Do we still need physical monitors? An evaluation of the usability of AR virtual monitors for productivity work

Leonardo Pavanatto*
Center for Human-Computer Interaction
Department of Computer Science
Virginia Tech, USA

Chris North†
Sanghani Center for AI and Analytics
Department of Computer Science
Virginia Tech, USA

Doug A. Bowman‡
Center for Human-Computer Interaction
Department of Computer Science
Virginia Tech, USA

Carmen Badea§
Microsoft Research
Redmond, WA, USA

Richard Stoakley¶
Microsoft Research
Redmond, WA, USA

Figure 1: AR virtual monitors could be used to provide extra screen space when users are working from mobile office locations.

Abstract

Physical monitors require space, lack flexibility, and can become expensive and less portable in large setups. Virtual monitors, on the other hand, can minimize those problems, but may be subject to technological limitations such as lower resolution and field of view. We investigate the impacts of using virtual monitors displayed on a current state-of-the-art augmented reality headset for conducting productivity work. We conducted a user study that compared physical monitors, virtual monitors, and a hybrid combination of both in terms of performance, accuracy, comfort, focus, preference, and confidence. Results show that virtual monitors are a feasible approach for performing serious productivity work, albeit currently constrained by technical limitations that lead to inferior usability and performance compared to physical monitors. We also discovered that, with current technology, the hybrid condition was a better trade-off between the familiarity and trustworthiness of physical monitors and the extra space provided by virtual monitors. We conclude by expressing the opportunity for designing strategies for mixing virtual and physical monitors into novel hybrid interfaces.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Mixed / augmented reality; Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Empirical studies in interaction design

*e-mail: lpavanat@vt.edu
†e-mail: north@vt.edu
‡e-mail: dbowman@vt.edu
§e-mail: carmen.badea@microsoft.com
¶e-mail: richard.stoakley@microsoft.com

1 Introduction

As computing processing capabilities grow, users have a more prominent need for more screen real estate to manage and visualize multiple windows at the same time [6]. Consider productivity tasks (such as creating or modifying documents, images, or videos) where users must compare content between two or more windows; when they need to keep track of the state of a task while working on another document; or when there is a need to transfer content between windows. These tasks may become more complicated when the user lacks sufficient space to place these windows side-by-side. Extra space allows for drag and drop operations and quick glances, taking advantage of external memory and physical navigation [2].

Setups such as large-scale displays [2], ultra-wide curved displays [10], and multi-monitor setups [4,6] provide high-resolution displays with wide aspect ratios that can fit large quantities of content side-by-side while taking advantage of the wide human field of view. The adoption of these devices is not without a cost, however. They occupy large amounts of physical space that may not be available or could be used for other ends. The bulk and weight of such setups also makes them less portable. Thus, a student in a classroom or a passenger on a train is unable to use large display setups and is limited to a small laptop display. The physical nature of the displays can also make them less flexible to changes in shape or size, and the cost of a large display setup, which is restricted to use in a single workspace, can become very large. The importance of portability becomes evident when we consider situations such as the ongoing COVID-19 pandemic, where many people are alternating between working from their office and home.

Augmented reality (AR) technologies have been explored as an alternative approach for substituting physical monitors [14, 17]. Instead of placing a physical device in the environment, users can see virtual monitors through a portable head-worn display (HWD). By using head tracking, these devices provide users with spatially registered content and a fully surrounding field of regard. These monitors do not occupy physical space, and they can be created, destroyed,

Table 1: Potential benefits and limitations of AR virtual monitors.

Potential Benefits	Potential Limitations
Do not occupy physical space	Lower readability
Flexible (size, shape)	Accommodation-vergence
Portable	Not fully opaque
Single cost	Smaller field of view
Privacy	Misleading occlusion cues
360-degree field of regard	Brightness differences
Any number of monitors	Color distortions

and modified (in shape and size) to match the user's needs. An example of usage can be seen in Figure 1, where a user is working with virtual monitors on a train. Finally, head-worn displays are likely to become much less expensive in the upcoming years, with the widespread use of pervasive AR [16, 19]. However, current state-of-the-art AR HWDs have technical limitations for characteristics such as resolution, field of view, and brightness that may make virtual monitors less usable. Some of the potential benefits and limitations of using AR virtual monitors are listed in Table 1.

The question, then, is how feasible is it to do real work with AR virtual monitors? What are the limitations posed by current state-of-art technology that impact productivity work? And how can we minimize this impact? In this research, we investigate the effects of replacing physical monitors with purely virtual representations, and a hybrid combination of both. We implemented a prototype that allows the use of a full-featured Windows 10 operating system with both a Microsoft HoloLens 2 and conventional monitors. We conducted a user study aimed at understanding the benefits and limitations of their use on productivity work, while considering aspects such as performance, accuracy, head movement, focus, comfort, confidence, readability, and user preference.

The contributions of this work include: (1) a quantifiable understanding of the usability of virtual monitors (based on a current state-of-the-art AR HWD) compared to physical monitors when conducting productivity tasks, (2) an understanding of the opportunities of using hybrid physical/virtual monitors for conducting productivity tasks, (3) outlining design implications for how to take advantage of the flexibility of virtual monitors for designing desktop UIs, and (4) a formal evaluation method that could be used or modified to consistently evaluate similar systems now and in the future.

2 Related Work

2.1 Lessons from Studies of Physical Displays

Previous research on large monitors has shown that when conducting cognitively difficult tasks, a display with larger screen space can provide a significant advantage in performance [10]. Enhanced performance was attributed to physical navigation and maintaining an overview context [3]. The location and visual appearance of content in large displays also become valuable clues to keep users aware of the organization of the space, as a type of external memory [1, 4]. This indicates the importance of having more screen real estate, and being able to access it using body motion instead of window or desktop switching.

However, known issues with large display setups include managing the mouse position and performing window/task management [25]. Differences in perception and usage can also impact the usage of such systems, where the designer needs to carefully consider placement strategies for mouse, keyboard and displays [12]. The existence of bezels between multiple monitors and the visual separation of displays at distinct depths create visual discontinuities that can impact task performance [27, 28]. Also, display size can affect task performance in complex tasks [9]. We expect similar effects to be present on virtual monitors.

2.2 Productivity Work in Mixed Reality

Existing research has explored the use of VR and AR technologies for conducting everyday tasks [5]. Early work placed 2D windows in the 3D real environment through the use of a see-through head-mounted display [13], allowing users to create links between windows and physical objects. Alternatively, the usage of cameras and projectors has been envisioned to augment existing office spaces, through the creation of spatially immersive display surfaces over physical objects [23]. Combining virtual displays with laptop/tablet touchscreens was shown to be a feasible approach to aid mobile workers [7]. Working with a combination of physical and digital documents through AR has also been shown to be feasible and accepted by users [18]. Existing work has also combined physical displays and virtual representations, such as displaying visualizations over tabletops [8], or large displays [20, 24].

Conducting office work in VR HWDs has been proposed to address challenges such as lack of space, the existence of surrounding noise, illumination issues, and privacy concerns [17,22]. VR reduced distraction of users working on open office environments, induced flow, and was preferred by users [26]. However, traditional anchored input devices such as keyboard, mouse, and touchpad can be hard to use with peripheral portions of the display space that are distant from the input devices, forcing additional head rotation that could possibly result in neck pain [17]. In situations of many displays organized horizontally in a cylinder around the user, it was shown that performing a virtual movement opposite to the physical user movement could reduce the amount of head movement required to access peripheral displays [21]. Context switching between physical and virtual, and focal distance switching between displays have also been shown to reduce task performance and induce visual fatigue [15]. Having multiple depth layers, such as when combining a HWD and a smartwatch, can also induce more errors when interacting [11].

To our knowledge, the usability of AR virtual displays for productivity work has not been studied. This is likely because AR displays were not advanced enough to make such systems feasible in the past. However, the current state-of-the-art AR displays such as HoloLens 2 make it possible to study this topic seriously for the first time, and thus explore many open questions. For instance, how do users perceive the differences between physical and virtual displays, and should they be used together in a complementary way? How do current technological limitations and human factors constrain the scope of what can be achieved with this approach? How could we overcome these limitations through interaction design?

3 User Study

We conducted a user study to investigate the usability of currently achievable virtual monitors when compared against physical monitors, to study the benefits of using a hybrid physical/virtual multi-monitor setup compared to a purely virtual one, and to obtain knowledge to help us understand how we can design a desktop UI to take advantage of the flexibility of virtual monitors. The rationale behind this study is the need to empirically quantify how different and feasible it is to complete an ecologically valid task with virtual monitors, which is essential to inform future design decisions. Once deemed viable, then this knowledge would enable us to explore the design space of virtual monitors, which could include novel UI paradigms and hybrid combinations of heterogeneous displays.

3.1 Conditions

Our study included three multi-monitor setups. A purely **Physical** setup, seen in Figure 2(a), was our baseline, with three physical monitors side-by-side. We considered this setup to be the gold standard. We opted for a multi-monitor setup because we were

Figure 2: (a) Physical condition had three monitors side-by-side; (b) Virtual condition had three monitors rendered through HoloLens; (c) Hybrid condition combined a central physical monitor with two peripheral virtual monitors.

also interested in understanding how limitations of virtual monitors would differ depending on their usage and spatial location, and because multi-monitor displays are one of the primary use cases for virtual monitors.

A purely **Virtual** setup, shown in Figure 2(b), was our second condition, with three floating virtual monitors rendered through the HWD. Given the fixed accommodation and lower resolution of the HWD, we chose to display these monitors at 2m from the user and at a larger size (1.5x larger in angular size compared to the central physical monitor). As we aimed to understand the limitations and benefits of this approach, these choices allowed us to achieve decent readability and reduce eye strain, thus maintaining the same capabilities for both conditions. Given the lower resolution of the HWD, we had to choose between lowering the resolution of the physical monitors or scaling up the virtual monitors to achieve the same readability. We opted for the more ecologically valid option, where we consider that the user would use the maximum resolution allowed by the physical display, and we adapted the virtual ones to match that requirement. We argue that the lower resolution of virtual monitors is an important aspect of existing technology, and must be considered in the analysis.

Finally, a **Hybrid** setup, shown in Figure 2(c), used a central physical monitor and two peripheral virtual monitors. We considered this setup to be a middle ground between the high specifications and user familiarity of physical monitors, with the potential benefits of virtual monitors.

3.2 Implementation and Apparatus

A Microsoft HoloLens 2 was used as the HWD in all conditions. The untethered device allowed free movement of the user while using optical see-through to combine virtual content with the real world. Since it does not rely on cameras or displays for visualizing the real world, this approach minimizes visual degradation of the physical scene compared to video see-through displays. It has a field of view of 43 degrees horizontally and 29 degrees vertically. It has a resolution of 2048x1080, with a 3:2 aspect ratio and about 2.5k light points per radian. Head tracking is done with four visible light cameras, while eye tracking uses two IR cameras. Participants wore the HoloLens 2 during all experimental conditions, including Physical. The device was used both for rendering virtual monitors when necessary and for obtaining metrics from the usage of the monitors, such as head rotation and eye-tracking. The physical monitors used in the study had 24 inches (central) and 21.5 inches (sides). Each individual physical and virtual monitor was rendered with a resolution of 1080p.

As our objective was to understand how the conditions impact users conducting productivity work, we used a full version of the Windows 10 operating system as the interaction environment. We designed our implementation using the Unity Engine, version 2019.2.21f1, the Mixed Reality Toolkit (MRTK), version 2.3.0, and the Windows Duplication API, ported to Unity through the uDesktopDuplication plugin. The unity scene had planes floating in an ellipse around the user. Each of these planes rendered an external texture of the monitor capture obtained from the duplication API. MRTK managed HoloLens integration, including spatial, hand, and eye tracking. We opted to use the Holographic Remoting Player for mirroring Windows 10 monitors on HoloLens. This application streams content from a computer to a HoloLens in real-time, through a Wi-Fi connection. The HoloLens sends the information obtained from its sensors (such as head and eye tracking) back to the PC, which uses them to make all necessary computations. The result is sent back to HoloLens and displayed to the user. For input, we used a standard wired keyboard and mouse in all conditions; the mouse could be moved naturally across all three monitors, without gaps between them.

The experiment was run on a PC with an AMD Ryzen 5 3600 6-Core 3.6GHz CPU, 16GB of 3200MHz DDR4 DRAM, a Samsung M.2 SSD, a EVGA GeForce RTX 2060 Super 8GB GPU, and a Asus TUF X570-Plus Motherboard, with integrated Intel Wireless-AC 9260 (2x2 antenna, 802.11ac, speeds up to 1.73Gbps).

3.3 Experimental Design

Our within-subjects independent variable was monitor type (Physical, Virtual and Hybrid), and we counterbalanced the order of presentation of the three conditions using a Latin Square. We recruited 18 participants from the general population that fit the following inclusion criteria: (1) were at least 18 years old, (2) had normal vision (corrected or uncorrected), (3) were proficient with the English language, and (4) used a computer daily for work.

Our dependent variables included performance (time to complete tasks), accuracy (correctness of task completion), confidence (how much the users would trust the system to do work), comfort (how much head movement they had to use), readability, and user preference.

3.4 Hypotheses

Given the current technical limitations of AR HWDs, our hypotheses were as follows:

H1. Using virtual monitors for conducting productivity work will lead to a measurable decrease in performance and accuracy when compared against physical monitors. We believed that a

Table 2: Blocks of questions in the main experimental task, with question type and requirements

Question Block	Type	Requirements
Rubrics	Physical transfer of content (text)	Copy and paste parts of the rubric document.
TA Grades	Physical transfer of content (numbers)	Copy and paste parts of each spreadsheet.
TA Consistency	Glance at content (Boolean)	Analyze if the grading was consistent or not.
Rubric Consistency	Cognitive transfer of content (text and numbers)	Analyze which rubrics were consistent or not.
Average	Heavy work on side monitor (text input)	Modify spreadsheets to find the overall average for all submissions.
Letter Grades	Heavy work on side monitor (manipulation)	Sort grades and count the number of students within a grade range.
Feedback	Glance at content and write on central monitor	Write freeform feedback to students.

purely virtual set of monitors would not perform as well as physical monitors, given limitations described in 1.

H2. Using the Hybrid condition will lead to a measurable increase in performance and accuracy when compared against a set of virtual monitors. While we acknowledge that there is a cost of context switching, we believed that the advantages of Hybrid would outweigh this cost, since having a central physical monitor would provide significant benefits for the central user task

H3. Users will have a measurably higher acceptance of the Hybrid condition than of the Virtual condition. We believed that the benefits of having higher resolution in the central monitor, with the glanceable space in the peripheral monitors, would lead to users preferring to work with the Hybrid condition over Virtual.

3.5 Experimental Task

We used an ecologically valid productivity task to test our hypotheses. The task's narrative was that the participant was the head teaching assistant (TA) of an undergraduate class, and it was their responsibility to provide students with performance feedback on their latest assignment. They were asked to fill out a feedback form that would be submitted to the instructor, and later to students.

Participants were informed that, while they were responsible for aggregating the feedback, two other TAs were responsible for actually grading the assignment. Each of them would have graded half of the submissions, and sent the grades to the participant. Participants were instructed that the grading should be consistent between the TAs, meaning that they needed to give similar grades for similar quality of submissions. All the information that they needed was available in five documents. The documents included a Word file with the assignment description, two Excel spreadsheets with the grades from each TA, a Word file with examples of feedback that had been given to students in the past, and a Word file with the rubric that the TAs used to grade the activity.

Participants answered seven blocks of questions, designed to require reading, information transfer, and interaction across all the documents and all three monitors. The question types and requirements are described in Table 2. The last question required using all the documents together to synthesize overall feedback to the class. A four-minute time limit was given for this question, to create a compromise between giving a complete answer and not spending too much time trying to find hidden data.

3.6 Procedure

The study was approved by the university's Institutional Review Board, including specific procedures regarding SARS-COV-2 exposure mitigation. During the time of the study, the spread of the virus was relatively under control in our community, with less than 30 registered new cases a day per 100k people, and a seven-week average of positivity rate averaging 3%.

The study took place face-to-face at our laboratory, in a single session of 90 minutes. We recruited participants through mailing lists, and asked them to complete a screening questionnaire for our inclusion criteria. They scheduled a session time, and received a digital copy of the consent form. They completed a screening questionnaire to evaluate their risk of SARS-COV-2 exposure, and the session would only be confirmed if the risk was low.

Before and after each session, tables, keyboards, mouses, pens, and the AR display were disinfected thoroughly. Sessions would be spaced by at at least 30 minutes, and only one person would use a given HoloLens device in the same day. Upon arrival, we greeted the participant outside the building and took their temperature. If it was lower than 100.4F, we proceeded with the study. During the entire study, participant and investigator wore face masks, stayed 6ft apart, and used hand sanitizer (participant) and gloves (investigator).

The participant then signed the consent form, provided contact information for our contact tracing log, and answered a brief background questionnaire on the computer. Next, they received general instructions. The participant then completed the standard HoloLens eye-calibration procedure. Before beginning the main task, the investigator presented the questions that would be asked during the task, giving participants time to ask any question they might have had about them. After that, we started the first condition.

For each condition, a calibration was made of the physical monitor position to the HoloLens coordinate system, for matching eye-tracking data and rendering virtual monitors in the same position across participants. After the calibration, the participant had some free time to explore the condition. In the first condition, this free time also included some instruction on how to conduct common Windows and Excel operations, such as copy and paste, counting cells, calculating the average, and sorting cells. In the main task, presented in Section 3.5, the participant answered questions displayed on the screen based on the documents they had on their monitors. This documents were at fixed locations and could not be moved by participants. The participants could, however, perform manipulations on these documents to find the results for the questions. After the main task, they answered a questionnaire for that specific condition. For each condition, we had distinct datasets to be sure that participants had to find the correct answers anew for each condition.

Once all conditions were completed, the participant answered a final questionnaire, where they ranked the conditions in terms of preference, readability, comfort, and confidence. The study was completed with a semi-structured interview. The interview was conducted with 10ft of distance between the participant and the investigator, and took about 10 minutes.

3.7 Participants

Eighteen participants (aged 18 to 32, three female) from the campus population took part in the experiment in individual sessions of around 90 minutes. One was a professional, three were graduate students, and 14 were undergraduate students. All participants used a computer daily for work, with 16 having at least intermediate experience with the Windows operating system, and 11 having low to intermediate experience using Excel. Of the participants, 12 had little to no experience with AR. Another 14 participants had never

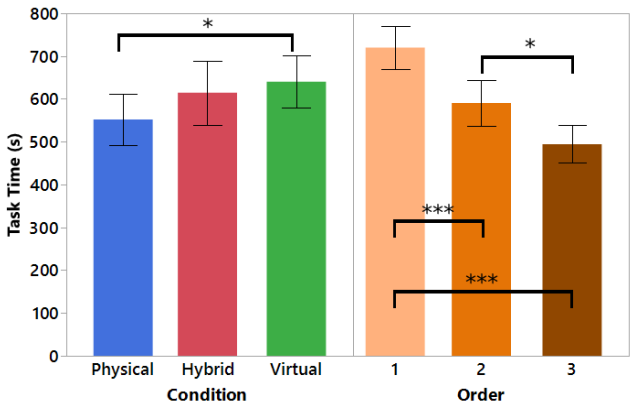

Figure 3: Total time in seconds to complete the experimental task on each condition (left) and order (right). Error bar represents the confidence interval (95%).

been an instructor or teaching assistant in any class.

4 Results

For each participant, we obtained data from three sources: an online survey, which included answers to the background, conditions, and final questionnaires, along with the answers and time to complete each question block in our main task; a frame-by-frame log file created by Unity, which included time, frame time, head orientation angle, gaze hit 3D coordinates, monitor and window being gazed at, cursor 2D coordinates, and foreground window (last clicked by cursor); and audio files, which included the in-depth responses given by participants during the semi-structured interviews, and were manually transcribed by the authors.

All files were saved in ".csv" format, and processed for analysis through *Python* scripts. We then proceeded to analyze the results statistically using the *JMP Pro 14* software. We used an α level of 0.05 in all significance tests. In the results figures, pairs that are significantly different are marked with * when $p \leq .05$, ** when $p \leq .01$, and *** when $p \leq .001$. For all the cases, we verified normality through Anderson-Darling tests and normal quantile plots, before deciding whether to apply ANOVA or non-parametric tests.

4.1 Performance

We conducted a two-way mixed-design factorial analysis of variance (ANOVA) for the total time taken in the main task. Task times and confidence interval related to order and condition can be seen in Figure 3. Our two factors were the condition being used (Physical, Hybrid, or Virtual) within-subjects, and the order that they were experienced (1st, 2nd, or 3rd) between-subjects. Although we tried to ensure that participants learned everything that they need to know about the task before starting the first condition, we still noticed in pilot testing that the first condition was slower than the other two. By including order as a factor, we can measure and isolate learning effects, understanding both how each of the conditions perform, and also how they perform after task repetition.

There was a significant main effect for condition ($F_{2,2} = 4.197, p = .021, \eta^2 = .078$). Pairwise comparison using Tukey HSD found a significant difference ($p = .019$) between Physical ($M = 551.94s, SD = 119.49$) and Virtual ($M = 640.40s, SD = 122.35$). There was also a significant main effect for order ($F_{2,2} = 25.760, p < .001, \eta^2 = .480$), with first ($M = 720.06s, SD = 100.54$), second ($M = 591.41s, SD = 106.23$), and third ($M = 495.45, SD = 88.18$). Pairwise comparison found significant difference between all placements, first and second ($p < .001$), first and third ($p < .001$), and second and third ($p = .010$). There was no significant interaction effect between factors ($p = .65$).

Table 3: Main effects for condition on each block of questions. Silver highlighted rows show at least marginal significance.

Block	$F_{2,2}$	p	η^2	Condition	Mean	SD
Rubrics	3.05	.057	.07	Physical Hybrid Virtual	49.28 58.94 68.06	21.93 22.77 40.72
TA Grades	6.24	.004	.19	Physical Hybrid Virtual	46.04 59.99 70.94	16.05 24.96 23.70
TA Consistency	0.66	.518	.02	Physical Hybrid Virtual	30.84 38.42 35.78	14.26 29.75 14.08
Rubric Consistency	0.38	.685	.01	Physical Hybrid Virtual	64.68 73.61 70.76	27.53 49.18 31.80
Average	3.40	.042	.09	Physical Hybrid Virtual	73.01 96.83 85.96	22.76 36.25 33.55
Letter Grades	0.24	.781	.007	Physical Hybrid Virtual	67.57 65.61 73.24	46.90 33.74 37.98
Feedback	2.52	.091	.08	Physical Hybrid Virtual	220.51 221.19 235.65	22.74 27.05 20.50

We also analyzed the blocks of questions in the task individually; the analysis results are shown in Table 3, and questions times organized by condition can be visualized in Figure 4. As seen in the table, we observed Physical being significantly faster than Virtual for the *Rubrics* and *TA Grades* blocks, and Physical being significantly faster than Hybrid for the *Average* block. Most of the question blocks showed a significant order effect, but we found no interaction effects between order and condition.

We can summarize results as Physical was 13.81% faster than Virtual. In terms of order, second condition was 21.7% faster than the first, while third condition was 19.36% faster than the second. In the "Rubrics" question block, Physical was 38.10% faster than Virtual. In "TA Grades", Physical was 54.08% faster than Virtual. And in "Average", Physical was 32.62% faster than Hybrid.

4.2 Accuracy

We analyzed the accuracy of the answers by giving them a score based on correctness. Answers were graded as right (1) or wrong (0). For subjective question ("Feedback"), answers were graded based on the quality of the response, namely if the question was answered correctly, and information was not missing (such as time ending before the participant completed a sentence). The total score, considering the sum of the scores of each question, was analyzed for significance. We conducted a Wilcoxon / Kruskal-Wallis test, and didn't find any main effect for either condition ($p = .739$) or order ($p = .462$). Accuracy levels were around 83% across all conditions and orders. We still conducted a ChiSquare test for proportions for each question individually, but did not find any significant differences (smallest $p > 0.3$).

4.3 Visual Attention

We also analyzed how frequently participants moved their gaze between windows. We used the eye-tracking data to cast a ray and estimate which window the participant was looking at during every frame of the task. We then counted the number of times that the user looked at a new window, but we did not count the times when they

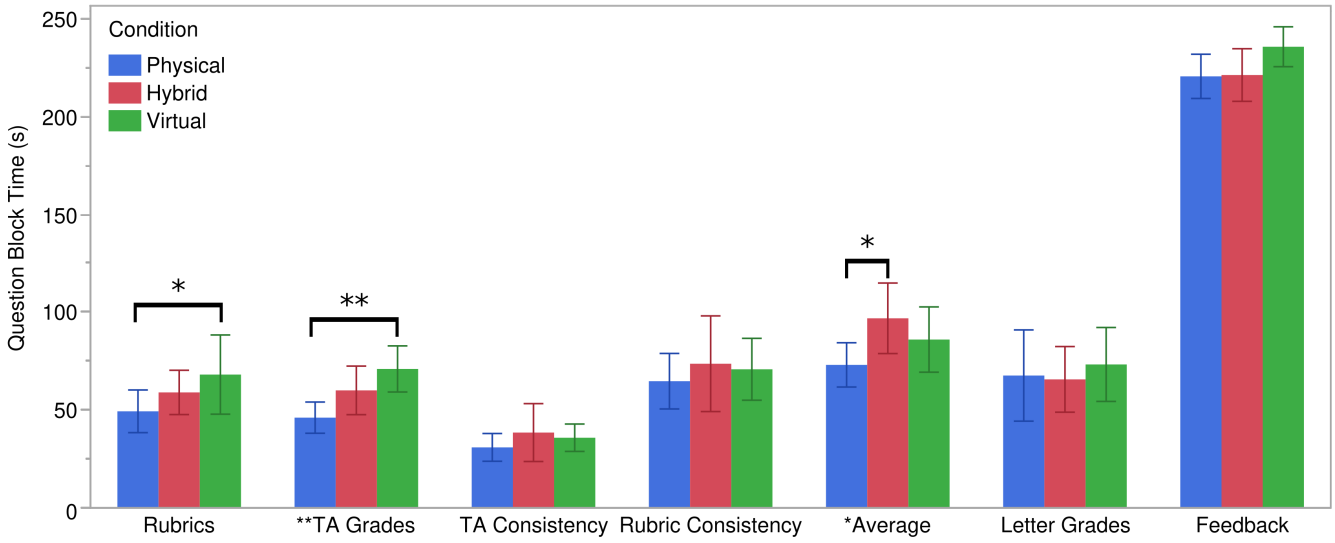

Figure 4: Time in seconds to complete each question block. Error bar represents the confidence interval (95%).

were looking at a given window and moved their gaze to the surrounding area. We then conducted a two-way ANOVA on the results. There was a significant main effect of condition ($F_{2,2} = 4.65, p = .015, \eta^2 = .13$), with Physical ($M = 296.56, SD = 76.95$), Hybrid ($M = 267.22, SD = 68.24$), and Virtual ($M = 235.17, SD = 56.09$). A pairwise analysis indicated a significant difference between Physical and Virtual ($p = .010$). There was also a significant main effect for order ($F_{2,2} = 7.41, p = .002, \eta^2 = .20$), with first ($M = 303.72, SD = 63.45$), second ($M = 268.89, SD = 65.95$), and third ($M = 226.33, SD = 64.52$).

4.4 Head Orientation

We analyzed the range of head movement used during the experimental task. We applied a median window filter of size 60 across the frame data from each user, and selected the minimum and maximum values for each direction. We chose 60 since the application ran at 60 frames per second, so it was big enough to eliminate small measurement errors and outliers, but small enough to not eliminate actual head rotations. We conducted a two-way mixed-design factorial analysis of variance (ANOVA) on the entries. The results for all angles can be seen in Table 4. As shown in the table, and as expected, the range of head turning was smallest for the Physical condition and largest, especially in Yaw, for the Virtual condition.

We also analyzed the total amount of head movement by applying a median window filter of size 60 across the frame data from each user and summing the absolute differences between each frame. We again conducted a two-way ANOVA on the results. There was a significant main effect of condition on Yaw ($F_{2,2} = 8.62, p < .001, \eta^2 = .23$), with Physical ($M = 4106.55, SD = 1060.96$), Hybrid ($M = 4974.85, SD = 1428.36$), and Virtual ($M = 6167.87, SD = 2061.38$). Pairwise comparison shows a significant effect between Physical and Virtual ($p < .001$), and a marginally significant effect between Hybrid and Virtual ($p = .053$).

4.5 Subjective Measures

After experiencing each condition, users were presented with statements that they had to rate using a seven-point Likert scale. The results are illustrated in Figure 5. We conducted Wilcoxon / Kruskal-Wallis Tests to evaluate effects of condition on the ratings.

"I found it easy to perform the task using this condition." There was a significant effect ($H = 12.67, p = .0018, \eta^2 = .26$), with Physical ($M = 6.11, SD = 0.76$), Hybrid ($M = 5.39, SD = 1.33$), and Virtual ($M = 4.28, SD = 1.67$). Multiple comparisons show significant differences between Virtual and Physical ($Z = -3.42, p < .001$), and Virtual and Hybrid ($Z = -2.02, p = .0433$).

Table 4: Main effects for condition on each axis of head rotation. Silver highlighted rows show at least marginal significance.

Angle	$F_{2,2}$	p	η^2	Condition	Mean	SD
- Pitch (X)	7.16	.002	.22	Physical	-12.46	5.18
				Hybrid	-17.36	4.42
				Virtual	-17.87	4.58
+Pitch (X)	0.85	.434	.03	Physical	18.30	7.90
				Hybrid	14.78	9.27
				Virtual	15.53	8.81
- Yaw (Y)	27.35	<.001	.55	Physical	-46.55	11.15
				Hybrid	-57.73	5.86
				Virtual	-70.71	11.24
+Yaw (Y)	41.59	<.001	.63	Physical	48.07	6.95
				Hybrid	64.46	8.67
				Virtual	71.73	7.52
- Roll (Z)	5.32	.008	.18	Physical	-14.08	4.56
				Hybrid	-16.90	3.44
				Virtual	-19.19	5.56
+ Roll(Z)	10.49	<.001	.30	Physical	16.02	4.24
				Hybrid	21.26	4.43
				Virtual	22.66	4.74

"I thought there were visual discrepancies between the monitors." There was a significant effect ($H = 6.41, p = .040, \eta^2 = .12$), with Hybrid ($M = 4.61, SD = 2.09$), Physical ($M = 3.5, SD = 2.04$), and Virtual ($M = 2.94, SD = 1.51$). Multiple comparisons show a significant difference between Virtual and Hybrid ($Z = -2.46, p = .0138$), and a marginally significant difference between Physical and Hybrid ($Z = -1.67, p = .095$).

"I found that I could see anything on the monitors at a glance." There was a significant effect ($H = 7.45, p = .024, \eta^2 = .13$), with Physical ($M = 4.94, SD = 1.89$), Hybrid ($M = 4.33, SD = 1.64$), and Virtual ($M = 3.33, SD = 1.78$). Multiple comparisons show a significant difference between Virtual and Physical ($Z = -2.48, p = .013$), and a marginally significant difference between Virtual and

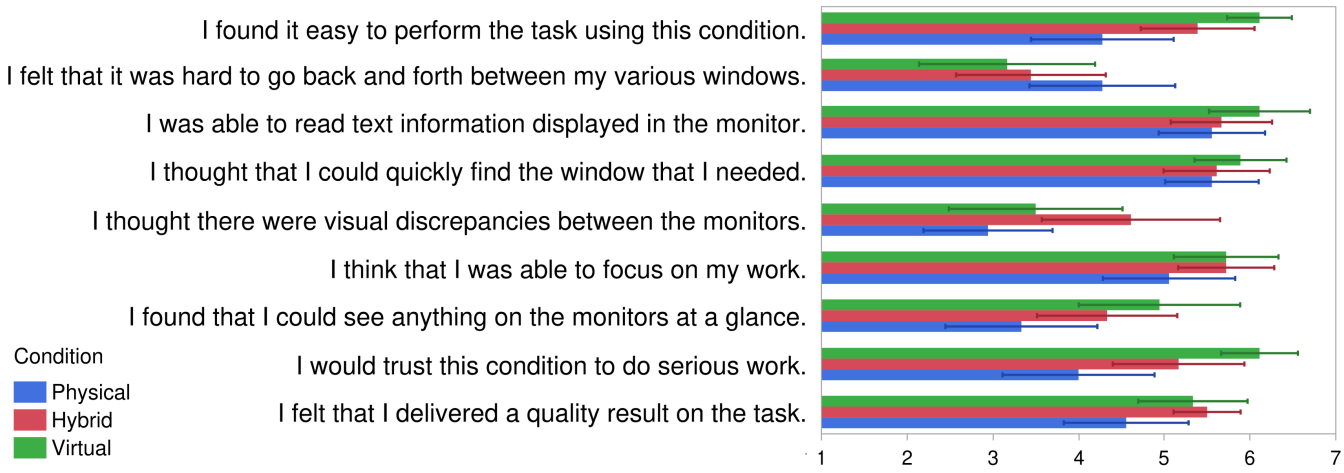

Figure 5: Participants ranked statements on a scale from 1 (Completely Disagree) to 7 (Completely Agree).

Hybrid ($Z-1.78, p=.074$).

"I would trust this condition to do serious work." There was a significant effect ($H=14.49, p=.001, \eta^2=.27$), with Physical ($M=6.11, SD=0.90$), Hybrid ($M=5.17, SD=1.54$), and Virtual ($M=4, SD=1.78$). Multiple comparisons show significant differences between Physical and Hybrid ($Z=2.04, p=.041$), Virtual and Hybrid ($Z=-1.98, p=.048$), and Virtual and Physical ($Z=-3.66, p<.001$).

There were no significant effects of condition on the other ratings.

4.6 Condition Ranking

We also asked participants to rank their preference, readability, comfort and confidence after they completed all conditions. Again we conducted Wilcoxon / Kruskal-Wallis Tests. Overall, results pointed for best rankings of the Physical condition, followed by Hybrid and then Virtual. Results can be seen in Table 5, and Figure 6.

5 Discussion

We hypothesized that Virtual would lead to a measurable decrease in performance and accuracy when compared against Physical (**H1**). Our results partially supported H1. There was a significant difference in overall time between the Virtual and Physical conditions, with an increase of about 14% in Virtual. However, we didn't find any main effect of condition on accuracy. This implies that even with the current state of the art in AR, virtual monitors are a feasible approach for performing productivity tasks, although with a measurable loss in performance. This loss is important, but could potentially be smaller than the performance decrease that would occur without the extra space provided by the virtual monitors. This should be verified in a future study. The results in the individual question blocks ("Rubrics", "TA Grades"), suggests that the performance loss could be tied to users taking longer to switch between monitors, since these questions required a quick copy and paste, and the information was easy to locate.

Our other data shed some light on why this performance gap occurs. In the Virtual condition, users rotated their heads through a larger range to see the peripheral monitors (which would be expected since virtual monitors are larger), and they made more accumulated rotations than the physical condition. This indicates that, since distances are larger, any back and forth between monitors in this condition means considerably more head movement, and that takes time. In addition, the limited FOV of the HWD meant that glancing at information in the periphery would require both head and eye movement in the virtual condition. We can see from our eye movement analysis that participants in the virtual condition moved their gaze between windows *less*. More head movement and less gaze switching means that Virtual required more head movement to navigate within each window. This could also indicate that participants chose to avoid switching attention to other windows in order to avoid even more head movement.

We also observed that some users lost the cursor a couple of times in the Virtual condition. Results show users had less confidence in Virtual. Virtual was rated lower than Physical on the statements "I found that I could see anything on the monitors at a glance" and "I would trust this condition to do serious work."

On the other hand, we didn't find differences in accuracy, which is supported by the lack of a significant effect of condition for the statements "I think that I was able to focus on my work" and "I felt that I delivered a quality result on the task," indicating that users also did not perceive any potential effect on accuracy. Overall, with current state of the art HWDs, the evidence shows that Virtual is less usable than Physical, but is still feasible to use for productivity tasks.

Our second hypothesis was that Hybrid would lead to a measurable increase in performance and accuracy when compared against Virtual (**H2**). Our results do not provide strong evidence to support H2, although they hint at some benefits of Hybrid. While we did not find any measurable difference in accuracy, there are some interesting findings in performance. Hybrid was significantly slower than Physical for the Average question block, which required users to do more interactions with the virtual monitor. However, there was a trend for Hybrid to perform as well as Physical on the Feedback question, which required more use of the central monitor and quick glances at the virtual side monitors. Our data suggests that Hybrid approximates Virtual in situations where there is more interaction, such as keyboard and mouse use, on the virtual monitors, while it gets closer to Physical when the usage of the virtual monitors is restricted to glanceable interaction. Future work is needed to statistically verify these trends.

Further, head orientation and eye focus measures showed no statistical significance between Hybrid and either Physical and Virtual. On the subjective metrics, Hybrid was rated statistically higher than Virtual for the statement "I found it easy to perform the task using this condition," with no statistical difference against Physical. It was rated worse than Virtual for the statement "I thought there were visual discrepancies between the monitors," which is expected considering the differences between the displays in terms of size, brightness, depth, and opacity. Finally, for "I found that I could see

Table 5: Main effects for rankings. Silver highlighted rows show at least marginal significance.

Ranking	H	p	η^2	Condition	Mean	SD
Preference	16.85	<.001	.32	Physical	1.39	.7
				Hybrid	2.11	.68
				Virtual	2.55	.71
Readability	18.16	<.001	.34	Physical	1.39	.72
				Hybrid	2.06	.72
				Virtual	2.56	.62
Comfort	30.91	<.001	.58	Physical	1.17	.51
				Hybrid	2.17	.51
				Virtual	2.67	.59
Confidence	19.14	<.001	.36	Physical	1.33	0.69
				Hybrid	2.17	0.51
				Virtual	2.5	0.79

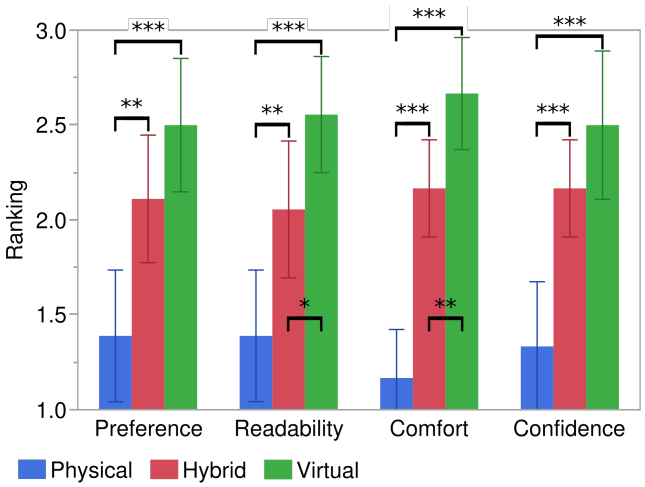

Figure 6: Average rankings for conditions. Ranking goes from 1 (Most Preferred) to 3 (Least Preferred). Error bars represent the confidence interval (95%).

anything on the monitors at a glance," Hybrid was rated higher than Virtual, finding no difference against Physical. This indicates that having the virtual monitors closer to the smaller central physical monitor made it easier to glance at the monitors.

Our third hypothesis was that users would have a measurably higher acceptance of Hybrid compared to Virtual (**H3**). Our results show some weak evidence to support H3. We found a marginally significant preference for Hybrid over Virtual on the ranking questionnaire. The statement "I would trust this condition to do serious work" was rated significantly higher for Hybrid than for Virtual. While Hybrid seems to be more accepted than Virtual, it is still less accepted than Physical. In our interview, the most preferred condition was Physical, and the second most preferred was Hybrid. Users liked the idea of having a single physical monitor and using virtual monitors to complement it, with the most frequent reason being lack of physical space and not being restricted on how much screen space they could add. There were complaints about the depth of the virtual monitors compared to the physical one, which matches prior research showing that displaying information in different depth layers leads to fatigue [11, 15].

Overall, our study suggests that virtual monitors created with current state of the art AR HWDs have decreased performance that is explained best by readability and head movement differences. Since we had to make our virtual monitors larger to make them readable, users had to turn their heads more to access and interact with the information on the side monitors in both the Virtual and Hybrid conditions. The low FOV of the AR HWD exacerbates this problem, and reduces the ability to use peripheral vision and rapid eye movements. Thus, we believe that increasing resolution should be the biggest priority for developers of AR headsets targeting this use case, as it would improve readability and reduce the need for larger virtual monitors. Increasing the field of view should also be prioritized.

6 Limitations

There are some limitations to the findings of this work that need to be considered. Our study compared different conditions of multi-monitor setups. We did not analyze how virtual monitors would perform under single monitor conditions, or how the hybrid extension of a physical monitor would compare against a single physical monitor. Based on our results, we believe that single virtual monitors would still face the same issues with low resolution and field of view, while a hybrid extension would present benefits over a single physical monitor, given the extra space. However, more research is required.

We also did not investigate the use of these conditions for long periods of time. Our experimental tasks had a length of about 10 minutes for each condition, which differs from full-time work schedules. Further research is needed to understand how some of the issues we discussed here manifest over larger periods, such as if the extended head turning in the virtual condition leads to neck fatigue, or if the change of depth in the hybrid condition leads to eye strain. Another limitation of this study is that we only evaluated a controlled task in the laboratory. Our results do not reveal anything about the user acceptance and social acceptability of using AR virtual monitors in real-world scenarios.

7 Conclusions and Future Work

In this work we investigated the feasibility of using AR virtual monitors for conducting productivity work. We implemented a prototype that enables the visualization of Windows 10 monitors both through physical monitors and a HoloLens 2 device. We conducted a user study to compare the use of Physical, Virtual, and Hybrid setups when performing an ecologically valid productivity task. We analyzed aspects such as performance, accuracy, head movements, and eye-gaze, and evaluated subjective user experience.

Results show that virtual monitors can be used now for real-world work, at least for short periods of time (discomfort could be a problem for longer work sessions). We did not find significant differences in task accuracy between our conditions. However, virtual monitors were both slower (about 14% more time overall) and required more head turning than Physical, while Hybrid represented a middle ground between the two.

For future work, it is important to study the impacts of using virtual monitors for longer periods of time, especially to understand if the increased head turning leads to fatigue. On the other hand, we believe that real benefits can be achieved by using virtual monitors to extend a single physical monitor into the periphery for productivity tasks, since such setups would represent an important win for people that use laptops for everything. As laptop screen sizes are small and high resolution, they could be used as primary monitors, with virtual monitors being used for other tasks. Additional research is needed to verify these hypothesized benefits.

Acknowledgments

This research is partially funded by a Microsoft Productivity Research grant and NSF grant #CSSI-2003387. Thanks to Lei Zhang for illustrating the concept.

REFERENCES

[1] C. Andrews, A. Endert, and C. North. Space to Think: Large High-Resolution Displays for Sensemaking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pp. 55–64. Association for Computing Machinery, New York, NY, USA, 2010. event-place: Atlanta, Georgia, USA. doi: 10.1145/1753326.1753336

[2] C. Andrews, A. Endert, B. Yost, and C. North. Information visualization on large, high-resolution displays: Issues, challenges, and opportunities. *Information Visualization*, 10(4):341–355, 2011. _eprint: https://doi.org/10.1177/1473871611415997. doi: 10.1177/1473871611415997

[3] R. Ball and C. North. Effects of Tiled High-Resolution Display on Basic Visualization and Navigation Tasks. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '05, pp. 1196–1199. Association for Computing Machinery, New York, NY, USA, 2005. event-place: Portland, OR, USA. doi: 10.1145/1056808.1056875

[4] R. Ball, C. North, and D. A. Bowman. Move to Improve: Promoting Physical Navigation to Increase User Performance with Large Displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pp. 191–200. Association for Computing Machinery, New York, NY, USA, 2007. event-place: San Jose, California, USA. doi: 10.1145/1240624.1240656

[5] M. Bellgardt, S. Pick, D. Zielasko, T. Vierjahn, B. Weyers, and T. W. Kuhlen. Utilizing immersive virtual reality in everydaywork. In *2017 IEEE 3rd Workshop on Everyday Virtual Reality (WEVR)*, pp. 1–4, 2017. doi: 10.1109/WEVR.2017.7957708

[6] X. Bi and R. Balakrishnan. Comparing Usage of a Large High-Resolution Display to Single or Dual Desktop Displays for Daily Work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pp. 1005–1014. Association for Computing Machinery, New York, NY, USA, 2009. event-place: Boston, MA, USA. doi: 10.1145/1518701.1518855

[7] V. Biener, D. Schneider, T. Gesslein, A. Otte, B. Kuth, P. O. Kristensson, E. Ofek, M. Pahud, and J. Grubert. Breaking the Screen: Interaction Across Touchscreen Boundaries in Virtual Reality for Mobile Knowledge Workers. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2020. doi: 10.1109/TVCG.2020.3023567

[8] S. Butscher, S. Hubenschmid, J. Müller, J. Fuchs, and H. Reiterer. Clusters, Trends, and Outliers: How Immersive Technologies Can Facilitate the Collaborative Analysis of Multidimensional Data. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 1–12. Association for Computing Machinery, New York, NY, USA, 2018. event-place: Montreal QC, Canada. doi: 10.1145/3173574.3173664

[9] G. Cetin, W. Stuerzlinger, and J. Dill. Visual Analytics on Large Displays: Exploring User Spatialization and How Size and Resolution Affect Task Performance. In *2018 International Symposium on Big Data Visual and Immersive Analytics (BDVA)*, pp. 1–10, 2018. doi: 10.1109/BDVA.2018.8534027

[10] M. Czerwinski, G. Smith, T. Regan, B. Meyers, G. G. Robertson, and G. K. Starkweather. Toward characterizing the productivity benefits of very large displays. In *Interact*, vol. 3, pp. 9–16, 2003.

[11] A. Eiberger, P. O. Kristensson, S. Mayr, M. Kranz, and J. Grubert. Effects of Depth Layer Switching between an Optical See-Through Head-Mounted Display and a Body-Proximate Display. In *Symposium on Spatial User Interaction*, SUI '19. Association for Computing Machinery, New York, NY, USA, 2019. event-place: New Orleans, LA, USA. doi: 10.1145/3357251.3357588

[12] A. Endert, L. Bradel, J. Zeitz, C. Andrews, and C. North. Designing Large High-Resolution Display Workspaces. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, pp. 58–65. Association for Computing Machinery, New York, NY, USA, 2012. event-place: Capri Island, Italy. doi: 10.1145/2254556.2254570

[13] S. Feiner, B. MacIntyre, M. Haupt, and E. Solomon. Windows on the World: 2D Windows for 3D Augmented Reality. In *Proceedings of the 6th Annual ACM Symposium on User Interface Software and Technology*, UIST '93, pp. 145–155. Association for Computing Machinery, New York, NY, USA, 1993. event-place: Atlanta, Georgia, USA. doi: 10.1145/168642.168657

[14] N. Fereydooni and B. N. Walker. Virtual Reality as a Remote Workspace Platform: Opportunities and Challenges. Aug. 2020.

[15] J. L. Gabbard, D. G. Mehra, and J. E. Swan. Effects of AR Display Context Switching and Focal Distance Switching on Human Performance. *IEEE Transactions on Visualization and Computer Graphics*, 25(6):2228–2241, 2019. doi: 10.1109/TVCG.2018.2832633

[16] J. Grubert, T. Langlotz, S. Zollmann, and H. Regenbrecht. Towards Pervasive Augmented Reality: Context-Awareness in Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics*, 23(6):1706–1724, 2017. doi: 10.1109/TVCG.2016.2543720

[17] J. Grubert, E. Ofek, M. Pahud, and P. O. Kristensson. The Office of the Future: Virtual, Portable, and Global. *IEEE Computer Graphics and Applications*, 38(6):125–133, 2018.

[18] Z. Li, M. Annett, K. Hinckley, K. Singh, and D. Wigdor. HoloDoc: Enabling Mixed Reality Workspaces That Harness Physical and Digital Content. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 1–14. Association for Computing Machinery, New York, NY, USA, 2019. event-place: Glasgow, Scotland Uk. doi: 10.1145/3290605.3300917

[19] F. Lu, S. Davari, L. Lisle, Y. Li, and D. A. Bowman. Glanceable AR: Evaluating Information Access Methods for Head-Worn Augmented Reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 930–939, 2020. doi: 10.1109/VR46266.2020.00113

[20] T. Mahmood, E. Butler, N. Davis, J. Huang, and A. Lu. Building Multiple Coordinated Spaces for Effective Immersive Analytics through Distributed Cognition. In *2018 International Symposium on Big Data Visual and Immersive Analytics (BDVA)*, pp. 1–11, 2018. doi: 10.1109/BDVA.2018.8533893

[21] M. Mcgill, A. Kehoe, E. Freeman, and S. Brewster. Expanding the Bounds of Seated Virtual Workspaces. *ACM Trans. Comput.-Hum. Interact.*, 27(3), May 2020. Place: New York, NY, USA Publisher: Association for Computing Machinery. doi: 10.1145/3380959

[22] E. Ofek, J. Grubert, M. Pahud, M. Phillips, and P. Kristensson. Towards a Practical Virtual Office for Mobile Knowledge Workers. In *Microsoft New Future of Work 2020 Symposium*.

[23] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stesin, and H. Fuchs. The Office of the Future: A Unified Approach to Image-Based Modeling and Spatially Immersive Displays. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '98, pp. 179–188. Association for Computing Machinery, New York, NY, USA, 1998. doi: 10.1145/280814.280861

[24] P. Reipschläger, T. Flemisch, and R. Dachselt. Personal Augmented Reality for Information Visualization on Large Interactive Displays. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2020. doi: 10.1109/TVCG.2020.3030460

[25] G. Robertson, M. Czerwinski, P. Baudisch, B. Meyers, D. Robbins, G. Smith, and D. Tan. The large-display user experience. *IEEE Computer Graphics and Applications*, 25(4):44–51, 2005. doi: 10.1109/MCG.2005.88

[26] A. Ruvimova, J. Kim, T. Fritz, M. Hancock, and D. C. Shepherd. "Transport Me Away": Fostering Flow in Open Offices through Virtual Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pp. 1–14. Association for Computing Machinery, New York, NY, USA, 2020. event-place: Honolulu, HI, USA. doi: 10.1145/3313831.3376724

[27] D. S. Tan, D. Gergle, P. Scupelli, and R. Pausch. With Similar Visual Angles, Larger Displays Improve Spatial Performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pp. 217–224. Association for Computing Machinery, New York, NY, USA, 2003. event-place: Ft. Lauderdale, Florida, USA. doi: 10.1145/642611.642650

[28] J. R. Wallace, D. Vogel, and E. Lank. Effect of Bezel Presence and Width on Visual Search. In *Proceedings of The International Symposium on Pervasive Displays*, PerDis '14, pp. 118–123. Association for Computing Machinery, New York, NY, USA, 2014. event-place: Copenhagen, Denmark. doi: 10.1145/2611009.2611019