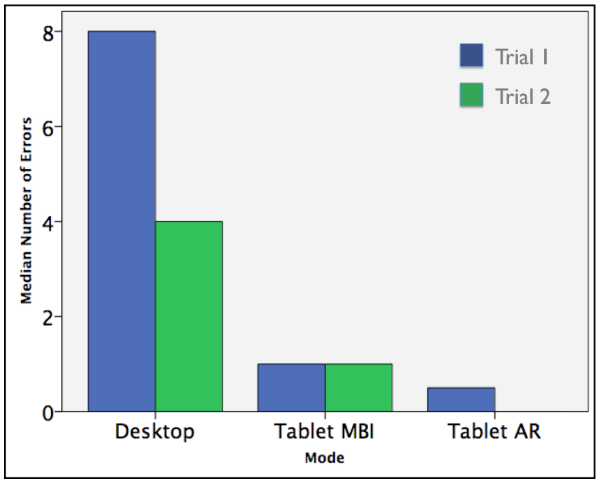

Figure 5: Errors by mode

When looking at the number of errors by task, Tablet AR had significantly lower errors on Steps 2, 3, and 11. Step 2 and 3 involved placing washers in a specific location. Step 11 involved the selection of one correct harness amongst similar incorrect harnesses. The significantly lower errors on these steps indicate AR may offer better information for exact placement and part selection tasks.

Fastest Time

The amount of time required completing a task as well as the learning curve are important measures of successful training. The total time was measured using a standard stopwatch and confirmed with the time stamped button clicks recorded. The time spent on each mode is shown in Figure 6 as minutes. The results indicate that most AR participants completed the assembly task faster the first time than with other modes. The variances are not equal in each mode (per Levene's Test), so the Welch ANOVA was used to show that Tablet AR trial 1 times (Blue bars on graph) are significantly lower than Desktop MBI (p = .01).

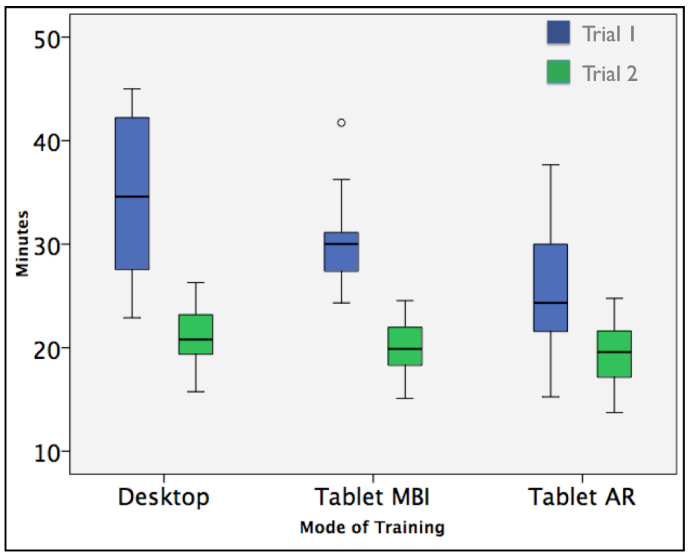

Figure 6: Time by mode

With the implementation of time stamped button clicks in each interface, it was possible to track the participant's advances and time through each step and the overall task as shown in Figure 7. Utilizing the computer to track tasks at sub-second accuracy provided higher fidelity data with fewer chances of errors. The human observer was marking events at a five second resolution. Even at this lower time resolution, the human observers often commented on the number of tasks they were attempting to perform simultaneously and the difficulty in performing all the tasks without error. The automated button click collection had the added benefit of providing researchers with confidence in the task time data and the resulting conclusions.

Step 5 (yellow box) shows a longer time taken than the other two modes, although not significantly. The extra time taken on this task was likely to be due to the lack of occlusion in the AR interface, as observers noted that participants had difficulty with vertical part ordering on this step. Occlusion provides a sense of depth by hiding parts of the virtual object that would be hidden by the real world object. Step 6 was another interesting time difference with Tablet AR's time being significantly lower than Desktop MBI (p < .003). Step 6 is represented by the red colored box in Figure 7.

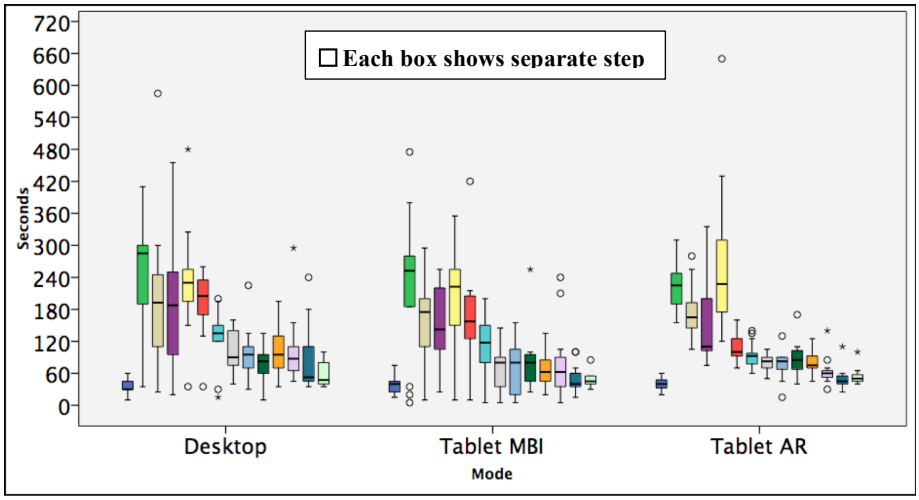

Figure 7: Time by Task

Worker Efficiency

The efficiency of a worker was looked at using the position and orientation of the both the tablet and the participant's head. This information allowed for the calculation of both the number of looks at the tablet during the task and the amount of time spent looking at the tablet. The human observer was not able to reliably log what the participant was looking at and for how long during the study. Instead the AR tracking system was leveraged to gather this information. The tracking system tracked the position and orientation of the helmet worn by participants

and the tablet holder. A criterion was developed for indicating a look at the tablet which required a ray emitted from the helmet's position and orientation to strike the tablet's plane within one meter of the tablet's center. One meter was allowed to handle small disparity with how the participant's wore the helmet.

The average number of looks at the tablet is shown in Figure 8. Both the first and second trials using the Tablet AR mode recorded significantly fewer average number of tablet looks (p = .004). The fewer number of looks meant that participants were not "bouncing" back and forth between the instructions and the physical task. They were able to focus on the actual assembly steps, as they understood the instructions more quickly.

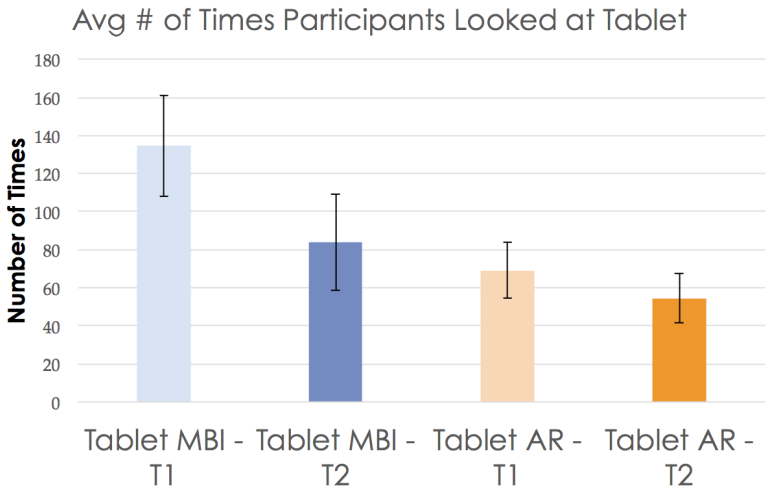

Figure 8: Average number of tablet looks

The average time per look was also calculated with the Tablet AR having a lower average time per look. The longer looks could indicate more focus on the task or more confusion interpreting the instructions. When combining the results of the average look time with the number of looks and the total task times, it would appear the longer look times would indicate more focus on the task.

Net Promoter Score

Among the self-reported data gathered was the Likert survey question, "I would recommend work instructions like this to a friend." Responses to this question on a 1-5 agree-disagree scale can be converted to a net promoter score (Reichheld, 2003) by subtracting the percentage of detractors (answers of 1, 2, or 3) from the percentage of promoters (answers of 5); answers of 4 are ignored. Net promoter scores range from -100% (worst) to 100% (best). According to Reichheld, the median net promoter score for over 400 companies in 28 industries was 16%. Net promoter scores for the three instruction modes were: Desktop, -20%; Tablet MBI, -47%, and Tablet AR, 44%. The difference between Tablet MBI and Tablet AR is dramatic, and is confirmed by statistical analysis of the data. A pairwise analysis using Dunn's (1964) procedure with a Bonferroni correction revealed statistically significant differences in recommend responses between Tablet AR and Tablet MBI (p = .015). One might suggest that the data is high for Tablet AR simply because of novelty. However, these responses were found to be significantly negatively correlated with the total number of errors in Trial 1, $r(45) = -.325$, $p < .031$ (one extreme outlier in errors was removed), which makes sense: participants who made fewer errors were more likely to recommend the system. After participants gave their recommendation rating, they were asked, "Why?" Their answers were critical for understanding the detractors. Detractors' answers fell into two primary categories: 1) complaints about the organization of the workflow (e.g., "Several steps could have been combined into one," or "I don't like the order [of steps]") and 2) complaints about the format and content of the instructions themselves (e.g., "Part names are way too confusing," or "It wasn't clear which direction to screw the bolts on"). This provided some of the most compelling evidence that sensor and self-reported data needed to both be present. From the net promoter scores, one could easily infer that Tablet MBI performed very poorly. In fact, for many of the tasks Tablet MBI scored significantly better than Desktop and often compared closely to Tablet AR, as has been shown previously. While it is important to consider the effect on each participant, it is equally important to consider how they performed using each mode.

DISCUSSION

In this study, three work instruction delivery methods were analyzed. Results covered in the previous section suggest that the use of augmented reality as a work instruction delivery method can increase first time quality while

reducing time on task. Data also indicates that using AR also led to more focus on each task by the participants. It was found that those using Desktop MBI and Tablet MBI spent a large amount of time traveling and confirming information by looking at the screen. Finally, it was found that specific tasks might benefit more strongly from the use of AR.

Traditional study methods including pre and post-surveys, video recording, a paper folding test, and qualitative, observational note taking were carried out for each participant of each instruction delivery mode under investigation. In addition to standard methods, non-traditional approaches including the use of head tracking and interface interaction were also recorded and used. Information such as the overall time, number of errors, and time for each task were possible to track using traditional methods of stopwatch and pen and paper. However, observers noted that it was difficult to keep up with recordings because of the large number of items they were asked to record, and that small errors may have been made due to this fact. Fortunately, by leveraging non-traditional quantitative measurements, namely the tracking system and interaction recordings, further analysis triangulated the data and confirmed situations causing potential discrepancy.

Using the tracking system also allowed the post-study analysis of trainee motion throughout the study. In some cases visualizing the participants' paths highlighted unique insights that otherwise would not have been captured. For example, it was very easy to pick out from the visualization that certain users would seldom move the mobile tablet. It is possible for the observer to note these facts, but with the observer pre-occupied, maintaining levels of accuracy and insight became increasingly difficult.

Another feature brought out by the post-processing tracking visualization was the commonality of participants making a back and forth motion between the Wing and tablet or desktop to confirm what they read or had viewed. Although not initially measured for by study design, this extra data provided a unique insight into why certain times were being observed and trends were coming out in the data.

A final and critical dynamic realized from data recorded in this manner was the ability to play back a participants' activities. This could be looked at by visualizing their paths or by viewing the full video. Because each participant was tracked throughout, data analysis can continue and further insights can be gleaned even though the study is over and participants are gone. New questions can still be asked and insights drawn from the data. One example of possible future work for this study would be an added feature to the visualization tool matching up timestamps of the participant interface interaction with the tracking data. This would allow analysis to be performed regarding how the participant was interacting with the device while in certain areas of the room. All of this can be done without requiring more trainees, more time, and subsequently more money and resources. This is not possible without the non-traditional study data collection.

Challenges and Recommendations

The described approach yielded many positives but difficulties arose while attempting to use more and more non-traditional data. For example, the extra work involved with recording this data. Two custom applications were built in order to use the tracking data (one for recording and saving the data and one for visualization). These can now be used for future work, but took valuable time to build as well as maintain.

In addition to the difficulty of acquisition, date storage represented a challenge. This study required the storage of five video feeds, tracking data, recorded device clicks, and observer notes, from separate computers. Aggregating all this data using manual sources (i.e. flash media, external hard drives) was a time consuming tasks. For a long-term study or large number of trials, it is worth considering networking the individual devices and automating the collection and storage processes. Finally, effectively using growing amounts of data is still a large problem both in military and industry today. The data are becoming easier to acquire, but using and understanding the data still remains difficult.

CONCLUDING REMARKS

In this work, traditional data collection methods for training analysis may have led to one or possibly multiple of the same conclusions. However, traditional methods would not have allowed for detection of tablet look time or accurate zone residence times, for example, which were used to gain greater insights into the traditional data

collected. By tracking participants and by recording interface interaction throughout each training session conducted, a major asset was added to the investigation.

As data continues to become easier to acquire through wearables, mobile devices, and other technologies, not leveraging their unique capabilities and ease of use represents a vast pool of unrealized data. In conclusion, it does not seem to be a question of whether to use these devices, but a question of how much, and to what extent.

REFERENCES

Caudell, T. P., & Mizell, D. W. (1992). Augmented reality: an application of heads-up display technology to manual manufacturing processes. *Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences*, 659–669 vol.2. doi:10.1109/HICSS.1992.183317

Creswell, J. W. (2003). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (2nd ed.). Thousand Oaks, CA: SAGE Publications.

Dunn, O.J. "Multiple comparisons using rank sums". Technometrics 6 (1964) pp. 241-252.

Durham, J., Tan, B.-K., & White, R. (2011). Utilizing Mixed Research Methods to Develop a Quantitative Assessment Tool: An Example From Explosive Remnants of a War Clearance Program. *Journal of Mixed Methods Research*, *5*(3), 212–226. doi:10.1177/1558689811402505

Fielding, N. G. (2012). Triangulation and Mixed Methods Designs: Data Integration With New Research Technologies. *Journal of Mixed Methods Research*, *6*(2), 124–136. doi:10.1177/1558689812437101

Hesse-Biber, S., & Johnson, R. B. (2013). Coming at Things Differently: Future Directions of Possible Engagement With Mixed Methods Research. *Journal of Mixed Methods Research*, *7*(2), 103–109. doi:10.1177/1558689813483987

Kaplan, B., & Duchon, D. (1988). Combining Qualitative and Quantitative Methods in Information Systems: A Case Study. *MIS Quarterly*, *12*(4), 571–586.

Kjeldskov, J., & Graham, C. (2003). A Review of Mobile HCI Research Methods. In Luca Chittaro (Ed.), *Human-Computer Interaction with Mobile Devices and Services* (pp. 317–335). Springer Berlin Heidelberg. doi:10.1007/978-3-540-45233-1_23

Lazar, J., Feng, J., & Hochheiser, H. (2010). *Research methods in human-computer interaction*. West Sussex, United Kingdom: John Wiley & Sons, Ltd.

Lukowicz, P., Timm-Giel, A., Lawo, M., & Herzog, O. (2007). WearIT@work: Toward Real-World Industrial Wearable Computing. *Pervasive Computing, …*, *6*(4), 8–13. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4343891

Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology*, *74*(4), 619-624.

Nakanishi, M., Ozeki, M., Akasaka, T., & Okada, Y. (2007). Human factor requirements for Applying Augmented reality to manuals in actual work situations. *2007 IEEE International Conference on Systems, Man and Cybernetics*, 2650–2655. doi:10.1109/ICSMC.2007.4413588

Nee, A. Y. C., Ong, S. K., Chryssolouris, G., & Mourtzis, D. (2012). Augmented reality applications in design and manufacturing. *CIRP Annals - Manufacturing Technology*, *61*(2), 657–679. doi:10.1016/j.cirp.2012.05.010

Newman, I., & Benz, C. (1998). *Qualitative-Quantitative Research Methodology: Exploring the Interactive Continuum*. USA: Southern Illinois University Press.

Orvis, K. L., Duchon, A., & DeCostanza, A. (2013, January). Developing Systems-based Performance Measures: A Rational Approach. In *The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)* (Vol. 2013, No. 1). National Training Systems Association.

Patton, Michael Quinn. Qualitative evaluation and research methods (2nd ed.). Thousand Oaks, CA, US: Sage Publications, Inc. (1990). 532 pp.

Reichheld, F. F. (2003). The one number you need to grow. *Harvard business review*, *81*(12), 46-55.

Shneiderman, B., & Plaisant, C. (2006). Strategies for evaluating information visualization tools. *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors Novel Evaluation Methods for Information Visualization - BELIV '06*, 1. doi:10.1145/1168149.1168158