



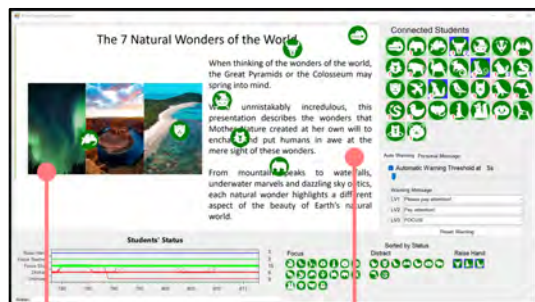
Tutor In-sight: Guiding and Visualizing Students' Attention with Mixed Reality Avatar Presentation Tools

Santawat Thanyadit*
job.santawat@gmail.com
Durham University
Durham, United Kingdom

Matthias Heintz
mmh21@leicester.ac.uk
University of Leicester
Leicester, United Kingdom

Effie Lai-Chong Law
lai-chong.law@durham.ac.uk
Durham University
Durham, United Kingdom

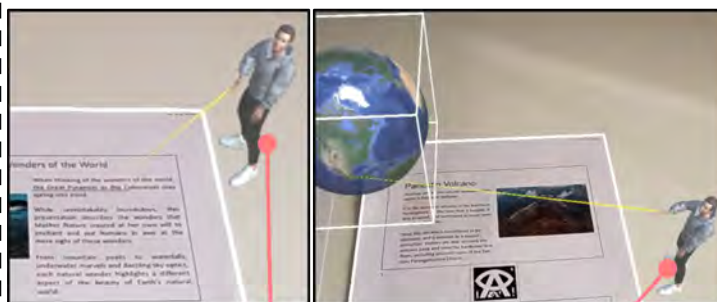
Teacher's Dashboard



The Dashboard is compatible with PowerPoint, allowing teachers to use familiar tools.

The Dashboard visualises students' eye gaze data, allowing the teacher to understand the class's attention.

Student's MR View



Avatar uses auto-generated gaze, gestures, and movement to direct the students' attention directly to the material being presented.

Avatar adjusts the position according to the material being presented and the student's point of view.

Figure 1: Tutor In-sight teacher's dashboard (left) and student's mixed reality view (right): The dashboard visualises students' eye gaze data, allowing the teacher to understand the class's attention. The in-situ avatar in the student's MR view uses auto-generated gaze, gestures, and movement to direct the students' attention directly to the material being presented.

Abstract

Remote conferencing systems are increasingly used to supplement or even replace in-person teaching. However, prevailing conferencing systems restrict the teacher's representation to a webcam live-stream, hamper the teacher's use of body-language, and result in students' decreased sense of co-presence and participation. While Virtual Reality (VR) systems may increase student engagement, the teacher may not have the time or expertise to conduct the lecture in VR. To address this issue and bridge the requirements between students and teachers, we have developed *Tutor In-sight*, a Mixed Reality (MR) avatar augmented into the student's workspace based on four design requirements derived from the existing literature, namely: integrated virtual with physical space, improved teacher's co-presence through avatar, direct attention with auto-generated body language, and usable workflow for teachers. Two user studies were conducted from the perspectives of students and teachers to

determine the advantages of Tutor In-sight in comparison to two existing conferencing systems, Zoom (video-based) and Mozilla Hubs (VR-based). The participants of both studies favoured Tutor In-sight. Among others, this main finding indicates that Tutor In-sight satisfied the needs of both teachers and students. In addition, the participants' feedback was used to empirically determine the four main teacher requirements and the four main student requirements in order to improve the future design of MR educational tools.

CCS Concepts

• **Applied computing** → **Computer-assisted instruction; Distance learning; • Human-centered computing** → **Mixed / augmented reality; User studies.**

Keywords

Virtual Avatar, Remote Presentation, Augmented Reality

*Also with King Mongkutt's University of Technology Thonburi, Thailand

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI '23, April 23–28, 2023, Hamburg, Germany

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9421-5/23/04...\$15.00

<https://doi.org/10.1145/3544548.3581069>

ACM Reference Format:

Santawat Thanyadit, Matthias Heintz, and Effie Lai-Chong Law. 2023. Tutor In-sight: Guiding and Visualizing Students' Attention with Mixed Reality Avatar Presentation Tools. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, April 23–28, 2023, Hamburg, Germany. ACM, New York, NY, USA, 20 pages. <https://doi.org/10.1145/3544548.3581069>

1 Introduction

Worldwide, schools, enterprises, and government agencies utilise videoconferencing tools such as Zoom, Google Meet, or Microsoft Teams on a daily basis. Videoconferencing system is therefore a major tool for many individuals to work, learn, teach, and meet. However, Riva et al. [60] claim that excessive use of videoconferencing system may result in burnout or "Zoom Fatigue." Furthermore, many university students, who had experienced videoconference-based learning during the pandemic, commented that educational quality has declined in comparison to face-to-face learning [1, 7, 22, 27], indicating the inability of these conferencing tools to reproduce the experience of co-presence between teacher and students. This well-recognised issue can be attributed to the lack of sensory cues such as gaze and posture [2, 12]. Since body language has been found to affect audience engagement and attention [4], decreased student engagement can be attributed to a reduced webcam feed of the teacher, which may obscure any body language used by the teacher (e.g., eye contact, gestures, and movement).

Virtual environments may alleviate the issue of low social presence and student engagement by providing teachers with interactive meeting places to engage with their students. In the virtual world, a teacher can be represented by a virtual avatar with head and hands tracking, providing the opportunity to express a wider range of body language and creating a greater sense of teacher's co-presence for students. Web-based virtual world platforms have become increasingly popular, with systems like VirBela and Mozilla Hubs being used to conduct academic conferences with hundreds of attendees (e.g., ISMAR, IEEEVR, iLRN). These platforms also include a plethora of social features, such as customised avatars, game-like pre-set gestures, and spatial voice chat, which enhance the immersive experience for users. Nonetheless, Jensen and Konradsen [34] found that current VR head-mounted displays (HMDs) are largely designed for entertainment. They are not intended for classroom usage and need a degree of technical expertise that many teachers find challenging. Moreover, recent surveys by Radianti et al. [56] and Yang et al. [75] identified pedagogical issues, including reluctance from schools and teachers, and the inflexibility of systems with regard to modifying instructional material. Additionally, wearing a VR headset, which is required to achieve immersion and co-presence on these platforms, may hinder students' ability to interact with learning materials in the real world, for example, their vision may be obstructed while taking notes, rendering these platforms problematic for educational settings. In addition, Radianti et al. [56] found that students' VR experiences vary substantially with personality variables, which have an impact on how virtual reality is perceived and thus on learning results.

These observations indicate that the existing systems are ineffective at meeting the requirements of both teachers and students. On the one hand, a videoconference system is an effective and familiar tool for teachers, but it can reduce the teacher's presence, causing students' motivation to dwindle. On the other hand, although VR provides immersion and motivates students, it is challenging for teachers to utilise.

Therefore, we are motivated to find a solution that satisfies these requirements. For students we propose to use a Mixed Reality (MR) HMD to augment a teacher's presence as a miniature 3D avatar

(dubbed Tutor In-sight) next to the students' printout of the materials in situ, allowing the students to focus on the materials and the teacher in their workspace simultaneously (Figure 1 right). The Tutor In-sight teacher avatar was designed to use a variety of body language, such as eye contact, gestures, and movement, to direct the students' attention to the presented information (e.g., text, images, and 3D objects). Our assumption is that the teacher's miniature avatar can enhance the sense of co-presence and the quality of interaction between the teacher and students, leading to improved student engagement. Furthermore, the actions of the avatar are automatically generated from the teacher's actions in Microsoft PowerPoint, ensuring compatibility with the teacher's presentation experience, especially as it allows the use of a familiar tool with minimal effort. In addition, students' eye tracking data is visualised in the teacher's dashboard (Figure 1 left) to help the teacher gain insight into the entire class with a regular graphical user interface on a 2D screen without the need to wear any HMD.

To create Tutor In-sight, we identified four design requirements from the relevant literature (Section 2) and then compared the initial prototype to a presentation delivered via an existing videoconferencing tool and VR conferencing tool (Mozilla Hubs) in two user studies, one with each user group, to understand how teachers use remote presentations and how students follow presentations in a remote setting across the different tools. Suggestions and observations from these comparisons were utilised to come up with refinements of the design of Tutor In-sight and gather requirements for future educational MR presentation systems.

The main contributions of our research work presented in this paper are four-fold:

- Four design requirements of Tutor In-sight derived from the literature review for creating a novel MR presentation system that fosters a sense of co-presence, improves engagement, and enhances the common ground-establishing process between a teacher and students during a remote presentation.
- Design, implementation, and evaluation of the Tutor In-sight's MR avatar presentation and body language (gaze, gesture, and movement) system, which is created from the presentation given by the teacher with a common presentation tool such as PowerPoint. The system can adapt autonomously and in real-time to the content being presented to guide students' attention in-situ.
- Design, implementation, and evaluation of the Tutor In-sight's accompanied teacher dashboard that visualises students' eye gaze data to enable a teacher to gain an overview of students' attention.
- Four main design requirements derived empirically from the teacher's evaluation feedback and four main design requirements from the students' evaluation feedback to enhance the future design of educational MR tools.

2 Related Work

In this section we present the literature review from which we have derived the four design requirements (R) for the design of Tutor In-sight (Section 3).

2.1 Integrate Virtual with Physical Learning Space (R1)

Real-time collaboration technologies can be deployed in a face-to-face, video-based, or audio-only context. For most collaborative tasks, face-to-face outperforms audio-only because the visibility of shared workspace enables collaborators to point directly to the workspace rather than describing the reference [28, 64]. However, the benefit of sharing workspace through videoconferencing is less clear. For instance, Fussell and Setlock [28] observed that adding a video to the remote collaborative task had no substantial benefit over audio-only since the video had limited visibility, and collaborators could not infer a target pointed at from their partners' gestures. With the advent of MR head-mounted displays, such as Microsoft HoloLens, researchers began using MR to combine remote virtual and local physical workspaces via a variety of remote embodiments, including sharing a remote user's arm [3, 30, 65], head [52, 54, 55], or body state [36, 48, 51].

In a typical remote lecture scenario, a teacher shares her or his presentation with students, who concurrently observe the teacher's presentation and perform tasks (e.g., taking notes) in their own workspace. Current research shows that videoconferencing may be insufficient to support this basic scenario. For instance, demonstrating on a computer screen could be difficult due to the restricted camera angle and the limited capacity to refer to or zoom in on physical objects on the workspace.

To address this problem, Villanueva et al. [69] deployed semi-automated robots with a screen that moved and augmented students' workspaces to assist circuit construction tasks. The robots could be controlled remotely by a teacher to improve the ground-establishing process. In another example, Thanyadit et al. [67] examined learning from a demonstration in VR in three different workspace setups: video (2D), avatar (3D), and transparent avatar (3D). The results suggested that participants preferred the transparent avatar condition over the video one because in the latter individuals had to divide their attention between their workspace and the video, aggravating the cognitive load. These works demonstrate the advantages of an integrated workspace that enables students to see guided instructions in their own workspace, enabling undivided attention. Tutor In-sight follows this design approach by putting an avatar of the teacher into the students' workspace to guide their attention.

2.2 Direct Attention with Body Language (R2)

Effective presenters demonstrate their abilities not just through their words, but also through nonverbal communication such as body gestures, eye contact, and even the movement on stage [13, 44]. Recent research has utilised body language to estimate the speaker's presenting abilities as well as predict audience engagement and attention [4, 24, 58], indicating that body language during a presentation is just as significant as the material being delivered.

Body language consists of five main components: gesture, facial expression, eye gaze, lips, and movement (proxemics) [13, 44]. This research focuses on pointing gestures, eye gaze, and movement since they are highly related to presentations. Pointing is a basic ground-establishing strategy that is often used in a remote guiding application [8, 53, 65], enabling the presenter to direct the viewer's

attention to the current content of interest. Eye gazing, like pointing, is used as an attention signal to help viewers understand the presenter's speech and develop collaborative attention [5, 8, 35, 52]. It is also used to monitor the viewer's attention in order to establish mutual gaze, which has been shown to enhance viewer engagement [29]. Movement of the presenter is often lost in remote presentations, and the presenter is often shown as an avatar with a generic walking animation, for example, in systems like VirBela. The presenter's avatar is used because the presenter's physical surroundings may mismatch the presentation space, preventing direct movement mapping. Nonetheless, this walking animation assists the viewer in directing their attention and allows them to infer the next point of attention [67]. The Tutor In-sight avatar incorporates pointing gestures, eye gaze, and movement next to the presentation material to better guide the viewer's attention.

2.3 Improve Co-presence through Avatar (R3)

Virtual embodiment, or virtual avatar, has been used to represent both viewers and presenters in a variety of virtual educational applications, including physical education [23, 26, 39], scientific laboratories [50, 67, 71], and safety training systems [20, 40, 43]. A theoretical framework of co-presence or social presence was introduced by Short et al. [63]. In a virtual environment, co-presence is commonly characterized as sense of being with another [10, 47]. The notion of co-presence is applicable to polyadic (multi-party) as well as dyadic interaction [9], which, in the context of our work, is between a teacher and a student. The related publications provide growing evidence that a virtual avatar can enhance co-presence by facilitating natural communication with improved social interaction, spatial awareness, and collaboration-based learning [16], as well as increased engagement [19]. The improved sense of co-presence provided by a virtual avatar is often associated with its capacity to communicate via body language (Section 2.2).

Another desirable feature of virtual avatars is their adaptability, as users may want to customise their avatar appearance by changing the style (cartoon-like, abstract, or realistic [76]), visible body parts (whole body or half-body [18]), and size [53]. Alternatively, instead of user-initiated adaptation, a virtual avatar can self-adapt to its surroundings or collaborators through a method known as "redirection." For example, ObserVAR [68] displays several students as virtual avatars, which can be repositioned according to the individual student's gaze to optimise visualisation, thereby reducing the teacher's cognitive load while viewing the whole virtual classroom. Kim et al. [38] show how to redirect the gaze of virtual avatars in order to boost the remote attendee's social presence. Mini-me [53] represented a remote helper using a pair of virtual avatars, one life-size and the other as a miniature representation. The miniature helper's pointing gesture is redirected to the same location as that of the life-size helper, facilitating communication between a remote VR user and a local MR user. A further study by the same research group [54] found that the local user preferred that the remote user would be represented by a miniature avatar with visual cue to show remote user's view frustum.

Tutor In-sight expands on previous work, particularly Mini-me, by using a miniature avatar that uses gaze, position, and gesture redirection methods to increase the viewer's engagement and adapt

to the viewer's workplace. One of the important novel aspects of Tutor In-sight is that its avatar's body language (gaze, gesture, and movement) is created from interactions of the teacher with the presentation slides (e.g., in Microsoft PowerPoint) and adapts autonomously to the content being presented, rather than representing actual human body language by body tracking, which can be costly in terms of resources, training, and setup.

2.4 Enhance Usable Design for Teacher (R4)

Regarding the use of VR and MR presentations in educational and corporate settings, recent surveys [32, 34, 56, 57] have suggested that many teachers and educators may lack the technical expertise or relevant experience required to produce a VR or MR experience. Additionally, prolonged use of VR and MR equipment may create discomfort for presenters, discouraging them from adopting MR as a presentation tool altogether. Therefore, a desktop alternative and compatibility with more established presentation tools such as PowerPoint are necessary in order to minimise the barrier of entry and improve adoption of MR presentation tools. Recent MR presentation and production research studies have started to prioritise this design component. For example, Nebeling et al. [45] alleviated these issues by providing pre-set and pre-calibrated configurations for the teacher, streamlining the live production and post-production processes associated with extended reality production. In another example, Woodworth et al. [73] integrated eye, head, and hand trackers into the presenter's desktop interface to generate avatar movement for a VR meeting application, allowing presenters to continue using a familiar desktop interface while their avatar presents in VR. Similarly, Tutor In-sight generates an avatar's body language from the mouse and keyboard inputs during the presentation, enabling the presenter to utilise established presentation tools. Furthermore, Tutor In-sight takes into account the material being presented in real-time as well as the student's head movement and eye gaze to produce avatar animation, allowing the Tutor In-sight avatar to establish eye-contact and remain inside the students' field of vision to promote co-presence.

Another barrier preventing teachers from employing VR or MR in the classroom is a reduced capacity to monitor students in virtual worlds. Teachers cannot monitor students as effectively as they could in the real world because bodily signs such as eye gazing, movement, and facial expression are typically portrayed loosely or not at all in the virtual environment. Furthermore, the lack of physical constraints enables student avatars to appear in an essentially limitless space, which often leads to confusion and a heavy cognitive burden for teachers. Recent research has begun to include a feature to boost teachers' awareness of the virtual environment. For instance, ObserVAR [68] grouped students' avatars based on the students' gaze targets to avoid virtual clutter. In another example, Broussard et al. [15] incorporated visual interfaces that displayed key information about students in the teachers' field of vision to assist teachers in better monitoring students in a VR classroom. Tutor In-sight has a teacher desktop dashboard to show students' eye gaze, enabling the teacher to gain an overview of the students' attentiveness throughout the lesson.

3 Design and Implementation of Tutor In-Sight

Tutor In-sight is composed of two subsystems: the teacher's subsystem and the students' subsystem (Figure 2). The teacher's subsystem consists of a dashboard that visualises students' eye gazing data and a PowerPoint plugin that transmits slide and mouse movement data to the students' subsystem. We developed the students' subsystem with the Unity engine 2020.3.13f1 and deployed it on the Microsoft HoloLens 2, which has an optical see-through display, spatial mapping, and eye tracking capabilities.

Tutor In-sight is designed around the premise of a remote real-time presentation between a teacher using a PowerPoint presentation and a dashboard (Figure 2 left) and students using an MR headset (Figure 2 right). The students are expected to obtain a copy of the teacher's slide (either digitally on a tablet or a physical printout) prior to the lecture for annotation purposes, which is a regular arrangement in multiple lecture-style classrooms. For such an MR-based presentation arrangement, we have created our system fulfilling the four design requirements (**R1**, **R2**, **R3**, and **R4**) (details see Section 2):

3.1 Teacher's subsystem (R4)

For the teacher's subsystem, we developed a PowerPoint plugin that processes slide contents and tracks mouse movements while the presentation is active. When the mouse is hovered over any media type (text, picture, 3D objects) on the slide, messages are generated and broadcast over the network; these messages are then received and processed by HoloLens 2 to generate the Tutor In-sight avatar's actions and augmentations based on the content of each message. The printout materials are marked with the Vuforia Image tag [70] (Vumarks), which encodes the surface size, page number, and number of slides per page, enabling presenters to specify the physical printout size for their presentation. We used two slides per page throughout this research as a good compromise between augmentation size and the number of printouts; nevertheless, 1-4 slides per page are valid alternatives. The teacher merely has to prepare the presentation as usual and then place the Vuforia tag on each printed slide page as an extra step to utilise Tutor In-sight, fulfilling R4.

We created a teacher's dashboard, inspired by the prior work on remote presentation systems [61] and real-time classroom monitoring [15, 33, 77], to enable teachers to quickly determine which students were paying attention (looking at the slides), and which were not, as a metaphor for the way a teacher scans the class when lecturing in a traditional classroom. The proposed attention dashboard should be used during the lecture, to offer real-time feedback to the teacher, and after the lecture, to complement a learning dashboard, which provides teachers with post-class statistics, like the ones presented by Mazza and Dimitrova [42] or Khakaj et al. [74].

Our proposed teacher dashboard visualises eye-gaze data as a proxy for students' attention since previous studies used eye-gaze to identify mind wandering [11], distraction [6], and interaction strategies [46]. However, since the dashboard presents the information in real-time, we have chosen to show the eye location data as they are. Figure 3 depicts the dashboard user interface in default mode and Figure 4 shows an example of changes when the teacher clicks on an individual student. The students' gaze panel (Figure 3a)

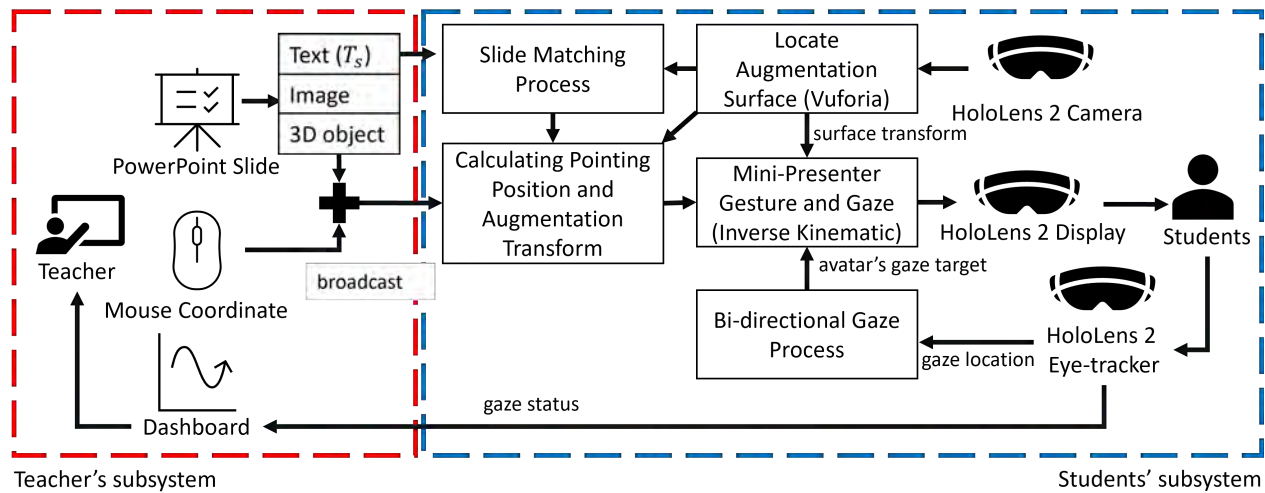


Figure 2: An overview of the Tutor In-sight system: the left-hand side red-dashed-line box shows the teacher's subsystem on a desktop PC, and the right-hand side blue-dashed-line box shows the students' subsystem on the Microsoft HoloLens 2.

displays where the students' eyes are concentrated on the presentation in real time, with each student represented by their profile picture (anonymous icons are used instead of students' pictures in the example). The overview panel (Figure 3c) displays the currently connected students' profile pictures, with the border of each student's profile picture being colour-coded [41] to reflect the current status of the student. The colour green indicates that the student is looking at the slide or the virtual avatar of the teacher. The blue colour indicates that the student wants to ask a question or has sent an inquiry through the Q & A panel (Figure 4c and Figure 5j). If the students have sent a question, a blue number appears next to the student's profile picture to indicate the number of inquiries (Figure 4a, lower right corner). The red border shows that the student's attention is elsewhere and that he or she may be distracted, whereas the intensity of the colour indicates how long the student has been distracted. The status graph [77] in the bottom left (Figure 3b) displays the number of students in each status in real time, allowing the teacher to better grasp the classroom dynamics. In addition to this graph, the sorting panel (Figure 3e) groups students according to their status, allowing the teacher to see who is distracted or requiring the teacher's attention at a glance. To assist the teacher in managing the students' attention, the warning panel (Figure 3d) may be used to create automated warning messages. An automated notification is then issued to students who are distracted for longer than the set time threshold. The teacher may prepare three different levels of warning based on how many times the students are distracted in one session. The current warning level is shown in red next to the student's profile picture to call for the teacher's attention (Figure 4a, lower left corner). When the teacher clicks on an individual student's picture, the Q & A panel (Figure 4c) appears, enabling the teacher to read the student's queries and write back to them; the student will see the teacher's private response as shown in Figure 5e. For other students, the teacher's avatar will be in an idle posture (Figure 5a), lip-syncing with the teacher and maintaining eye contact with students who gaze at the avatar. Furthermore,

the individual student's status history graph (Figure 4b) is shown in place of the students' status graph (Figure 3b) to assist the teacher in understanding the selected student's attention during the whole lecture.

3.2 Augmenting the Students' Workspace (R1)

To meet R1, Tutor In-sight must first find the student's workspace and discover what the student has available in their workspace. To find the student's workspace, the cameras on HoloLens 2 are used to detect the Vuforia tag on the printed slide in order to generate the augmentation surface for all other augmentations to be put on (depicted as the white frame in Figure 5a). If the slide in front of the student corresponds to the slide currently presented, the mouse coordinates from PowerPoint are transformed into the physical slide coordinates that the Tutor In-sight avatar will be pointing to. If the teacher is pointing to the text, an augmentation underlining the text appears (Figure 5b). When the teacher points to an image in the slide, a frame is created around the image on the physical slide to draw the student's attention to it (Figure 5f). If the teacher does not point to anything on the slide, the Tutor In-sight avatar will face the student. The avatar is lip-synced with the teacher to create a realistic representation of the presenting teacher. To make better use of the MR environment, the teacher may add 3D objects to the presentation and provide a link to each 3D item in the alternative text option. Tutor In-sight will then use this link to display the 3D objects at runtime when the teacher points to the 3D object on the slide (Figure 5g). The student may move, rotate, or scale the 3D object to get a better look.

According to cognitive theories [37], the amount of attention paid to an event is a good predictor of whether or not it will be recalled consciously later; hence, Tutor In-sight is designed to encourage the student to concentrate as much as possible on the avatar and the presentation material. If students look elsewhere for a period of time that exceeds a threshold set in the teacher dashboard, a notification (Figure 5h) is displayed in front of them to remind

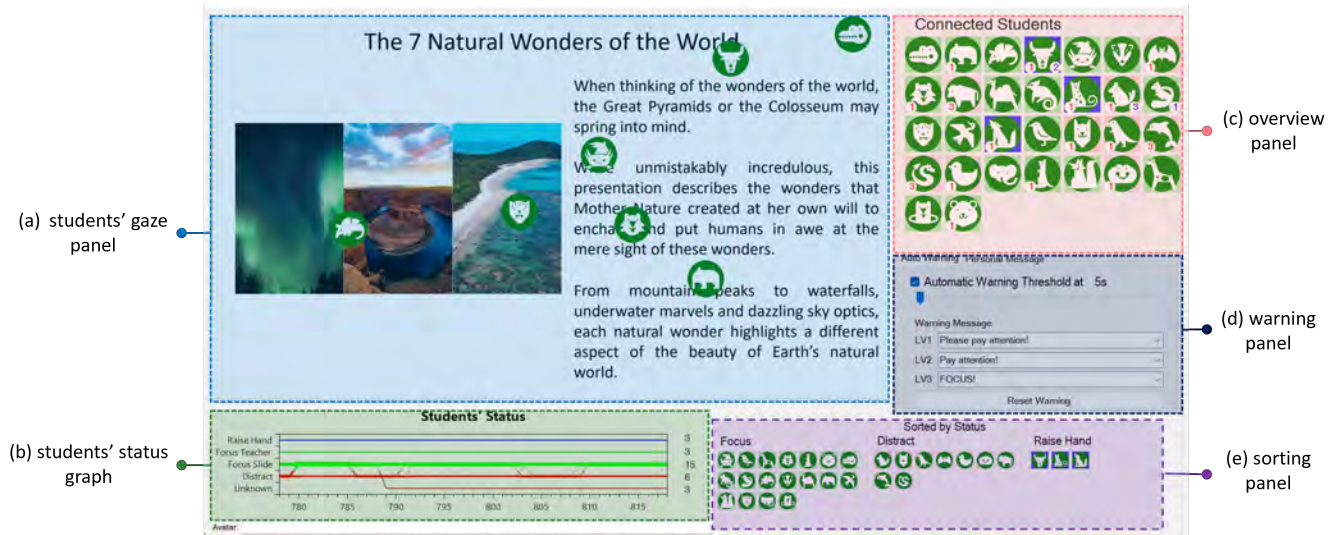


Figure 3: (a) The students' gaze panel shows students' eye gaze position in real-time (b) The students' status graph displays the dynamic of the class (i.e., aggregated data of all students) over time (c) The overview panel displays the currently connected students' profile pictures, colour-coded based on their status (d) The warning panel allows the teacher to configure automatic warning messages (e) The sorting panel groups students according to their status.

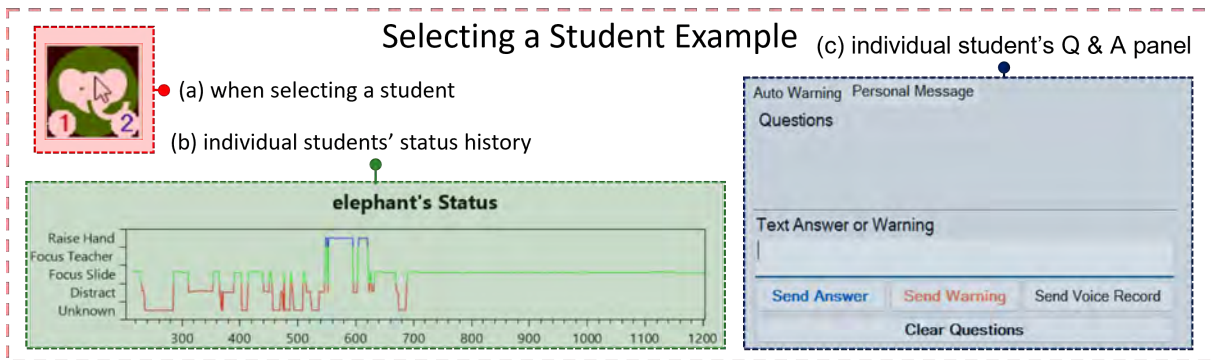


Figure 4: Specific student data example: (a) When a teacher selects a student, the warning panel and status graph change to show the selected student's information (b) The panel shows the history of student's status during the entire lecture (c) The Q & A panel allows the teacher to respond to the student's questions.

the students and lead their attention back to the current teaching topic. Additionally, an automated warning message is shown as a speech bubble on the teacher's avatar with the student's name on it so that students feel that the warning is for them (Figure 5e). An automated warning message is required to draw students' attention back to the materials in 3D spaces so that the key information that the teacher is highlighting is not missed. It also aims to reduce the need for the teacher to provide a verbal warning to catch the students' attention.

Furthermore, if the slide is not available in front of the student, Tutor In-sight displays a warning in the form of a textbox and suggests that the student turn the pages. To increase the flexibility in how the students follow the presentation (e.g., allowing them to look back to already presented slides), we add a "blackboard"-like

text augmentation to show the information of the presenting slide when the student's physical slide does not match the teacher's slide, as seen in Figure 5i. The Tutor In-sight avatar also points to these text augmentations based on the mouse locations in the PowerPoint slide. This text augmentation should keep the students informed when perusing other slides or searching for the current slide.

3.3 Creating The Avatar and Body Language (R2, R3)

We created a full-body avatar for Tutor In-sight using an open-online avatar building platform, Ready Player Me [59], which allows teachers to quickly create and personalise their avatar for their presentation. The full-body avatar is generated at run-time using Ready Player Me's link, which is then scaled down to one-tenth of

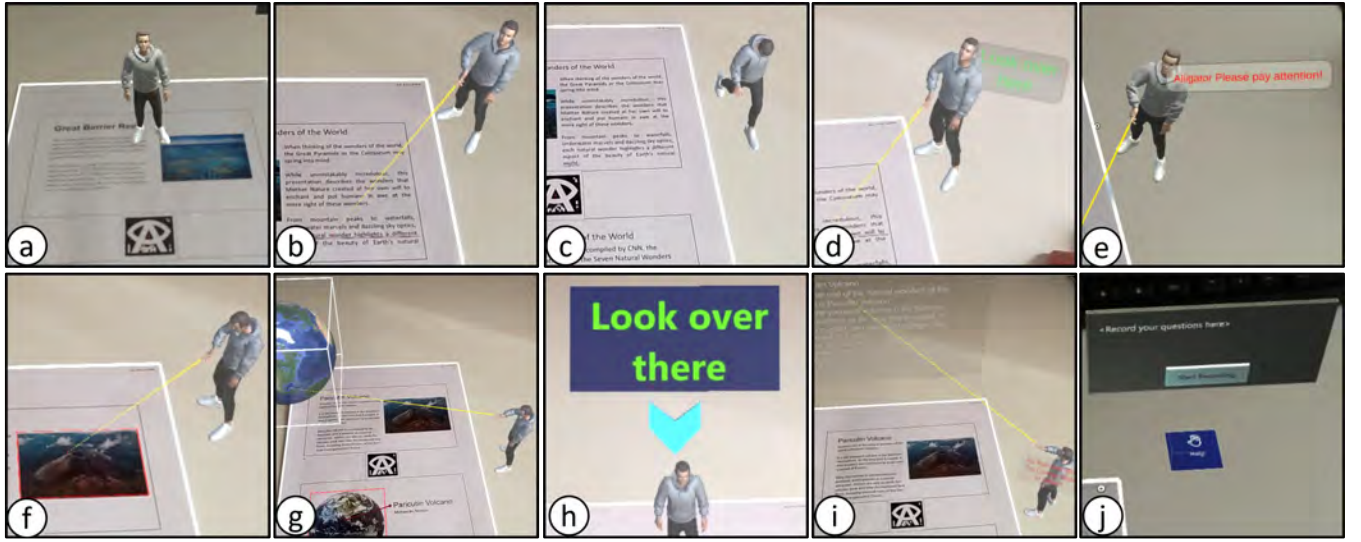


Figure 5: (a) Tutor In-sight’s avatar in an idle posture when not presenting; (b) the avatar points to the information currently presented during the presentation; (c) the avatar walks to the next slide; (d) the avatar reminds the student to focus on the slide via textbox; (e) the avatar warns the students according to the teacher’s setting or replies to a student’s question; (f) the avatar points to the same image the teacher is currently pointing their mouse pointer at; (g) the avatar points to the 3D globe at the same mouse position as on the 3D globe in PowerPoint; (h) a large textbox points back to the slide when the student is distracted from the presentation; (i) temporary “blackboard” augmentation for viewing current presentation content when a physical slide is not available in the student’s workspace e.g., because they read ahead; (j) Q & A panel for the students to ask questions or raise hand.

Table 1: Left: Mean gaze lengths of avatar on the gaze target in each phase. Right: Probabilities of avatar gazing toward each target in each phase. Data derived from the study by Andrist et al. [5].

Gaze Fixation Length (seconds)				Gaze Shift Probabilities			
phase	pointing position	student’s head	student’s gaze location	phase	to pointing position	to student’s head	to student’s gaze location
pointing	1.1	0.6	0.5	pointing	0.48	0.11	0.41
idle	1.2	1.7	0.6	idle	0.49	0.15	0.36

its original size to fit inside the students’ field of vision (R3). Furthermore, the Ready Player Me avatar supports sound processing, allowing lip-sync with the teacher’s voice to be conducted.

To meet R2, the Tutor In-sight avatar guides students using three forms of body language: pointing gesture, mutual gaze, and movement. Pointing is a basic ground-establishing gesture that is commonly employed in collaborative applications [52, 53, 65]. PowerPoint’s mouse coordinates are transformed into the physical slide’s pointing target. Inverse kinematics is then applied to the Tutor In-sight avatar’s shoulder, elbow, forearm, and upper torso to make its finger point to the same pointing target as on the PowerPoint slide. Additionally, a yellow line is added between the avatar’s finger and the pointing target for improved visibility. The Tutor In-sight avatar points to the slide when the teacher moves the mouse in PowerPoint until the teacher stops moving the mouse, then it returns to its idle position.

We utilised the eye tracker on HoloLens 2 to enable Tutor In-sight and the students to coordinate their gaze. In a visual environment, gaze is especially important for social presence [5, 38, 52]. During the presentation, we identified three probable avatar gaze targets: pointing position, student’s head, and student’s gaze location. Inspired by the work of Andrist et al. [5], a stochastic finite-state machine is used to determine the avatar’s gaze target and duration. Additionally, the state machine is divided into two phases in accordance with the teacher’s actions (pointing and idle). Gaze shifting and gaze duration are decided in each phase by the probability associated with each target, as indicated in Table 1 (selected from [5] human-human gaze data). After completing one gaze fixation, the next gaze target is chosen using a weighted random sampling. As with [5], heuristics are used to override the state machine when the student changes gaze target in order to better react to the student’s gaze during the pointing phase. The heuristics are as follows:

- (1) When the student gazes toward the pointing position, the avatar also gazes toward the pointing position.
- (2) When the student gazes toward another object while the avatar is pointing, the avatar gazes toward the pointing position. If the student looks at other objects for more than two seconds (based on the average gaze fixation length [5]) without looking at the presentation while the avatar is pointing, the avatar gazes toward the student and a warning is shown to guide the student's attention back to the presentation.
- (3) When the student gazes toward the avatar, the avatar gazes back toward the student. If the student's gaze fixates on the avatar for more than three seconds while the avatar is in the pointing phase, a small textbox (Figure 5d) is shown, asking the student to look at the pointing location.

In addition, the Tutor In-sight's avatar should reposition itself to remain inside the student's field of vision in order to ensure visibility and promote social presence. If the slide is detected in the student's workspace, the avatar will relocate to the top right-hand corner of the slide (Figure 5c) and will remain there until the teacher changes slides. In a pilot study, we discovered that the Tutor In-sight avatar should only move when the teacher changes a slide; otherwise, the movement becomes a distraction. On the other hand, if the slide is not detected, the Tutor In-sight avatar will utilise the blackboard feature (described in Section 3.2), and the avatar will reposition itself such that it is in-between the student's head and the blackboard, ensuring that the student can see both the pointing target on the blackboard and the avatar simultaneously.

4 User Study

To better understand the perspectives of teachers and students in using remote presentation tools across different setups from commercially available tools (videoconference and VR) to our proposed Tutor In-sight system, we conducted two user studies, one for teachers and one for students.

The research goals of the teacher user study were to:

- verify previously recognized impediments for adopting VR technology in remote presentations
- investigate if Tutor In-sight could mitigate these issues
- understand how teachers monitor students in different existing presentation tools
- determine if the Tutor In-sight's dashboard enhances this capability

The research goals of the student user study were to:

- better understand the influence of the teacher's representation on student engagement.
- compare our proposed Tutor In-sight avatar (with auto-generated body language) to two existing presentation tools in terms of usability, co-presence, and engagement.

In each of the studies, we performed a within-subjects study comparing the Tutor In-sight condition (TI) to the videoconference condition (VC) and the Mozilla Hubs VR condition (MH).

4.1 Teacher's perspective study

4.1.1 Setups of three conditions

To be able to collect feedback from teachers on AR and VR presenting techniques that they may not be familiar with, our study

was split into two phases: introduction and presentation. During the introduction phase, participants were given training on the fundamental capabilities of the presentation tools until they were able to do a basic presentation in each of the tools. To equip participants with an overview of the whole system, a brief explanation and a video introduction of the students' perspectives was also presented. Following a brief break, participants were instructed to make a mock presentation using a slide of their choice from the given PowerPoint file, each slide containing roughly 100 words. During the mock presentation, participants were told to keep an eye on the students' attention and to engage with them as if they were in a real remote classroom situation. In each of the remote presentation tools, we mimicked the presence of the students to make the environment as realistic as possible. We employed the recording feature of each presentation tool to ensure that the students would behave the same across participants.

For the VC condition, we imitated students' webcam streams by editing and combining volunteers' video feeds captured during a local university lecture or videoconference. The resulting video stream was shown next to the slide to resemble videoconferencing software when the share screen feature is used during a presentation (Figure 6a). For the TI condition, we captured the eye gaze data of the students and played it back to the participants throughout the presentation phase (Figure 3). During the presentation phase, an actor portraying a student connected in real-time to the Tutor In-sight system in order to replicate more accurately the behaviour of actual students attending the session. The actor was instructed to concentrate on the slide the teacher participant was talking about to behave as an attentive student. At the same time, the teacher participant also viewed pre-recorded random eye-gaze, representing inattentive students. Since the recording capability is unavailable in the MH condition, avatars with simple eye animation were deployed in the Mozilla Hubs room to mimic the students' presence (Figure 6b). Similar to the TI condition, an actor was connected in real-time to the Mozilla Hubs in order to simulate a student for the participants to observe live. We simulated 30 students in each condition to represent the average student number in a real school classroom [25, 66].

4.1.2 Procedures and participants

Six teacher volunteers were invited to participate in an in-person study at a local university (local participants). Seven additional teacher volunteers were invited to participate from remote locations (remote participants). The demographic data of the two groups are summarised in Table 2. All participants were university lecturers who had either never used AR or VR technology or considered themselves novices. However, all of the participants ranked their videoconference presentation skills as average or higher.

Their participation was voluntary without involving any reward or compensation. The main difference between the local participants and remote participants was that the local participants were asked to perform a mock presentation in VR (using Meta Quest 2) during the presentation phase of the MH condition, while remote participants used the web interfaces of Mozilla Hubs to perform the presentation in the MH condition instead due to the limited availability of VR-HMD.



Figure 6: (a) A simulated video conference presentation for the teacher's perspective (b) a simulated environment for VR presentation in the Mozilla Hubs.

Table 2: Demographic data of teacher participants (Note: [frequency] : [attribute])

	Local (n=6)	Remote (n=7)
Gender	4: female, 2: male	3: female, 4: male
Age	5: 31-40 yrs, 1: 41-50 yrs	5: 41-50 yrs, 1: 31-40 yrs, 1: 50-60 yrs
Teaching Subject	5: STEM, 1: Social Sciences	6: STEM, 1: Social Sciences
Teaching Experience	1: Pre-service 4: <5 years 1: >10 years	1: Pre-service, 1: <5 years, 1: 5-10 years, 4: >10 years

This research project was conducted in the context of the EU project ARETE. Participants were asked to sign a consent form and completed a demographic questionnaire upon arriving at the site of the evaluation or upon joining remotely. To avoid learning bias, the participants used the presentation tools (VC, TI, and MH) in a random order. Due to their familiarity with the VC condition, each participant spent less than five minutes listening to the introduction and giving a mock presentation. In the TI condition, participants spent around seven minutes on the introduction, including time to ask questions, but less than three minutes on the mock presentation, since they were accustomed with the 2D UIs and PowerPoint. In the MH condition, participants likewise spent roughly seven minutes on the introduction and questions; however, they spent about ten minutes on the mock presentation part, depending on how much difficulty they had navigating the 3D UIs. Following each presentation, participants completed a post-experiment questionnaire to collect subjective data. After participants performed presentations in all three conditions, we conducted a semi-structured interview with each participant to gather feedback using questions such as "What do you like most/least?" and "Which issue would you like to see improved the most?" The entire evaluation took roughly 50 minutes to complete. In the post-experiment questionnaire, we collected the subjective assessments shown in Table 3.

4.1.3 Hypotheses

Previous studies [34, 56] indicate that participants will have difficulty presenting in the VR (MH) condition. In contrast, even if

participants have never used Tutor In-sight before, we assumed they would be able to utilize it easily since it was designed to integrate with PowerPoint (H1). In addition, Tutor In-sight's dashboard was assumed to be able to enhance the monitoring ability of the participants (H2). Additionally, since participants in the MH condition can see the avatars of the students, they would have a stronger sense of student's presence in the MH condition (H3). Based on these assumptions and our design requirements, we hypothesized the following for the teacher evaluation for both local and remote groups:

H1 : There are significant differences among the three conditions in (a) task difficulty rating, (b) mental effort rating, (c) SUS score, (d) acceptable setup time consideration, favouring the VC and TI over MH condition.

H2 : There are significant differences among the three conditions in monitoring ability, favouring the TI condition.

H3 : There are significant differences among the three conditions in (a) perception of co-presence, (b) attention allocation, and (c) perceived message understanding, favouring the MH condition.

4.1.4 Results

We separated the data analysis between the local participants and remote participants due to the differences in user study procedure (Section 4.1.2). Align-and-rank ANOVA ($\alpha = 0.05$) [72] was used for non-parametric analysis, given that the data are not normally distributed, as shown by the results of Shapiro-Wilk test ($p < 0.05$). Bonferroni correction was used for post-hoc pairwise comparisons. Table 4 shows the descriptive statistics of the variables. We categorized each variable as system usability, cognitive effects, or experiential responses.

System Usability

Task difficulty: There were no significant differences in either local ($F_{2,10} = 1.50, p = 0.269, \eta p^2 = 0.231$) or remote participants ($F_{2,12} = 2.092, p = 0.166, \eta p^2 = 0.258$).

Subjective Mental effort: There were no significant differences in either local ($F_{2,10} = 3.003, p = 0.095, \eta p^2 = 0.375$) or remote participants ($F_{2,12} = 3.618, p = 0.059, \eta p^2 = 0.376$).

System Usability Score (SUS): The local SUS score showed that VC was rated the best in terms of system usability, followed by TI and MH. There was a significant difference ($F_{2,10} = 9.488, p <$

Table 3: Subjective rating descriptions in our post experiment questionnaire.

#	Variable	Statements	Source
1	Task difficulty	Overall, this task was	Sauro and Dumas (2009) [62]
2	Enjoyment	I enjoyed the experience	Homegrown
3	Focus	I was able to focus on the task activities	
4	Mental effort	Please rate your mental effort in this task according to the scale provided	Zijlstra and Van Doorn (1985) [78]
5	System Usability Scale	I think that I would like to use this system frequently.	Brooke (2013) [14]
6		I found the system unnecessarily complex.	
7		I thought the system was easy to use.	
8		I think that I would need the support of a technical person to be able to use this system.	
9		I found the various functions in this system were well integrated.	
10		I thought there was too much inconsistency in this system.	
11		I would imagine that most people would learn to use this system very quickly.	
12		I found the system very cumbersome to use.	
13		I felt very confident using the system.	
14		I needed to learn a lot of things before I could get going with this system.	
15	Co-presence	I noticed my students.	Harms and Biocca (2004) [31]
16		My students' presence was obvious to me.	
17		My students caught my attention.	
18	Attention allocation	I was easily distracted from my students when other things were going on.	
19		I remained focus on my students throughout our interaction.	
20		My students did not receive my full attention.	
21	Perceived message understanding	I understood where my students' focus was on	
22		My students' thoughts were clear to me	
23		It was easy to understand my students	
24		Understanding my students was difficult	
25	Monitoring ability	How do you rate this remote presentation system's ability to support you to monitor your students' attention? (Teacher study only)	Homegrown
26	Setup time requirement	How do you rate acceptability of the time required to set up this remote presentation system? (Teacher study only)	
27	Note taking ability	How do you rate this presentation system's ability to support you in annotating slides or taking notes? (Student study only)	
28	Preference	Please order the remote presenting systems according to your preferences	

0.005, $\eta p^2 = 0.655$). However, post-hoc analysis did not show any pair-wise difference. There were no significant differences for remote participants ($F_{2,12} = 3.084, p = 0.083, \eta p^2 = 0.340$).

Setup Time Requirement: The local participants rated the setup time requirement for VC and TI as more acceptable than MH. There were significant differences ($F_{2,10} = 5.677, p < 0.05, \eta p^2 = 0.532$) in this variable between VC-MH ($p < 0.05$). The remote participants rated the setup time for TI as more acceptable than VC followed by MH. There was a significant difference ($F_{2,12} = 5.213, p < 0.05, \eta p^2 = 0.465$) between TI and MH ($p < 0.05$).

Cognitive Effects

Focus: There were no significant differences in either local ($F_{2,10} = 1.872, p = 0.204, \eta p^2 = 0.272$) or remote participants ($F_{2,12} = 0.377, p = 0.694, \eta p^2 = 0.059$).

Attention Allocation: There were no significant differences in either local ($F_{2,10} = 0.930, p = 0.426, \eta p^2 = 0.156$) or remote participants ($F_{2,12} = 3.613, p = 0.059, \eta p^2 = 0.376$).

Perceived Message Understanding: There were no significant differences for local participants ($F_{2,10} = 1.305, p = 0.314, \eta p^2 = 0.207$).

The remote participants understood the students better in TI followed by MH and VC condition. There was a significant difference ($F_{2,12} = 4.383, p < 0.05, \eta p^2 = 0.422$) between TI and VC ($p < 0.05$). **Monitoring Ability:** There were no significant differences for local participants ($F_{2,10} = 2.031, p = 0.1818, \eta p^2 = 0.289$). The remote participants rated this ability higher for TI followed by MH and VC. There was a significant difference ($F_{2,12} = 5.213, p < 0.05, \eta p^2 = 0.442$) between TI and MH ($p < 0.05$).

Experiential Responses

Co-presence: There were no significant differences for local participants ($F_{2,10} = 0.696, p = 0.5211, \eta p^2 = 0.122$). On the other hand, the remote participants felt the strongest students' presence in TI followed by MH and VC. There were significant differences ($F_{2,12} = 11.339, p = 0.002, \eta p^2 = 0.654$) between TI-MH ($p < 0.05$) and TI-VC ($p < 0.05$).

Enjoyment: There were no significant differences in either local ($F_{2,10} = 1.017, p = 0.396, \eta p^2 = 0.169$) or remote participants ($F_{2,12} = 1.233, p = 0.326, \eta p^2 = 0.170$).

Table 4: Teachers' questionnaire results for each condition (L = local participants, R = remote participants).

#	Variable	Location	Videoconferencing		Tutor In-sight		Mozilla Hubs		P
			M	SD	M	SD	M	SD	
System Usability									
1	Task difficulty (1: very difficult – 5: very easy)	L	4.00	0.76	3.50	0.96	2.83	0.90	0.269
		R	3.29	1.11	3.86	0.99	2.86	0.99	0.166
2	Mental effort (0: not at all hard to do – 150: ‘tremendously hard to do’)	L	18.33	18.07	47.50	34.37	58.33	31.71	0.095
		R	43.57	31.70	31.86	29.72	54.29	30.87	0.059
3	System Usability Scale (out of 100)	L	75.42	12.73	64.17	21.10	45.83	18.63	*0.005
		R	56.79	13.92	65.36	10.21	49.64	9.77	0.083
4	Setup time requirement (1: very low–5: very high)	L	4.00	0.93	4.00	0.58	2.50	0.96	*0.022
		R	3.43	0.70	4.00	0.00	3.00	0.93	*0.024
Cognitive Effects									
5	Focus (1: strongly disagree – 5: strongly agree)	L	4.50	0.73	3.50	1.26	3.33	1.11	0.204
		R	3.14	1.45	3.86	0.83	3.43	1.29	0.694
6	Attention allocation (1: strongly disagree – 7: strongly agree)	L	3.92	0.83	3.88	1.49	2.96	1.11	0.426
		R	3.54	1.05	4.82	0.32	3.79	1.15	0.059
7	Perceived message understanding (1: strongly disagree – 7: strongly agree)	L	3.92	0.83	3.88	1.49	2.96	1.11	0.314
		R	2.81	0.98	5.19	1.30	3.48	1.74	*0.037
8	Monitoring ability (1: very low–5: very high)	L	3.33	1.12	4.17	1.07	2.83	1.21	0.182
		R	2.43	1.00	4.00	0.76	3.00	1.20	*0.030
Experiential Responses									
9	Co-presence (1: strongly disagree – 7: strongly agree)	L	5.33	1.41	4.67	1.80	4.44	1.63	0.521
		R	2.95	0.95	5.95	0.28	4.14	1.65	*0.002
10	Enjoyment (1: strongly disagree – 5: strongly agree)	L	4.00	0.53	3.67	0.75	3.83	1.34	0.396
		R	3.57	1.22	4.29	1.16	3.43	1.29	0.325
11	Preferences (1: best – 3: worst)	L	<div><div>2</div><div>2</div><div>2</div></div>	<div><div>3</div><div>2</div><div>1</div></div>	<div><div>1</div><div>2</div><div>3</div></div>	<div><div>Rank 1</div><div>Rank 2</div><div>Rank 3</div></div>			
		R	<div><div>2</div><div>3</div><div>2</div></div>	<div><div>5</div><div>1</div><div>1</div></div>	<div><div>1</div><div>3</div><div>3</div></div>				

Preferences: Local participants selected the TI condition as the best, giving it three first-place votes, two second-place votes, and just one third-place vote. Local participants placed the VC condition second with two first-place votes, two second-place votes, and two third-place votes. The MH condition was rated last, with just one vote for first-place, two votes for second-place, and three votes for third-place. The majority of remote participants likewise preferred the TI condition with five first-place votes, one second-place vote, and one third-place vote. The VC condition followed the TI condition with two votes for first-place, three votes for second-place, and two votes for third-place. The MH condition was placed last by remote participants, with just one first-place vote, three second-place votes, and three third-place votes. Note that one remote participant ranked the VC and TI as tied for the first-place.

4.1.5 Discussion

System Usability

In terms of usability, we found a difference in SUS and Mental Effort ratings between the local and remote participants for the VC condition, even though the study procedure was the same for both groups. The semi-structured interviews suggested that participants considered their real-life experience of teaching using VC, which resulted in a variety of responses. Nevertheless, we saw a comparable response between the two groups for the TI and MH conditions since all participants had no previous experience conducting lectures in AR and VR.

The questionnaire responses partly support H1 since participants rated the MH condition as the lowest in terms of SUS score (c, local) and acceptable setup time (d, both). Moreover, local participants who were required to present via VR-HMD appeared to rate the usability score lower than remote participants, as five local participants noted that the user interfaces within the VR-HMD required adjustment time and that they felt as though they needed to learn an additional programme. “*I feel like I have to adjust to the new environment after wearing the HMD*” – LP2. “*I cannot use it in my office since I feel like I will crash into things*” – LP3. Despite having a lower usability score than the other conditions, eight participants said that they enjoyed using MH due to the immersion in the VR and believed that, with practice, they would become proficient with the system. “*I feel like I am in the real classroom*” – LP1. “*I like that the class feels very interactive*” – LP5. However, six participants believed that the benefits of MH did not outweigh the effort for daily teaching practice. As one participant stated, “*I might use this (MH) when I have to teach a special class about something that requires 3D models, but when I am giving a regular lecture, I probably prefer other conditions more, as this (MH) requires a lot of time to setup.*” – RP7. Another participant stated, “*I can see it being used in some settings, but I don't think it will have regular use for university students; it could be used to engage younger students.*” – LP4. This might explain why the majority (11) of participants ranked MH as either second or third when it came to preference. Participant feedback indicated that teachers did not rate teaching tools based purely on usability, but also examined trade-offs between efficiency (time for technical

setup) and student learning experience. However, the existing VR presentation system's disadvantages do not yet exceed its benefits in a classroom setting, which is consistent with the findings of Jensen and Konradsen [34] and Radiani et al. [56]

There was no significant difference between the VC and TI conditions in terms of task difficulty, mental effort, SUS score, and setup time, showing that the participants understood the Tutor In-sight's dashboard and its functionality as much as the videoconference system. In addition, during the semi-structured interviews, nine participants gave positive responses to the TI condition, stating that the user interfaces were "simple to use", and eye tracking data was useful for monitoring students. *"I think in real class, if I saw students' icons follow my pointing, even 50 percent of the class, I would get feedback that the students were still paying attention"* – RP1. They also thought that the functionality of colour-coding students' profile pictures and sorting their attention states into groups would be helpful for understanding the class status. *"I feel like it's easy to detect distracted students using this (UI)"* – LP2. Five participants praised the automated warning message, stating that *"It is a useful tool that helps manage students and reduces my burden"* – RP4. In fact, when we explained to the participants about setup requirements and advanced features (3D objects) of the TI condition, none of the participants considered that it would lead to a significant increase in their workload. *"I think I will manage since it is PowerPoint, and the workflow is similar to the current workflow"* – LP4. Participants' positive responses supported the premise of **H1**, as they believed they could easily incorporate Tutor In-sight into their current presentation process without a negative impact on the setup time.

As suggested by the participants' feedback, they preferred the TI condition over the VC and MH conditions due to its ease of use and additional benefits such as being able to easily identify students needing help due to the colour-coding and grouping of students' profile pictures. However, seven participants were concerned about the eye-tracking data, arguing that the students' eye movements were too fast to observe and might cause distractions. *"I preferred to look at presentation on the dashboard, so I can see the students at the same time, so it would be nice if the icons are (semi)transparent"* – LP4. *"The icons are a bit too fast; I think it's better to slow them down or provide a summary instead."* – RP2. Nevertheless, two participants mentioned that if they were acquainted with the materials, they would be able to pay more attention to the students since they would not need to read from the slides. *"If it is a lecture that I done before, I probably focus more on the students like I do in the real classroom"* – RP5. Four participants were concerned about the Tutor In-sight's scalability, with one commenting that *"I am teaching a big module (classroom) of around 100 pupils, and I believe that the display picture of students may be too tiny in that scenario."* – RP6. A common suggestion for improving Tutor In-sight is to consolidate eye tracking data to decrease the dashboard movement and prevent distraction. We address these concerns in the redesign suggestion of the Tutor In-sight's dashboard described in Section 4.1.6.

Monitoring Ability and Student's Presence

In comparison to the other conditions, the TI condition was rated favourably in terms of co-presence, perceived message understanding, and ability to monitor students' attention. These questionnaire

responses provided evidence to **H2**. This finding was further supported by the semi-structured interviews, which revealed that the majority of participants believed that eye-tracking data was a practical solution for monitoring students' attentiveness. In addition, they explained the weaknesses of VC and MH. In a practical remote classroom, ten participants pointed out that students seldom use cameras, making it impossible to monitor their focus. One participant noted, *"We cannot ask the students to open the camera since some students may not have a private room and opening the camera might violate their family's privacy."* – RP1. Furthermore, six participants pointed out that even when the students turn on the camera, the teachers are unsure of what the students are viewing since they may be multitasking while gazing at the screen [17]. *"Even though the students turn on the camera, it is very hard to understand what they are watching. I feel like I speak alone most of the time; even when I ask questions, I am often met with silence."* – RP3. Therefore, the participants ranked the VC condition lower than TI, despite their familiarity with the VC condition.

While we anticipated that the immersion in MH would increase the perceived students' co-presence for the participants, particularly for the local ones who wore a VR-HMD throughout the presentation phase, the perception of three local participants refuted our assumption as they noted that the students' avatars lacked facial expressions and the avatars alone did not add to the sense of being with other humans. *"I felt like I talked to cartoon characters"* – LP2. The participant comments and lack of significant results in the questionnaire responses in terms of co-presence, attention allocation, and perceived message understanding led us to reject **H3**. This finding indicated that immersion alone did not necessarily increase social presence for the teachers.

Teachers also noted that they could only interpret students' gaze based on the head movement of the avatar in the MH condition, which required more mental effort than the Tutor In-sight's direct eye gaze visualisation. *"While I liked the VR immersion, I feel like I need to interpret the avatar head to understand the students' focus, compared to the TI condition where I can directly see where the students are looking."* – LP6. Three remote participants also identified the same flaw in the MH condition as in the VC condition: *"students might just login to the VR and do another job."* – RP4. Due to these factors, the majority of participants favoured the TI condition over the others for monitoring students, with some stating, *"At least we know that the students are paying attention to what we are saying."* – RP1.

When questioned about the privacy concern with eye tracking, all participants who preferred the TI condition believed that eye tracking was a fair compromise between forcing students to turn on the webcam and the current situation that teachers cannot monitor students' attention. Six participants from this group, however, believed that individuals' eye tracking data might be too detailed and cause concern for students. Hence, they proposed that the eye tracking data should be analysed and reported as a summary report to provide an overview and notify teachers only about students who may have problems. *"I feel like the moving graph is a bit distracting, I think it's better to compile and show it after class; during the class, it is probably better to just show me students that need my attention"* – RP3. These redesign suggestions are discussed further in the following section.

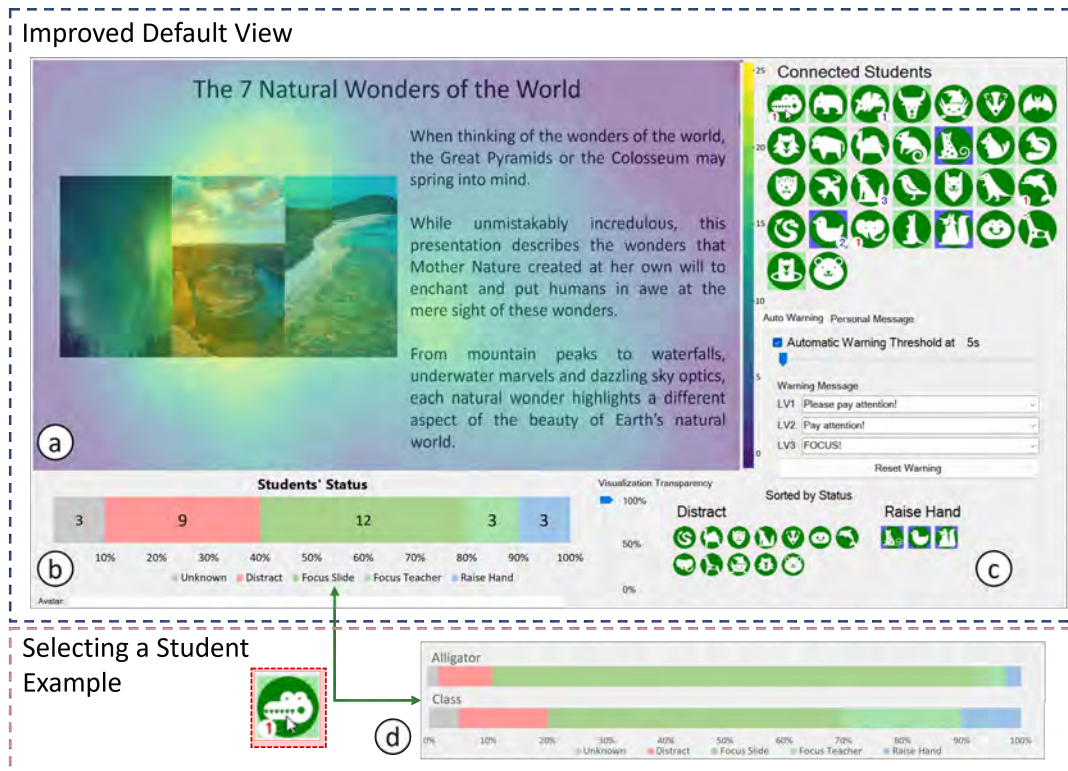


Figure 7: Dashboard redesign based on teachers' suggestions: (a) a heat-map to visualise students' eye gaze data; (b) a stacked bar graph to provide an overview of the class status; (c) a sorting panel that shows students that are distracted or having questions; (d) when selecting a student, two stacked bar graphs are displayed to show the selected student's average status compared to the class average.

4.1.6 Dashboard Redesign

Teacher participants had two concerns with the dashboard of Tutor In-sight. First, the data from students' eye tracking was too fast-paced and too many data points were visualised, which may create distractions during the presentation. Second, the eye tracking data was overly specific, which may lead to students feeling singled out or concerned about their privacy. Furthermore, participants preferred to get an average of data throughout the whole lecture session rather than focussing on minute details. Consequently, we updated the dashboard design. We addressed the first concern by using a heat-map overlay on the slide to summarise students' eye gaze instead of showing every student's eye gaze data (Figure 7a). This heat-map visualisation should help smoothen out the eye tracking data movements and reduce distractions. As all teacher participants used the dashboard slide to track both the slide and the students' status (similar to the presenter view feature in PowerPoint), we added a slider to control the heat-map visualisation transparency, which enables teachers to prioritize their attention on the fly between the slide content and the students' eye-gaze data. Furthermore, the status graph in the bottom left was replaced by the stacked-bar graph (Figure 7b) that shows the current classroom status in percentages, providing an overview of the class and addressing the second concern. We also removed the 'Focus' group from the sorting panel (Figure 7c) to create additional space for the "Distract"

and "Raise Hand" groups, as participants pointed out that they prefer to only see the students that require immediate attention. In addition, when selecting a student, two stacked bar graphs are displayed to show the selected student's average status compared to the class average (Figure 7d). As suggested by the participants, another important feature is the post-class statistics, which should be visualised using a separate post-class dashboard as described in Mazza and Dimitrova [42] or Xhakaj et al. [74]. This dashboard could potentially integrate advanced techniques [6, 11] to analyse eye-gaze sequence to quantify student's attention and detect mind wandering or distraction. This post-class dashboard design and development, however, is beyond the scope of this research and will be left for future development.

4.2 Student's Perspective Study

4.2.1 Setups of three conditions

To evaluate the design of Tutor In-sight from the student's perspective, we performed a within-subjects study comparing the Tutor In-sight condition (TI) to the standard videoconference condition (VC) and Mozilla Hubs condition (MH). Three mock presentations on the topic of "The 7 Natural Wonders of the World"¹ were prepared for the participants to view as learning and teaching materials.

¹<https://www.worldatlas.com/places/the-7-natural-wonders-of-the-world.html>

This topic was selected because it should be easily comprehended by participants and because it exemplifies a presentation with a variety of content types, including text, images, and 3D maps. Each presentation discusses two of the world's seven natural wonders in order to avoid repetition during the within-subjects study, and is structured similarly in terms of presentation time, amount of text, number of images, and maps. A recording was used to simulate the live learning setting to ensure that all participants would receive the same presentation. We created a custom software to capture and replay the presentation in the TI condition. Since Mozilla Hubs does not support the playback of pre-recorded presentations, we also created a Mozilla Hubs-like virtual environment software for participants to view the presentation in VR. This Mozilla Hubs-like software records the presenter's head and hand movements to create a half-body avatar in the VR that is lip-synced with the presenter's voice, as shown in Figure 8b. We also recorded the presentations using screen recording software and a webcam to capture the presenter's face in order to make a video comparable to the videoconference setup in VC (Figure 8a). Presentations in all of the conditions were created by the same presenter and script for consistency. This enables us to do an unbiased comparison of the presenter's social presence and the usability of each presentation tool.

4.2.2 Procedures and participants

Eleven student volunteers (5 female and 6 male, 21 – 30 years of age) were recruited from a local university. The participants came from a variety of disciplinary backgrounds, ranging from undergraduate art students to post-graduate engineering students. Three individuals had never used MR or VR before, seven regarded themselves as beginners, and one was an expert user. All participations were voluntary without involving any reward or compensation.

Prior to the presentation, HoloLens 2's eye calibration apps were used to calibrate the MR headset to the participants' eyes. Participants were asked to focus on the presentations, which lasted around five minutes each. To avoid learning bias, the participants used the presentation tools (VC, TI, and MH) and viewed the contents in a random order. We used a post-experiment questionnaire (Table 3) and a semi-structured interview with similar procedures as in the teacher's perspective study. A session of the student's perspective study took roughly 40 minutes to complete.

In the TI condition, HoloLens 2 was connected through a local WIFI network to a laptop (CPU AMD Ryzen7 3.20GHz, 32GB of RAM, NVIDIA GPU RTX3080 laptop) for playback of the recorded presentation on HoloLens 2. In the MH condition, a Meta Quest 2 was connected to the same laptop to display the recorded presentation in the VR. For the VC condition, a recorded video presentation was displayed on a 27-inch display monitor connected to the aforementioned laptop with a 2560x1440 resolution. In all conditions, participants were given printouts of the slides with a Vuforia tag and were instructed to freely annotate the printed slides as if they were viewing a live lecture.

4.2.3 Hypotheses

Since students were familiar with the VC condition but not the TI or MH conditions, it would be simpler for participants to use VC (**H4**). However, due to Zoom fatigue [60] and restricted body language with webcam stream, participants would not enjoy learning through

VC (**H5**) and would have a diminished sense of the teacher's co-presence. Moreover, we anticipated that the body language of the computer-generated avatar in the TI condition would have the same impact in term of teacher's co-presence as the body language of the body-tracking avatar in the MH condition (**H6**). We also assumed that student's note-taking capability would be at its lowest in the MH condition due to the VR-HMD blocking the participants' view (**H7**).

Based on these assumptions and our design requirements, we hypothesized the following for the student evaluation:

- H4** : There are significant differences among the three conditions in (a) task difficulty rating, (b) mental effort rating, (c) SUS score, favouring the VC over TI and MH conditions.
- H5** : There are significant difference among the three conditions in enjoyment, favouring TI and MH over VC condition.
- H6** : There are significant differences in the rating of a) co-presence and b) attention allocation, favouring TI and MH over VC condition.
- H7** : There are significant differences among the three conditions in terms of note-taking ability, favouring TI and VC over MH condition.

4.2.4 Results

Similar to the teacher's perspective study, Align-and-rank ANOVA ($\alpha = 0.05$) [72] was used for non-parametric analysis, given that the data are not normally distributed, as shown by the results of Shapiro-Wilk test ($p < 0.05$). Bonferroni correction was used for post-hoc pairwise comparisons. Table 5 shows the descriptive statistics of the variables.

System Usability

Task difficulty (SEQ): There were no significant differences ($F_{2,20} = 1.482, p = 0.251, \eta_p^2 = 0.129$).

Mental effort: There were no significant differences ($F_{2,20} = 2.712, p = 0.091, \eta_p^2 = 0.213$).

System Usability Score (SUS): There were no significant differences ($F_{2,20} = 1.097, p = 0.353, \eta_p^2 = 0.099$).

Note Taking Ability: The participants rated the note taking ability for VC and TI as higher than MH. There were significant differences ($F_{2,20} = 5.232, p < 0.05, \eta_p^2 = 0.343$) in this variable between the TI-MH ($p < 0.05$) and VC-MH ($p < 0.05$). This confirms **H7**.

Cognitive Effects

Focus: There were no significant differences ($F_{2,20} = 1.914, p = 0.174, \eta_p^2 = 0.161$).

Attention Allocation: The participants allocated attention to the teacher similarly in MH and TI followed by VC. There were significant differences ($F_{2,20} = 10.535, p < 0.001, \eta_p^2 = 0.513$) in this variable between TI-VC ($p < 0.01$) and MH-VC ($p < 0.01$). This confirms **H6b**.

Perceived Message Understanding: The participants rated all three conditions similarly, and there were no significant differences ($F_{2,20} = 0.084, p = 0.92, \eta_p^2 = 0.008$).

Experiential Responses

Co-presence: The participants felt the teacher's presence similarly in TI and MH, followed by VC condition. There were significant differences ($F_{2,20} = 6.960, p < 0.005, \eta_p^2 = 0.410$) in this variable

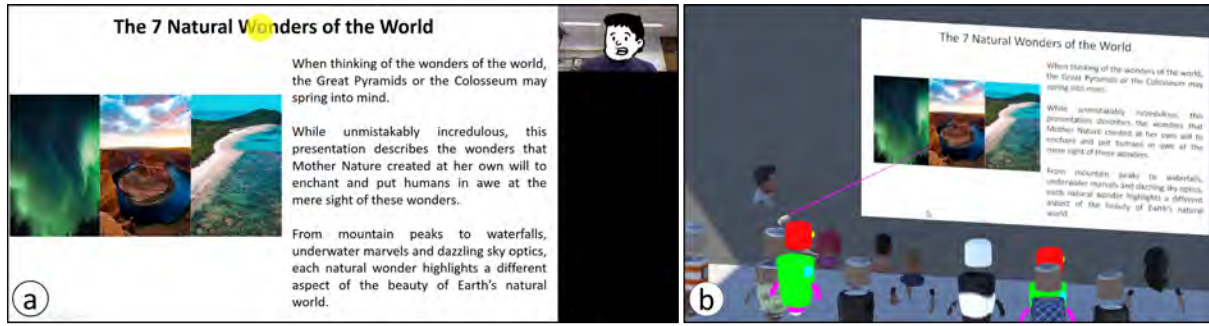


Figure 8: (a) A simulated videoconference presentation for the student's perspective (b) Mozilla Hubs-like virtual environment that supports presenter record and playback features to create a simulated VR presentation for the student's perspective.

Table 5: Student's questionnaire results for each condition.

#	Variables	Videoconferencing		Tutor In-sight		Mozilla Hubs		p
		M	SD	M	SD	M	SD	
System Usability								
1	Task difficulty	4.45	0.99	4.00	0.74	4.00	0.95	0.251
2	Mental effort	15.45	26.41	22.73	19.47	19.55	15.59	0.091
3	System Usability Scale	77.50	10.28	69.09	10.67	68.64	18.10	0.353
4	Note Taking Ability	3.82	0.94	3.64	0.64	2.45	1.08	*0.015
Cognitive Effects								
5	Focus	3.55	1.37	3.91	0.90	4.45	0.50	0.174
6	Attention allocation	3.91	1.16	5.43	0.89	5.45	0.88	*0.001
7	Perceived message understanding	5.67	1.16	5.76	1.07	5.91	1.04	0.919
Experiential Responses								
8	Co-presence	4.36	1.29	5.79	0.90	5.79	1.10	*0.005
9	Enjoyment	3.73	1.14	4.91	0.29	4.64	0.64	*0.006
10	Preferences	<div><div>2</div><div>2</div><div>7</div></div>		<div><div>6</div><div>5</div></div>		<div><div>3</div><div>4</div><div>4</div></div>		<div><div>Rank 1</div><div>Rank 2</div><div>Rank 3</div></div>

between TI-VC ($p < 0.05$) and MH-VC ($p < 0.05$). This confirms **H6a**.

Enjoyment: The participants enjoyed using TI the most, followed by MH and VC. There was a significant difference ($F_{2,20} = 6.609, p = 0.006, \eta_p^2 = 0.398$) between TI and VC ($p < 0.05$) conditions. This partly confirms **H5**.

Preferences: The TI condition was ranked the best by the participants, with six participants placing the TI as the best, and the other five participants placing the TI as the second best. The MH condition placed second in the ranking by receiving three first-place votes, four second-place votes, and four third-place votes. The VC ranked last with two first-place votes, two second-place votes, and seven third-place votes.

4.2.5 Discussion

The student participants rated all conditions similarly in terms of system usability, i.e., SUS score, mental effort, and task difficulty; thus, there is no evidence to support **H4**. However, the majority (7) of participants favoured the TI and MH conditions over VC. This

indicates that the participants preferred the sense of co-presence, enjoyment, and attention allocation provided by TI and MH. The results of semi-structured interviews also confirmed this finding, since seven participants noted the lack of teacher interaction in the VC condition. During the videoconference presentation, they said that they only read the presentation materials and rarely looked at the teacher. "I don't actually feel the need to follow the teacher's presentation" – P10. Four participants also admitted to being easily distracted and finding the videoconference session uninteresting. "I feel like I lost focus when I take my eye of screen to take some note" – P4. These participant remarks and significant lower ratings of enjoyment, co-presence, and attention allocation in the VC condition provided evidence for Zoom fatigue (**H5**) and impeded teachers' body language on the webcam broadcast (**H6**). Two participants, however, preferred the VC condition because they were more comfortable with the technology and would rather observe the teacher's facial expressions.

The TI and MH conditions were assessed similarly in the post-experiment questionnaire, with the exception of the note-taking

ability, which was hampered by the VR-HMD in the MH condition, according to nine participants. *"I cannot take note in VR at all, I can only listen and listen"* – P1. This observation was corroborated by a significantly lower rating for note taking ability in the MH condition, resulting in the acceptance of **H7**. Nevertheless, three other individuals who preferred the MH setting said that they liked to concentrate on the lecture first and then take notes after class. On the other hand, six participants appreciated the integrated virtual and physical workspace in the TI condition, which allowed them to concentrate on a single workspace and improved their note-taking capabilities. *"I like this (TI) condition since it is convenient for me to listen and take note simultaneously, because I can focus on single location"* – P4. *"In the AR, I can completely focus on that (slide) and pointing animation show me the point to focus, in addition, I am in the real-world that I can still take-note"* – P7. They claimed that the note-taking skills would be vital to their learning process, implying that the note-taking ability has a strong influence on participants' preferences.

The semi-structured interview also revealed differing opinions on VR in the MH condition, with four participants commenting on VR positively for its immersiveness, which contributes to their focus. *"I feel like I am in another world that help me focus"* – P5. However, four other participants argued that the virtual environment caused them to be distracted and anxious because they had to adjust to the virtual world and their view of the real-world was blocked off. *"It's easy to distract in the VR, I found myself looking at other students and (virtual) environments, there is so much to look at"* – P10. *"I felt uncomfortable and anxious because I don't know what happened in the real-world"* – P6. The opposing viewpoints reinforced Radianti et al. [56]'s observation that VR experiences varied from student to student, potentially influencing learning outcomes.

Similar ratings in the TI and MH conditions also indicated that the avatar with generated body language in the TI condition had similar impacts in terms of co-presence, enjoyment, and attention allocation as the recorded body-tracking avatar in the MH condition, providing support to **H6**. The semi-structured interviews backed up our findings, as seven participants characterised the avatar in the TI condition as interactive with its body language (gaze, gesture, and movement) that kept their attention throughout the presentation. *"I feel like the instructor is there with me, and it feel more or less the same between AR and VR"* – P7. Three participants also stated that they thought that the teacher was paying attention to them, which could be attributed to the avatar's mutual gaze. *"I like that the teacher is walking around and look at me, felt like the teacher is looking and I have to focus"* – P9.

In addition to the avatar, several participants positively commented on other TI condition aspects. For example, three participants commended the notification feature that reminded them to pay attention to the topic being presented (Figure 5e, 5h). *"I feel like the instructor interact with me when I saw the pop-up message"* – P10. Two additional participants liked the "blackboard" features (Figure 5i) since it reminded them to change the slide and kept them informed while doing so. *"The pop-up helped remind me to keep up with the slide and also provided information about the correct slide"* – P11.

Despite the lack of statistically significant differences between the TI and MH conditions in the quantitative data, semi-structured

interviews suggested that students preferred TI over MH because of its ability to support note-taking, keep participants aware of their surroundings, provide automatic notifications to keep participants engaged, and provide additional avatar personalization with mutual gaze.

5 Differences in Teacher and student requirements

The findings of both user studies (Table 6 **H1**, **H5**, **H6**, **H7**) supported initial observations in previous literature [32, 34, 56, 57], suggesting that the existing systems are inadequate at meeting the needs of both teachers and students. According to the results of the teacher study, teacher participants prioritised setup time, usability, and low learning curve, causing the majority of them to disregard VR presentation tools in their routine usage (**H1**), despite the fact that they liked the immersion that VR provided. This finding is consistent with previous research [34, 56] and establishes that there are barriers to employing VR technology in remote presentation settings and that usable alternatives are required to enhance acceptance.

Another intriguing finding is that, contrary to previous research [49, 53], the avatar of students in VR does not increase co-presence for teachers (**H3**), as some teacher participants commented on the student avatar's lack of facial expression and cartoon-like style. Teacher participants, on the other hand, rated the TI condition, which only had students' eye gaze, as comparable in terms of co-presence. This implies the functioning of the students' representation is the priority for the teacher participants, and a more realistic avatar with eye-gaze and facial expressions might be needed for VR to be used in the education setting.

Teacher participants also highlighted the difficulties of supervising students using the videoconferencing technology since typically in real-life teaching sessions students seldom turn on the web camera owing to privacy concerns. As a result, teacher participants preferred Tutor In-sight, which allows them to monitor the students and receive real-time feedback (**H2**). Another requirement that teacher participants desired is tools that reduce their workload, such as automated warning messages that help manage students and summary reports that convey important classroom information after the session.

The student participants' feedback also supported the "Zoom Fatigue" claims by Riva et al. [60] (i.e., students begin to experience burnout from excessive use of videoconferencing software, **H5**) and decreased engagement as a result of restricted body language owing to the limited web camera stream (**H6**), as evidenced by the low rating of videoconferencing systems in terms of preferences, enjoyment, co-presence, and attention allocation. The majority of student participants preferred the Tutor In-sight and Mozilla Hubs presentations, which facilitate interactive learning, enhance teacher co-presence, and in general provide a more engaging presentation. The student participants in our study attributed the increased interactivity, co-presence, and engagement to the avatars' body language, indicating that teacher body language is vital for students' engagement and should be supported in remote presentation systems. In contrast to the comments of the teacher participants, the presence of the avatar resulted in increased co-presence for

Table 6: Summary of hypotheses (H) based on results

H	Variables	Results	Summary
Teacher Study			
H1	(A) Task Difficulty	Rejected	While teacher participants felt they could be proficient VR presenters, many believe the setup time is not worthwhile for frequent use.
	(B) Mental Effort	Rejected	
	(C) SUS	Accepted (Local)	
		Rejected (Remote)	
(D) Setup Time	Accepted		
H2	Monitoring Ability	Accepted	Participants found Tutor In-sight’s dashboard with eye-tracking to be an effective alternative to videoconferencing and virtual reality for monitoring students.
H3	(A) Co-Presence	Rejected	Teachers’ social presence was not always enhanced by students’ avatars and immersion.
	(B)Attention Allocation	Rejected	
	(C) Perceived Message Understanding	Rejected	
Student Study			
H4	(A) Task Difficulty	Rejected	Participants had minimal trouble adapting to VR and Tutor In-sight.
	(B) Mental Effort	Rejected	
	(C) SUS	Rejected	
H5	Enjoyment	Accepted	Students did not enjoy the videoconference condition, suggesting Zoom fatigue.
H6	(A) Co-Presence	Accepted	Body language is limited during video conferences, resulting in reduced co-presence and attention allocation. In addition, auto-generated avatars (Tutor-In-sight) and body-tracking avatars (VR) provide comparable co-presence and attention allocation.
	(B) Attention Allocation	Accepted	
H7	Note taking ability	Accepted	The ability to take notes was hampered by the VR-HMD.

the students, which is consistent with previous studies [49, 53]. It showed that the student participants prioritized the interactivity of the teacher's avatar. Furthermore, students made no complaints about the cartoon-like appearance of the avatars, despite the fact that they were produced from the same avatar generation platform (ReadyPlayerMe [59]), implying that the students' standards for the teacher's avatar are more relaxed than the teachers' standards for the students' avatar.

At the same time, the presentation system should support and enhance note-taking abilities (H7) since many students utilise note-taking during a lecture as a form of active learning [21]. The student participants also noted that features that helped guide their attention, such as notifications or visual cues that highlight the current teaching material, were useful to keep their focus.

The student participants' varied responses to the VR system also demonstrate various effects of immersion [56] on participants' concentration and attention allocation, as some students feel more concentrated, while others feel anxious or their thoughts wander in the environment. Because these variations may have an influence on the learning result, the virtual environment should be carefully designed or enable students to personalize the environment to

boost their concentration. Student participants' response to the MR system (Tutor In-sight) on the other hand, was more consistent since the participants were still seeing the physical environment.

We summarized the top four design requirements for developing an effective remote presentation system for both teachers and students based on the above discussion in Table 7.

According to the results of the two studies with teachers and students, Tutor In-sight meets the requirements of both teachers and students to a considerable extent; therefore, the participants of both studies preferred Tutor In-sight, indicating that Tutor In-sight has the potential to serve as an alternative to MR presentation for both teachers and students. Future development of remote MR presentation should also take the requirements in Table 7 into consideration to create effective presentation tools for both teachers and students.

6 Limitation and Future Work

Like most other empirical studies, our work has limitations. Here we discuss them to infer implications for further research. First, even though Tutor In-sight has proved in our user tests to improve remote presentation for both teachers and students, the design

Table 7: Teachers' and students' future requirements for MR remote presentation system.

Teachers' Requirements	Students' Requirements
TR1 - Enabling teachers to weigh trade-offs between efficiency (time for technical setup) and student learning experience (extent of immersion)	SR1 - Interactive teacher avatar that supports co-presence via body language (gaze, gesture, and movement)
TR2 - Ability to monitor students' engagement through facial expressions and other attributes	SR2 – Enhancing note taking ability without distraction
TR3 – Automating tools for student attention management to reduce workload	SR3 – Notifications and visual cues to guide attention to the current teaching material
TR4 – Providing post-class statistics to supplement real-time feedback of students' learning experience	SR4 – Integrated virtual and physical workspace to improve concentration in a single workspace

of Tutor In-sight focuses on enhancing lecture-style instruction and supporting teacher-students interactions (one-to-many), which is the prevalent pedagogical approach in remote education. We propose that further research and additional features are necessary for Tutor In-sight to enable students-students interactions (many-to-many). Furthermore, features that enable teachers to check students' workspaces should be implemented in the future to improve two-way communication.

Second, Tutor In-sight focuses only on eye tracking data to visualise students' eye-gaze; nevertheless, there are ongoing research areas such as emotion recognition [61, 77] that might be valuable for enhancing remote classroom awareness for teachers. However, feedback from teachers on eye tracking already indicates that excessive detail might raise privacy concerns and may not be essential for live monitoring of students' attentiveness. Therefore, further research is needed to establish if the inclusion of other data, such as students' emotions, may be advantageous or essential for monitoring students' attention during a remote presentation.

Third, Tutor In-sight was developed and tested with PowerPoint, as it is a prevalent presentation tool that teachers are familiar with. In addition, we based the system design around students' printouts; thus, our system is limited to static slides. Further development and testing are needed to support animated slides and other existing presentation tools.

Fourth, to control any extraneous variations in live presentations, it was necessary to use a pre-recorded presentation in the student study. A human presenter might unintentionally vary their way of presenting the same material with the same tool, and this would confound the effects of the presentation tools used in our study. Nonetheless, the only noticeable impact of this arrangement for the students was that the teacher could not respond to their questions. The Tutor In-sight avatar performed its actions based on the algorithm described above with the pre-recorded mouse movement and pointing positions as input, but its behaviour would have been identical if the pre-recorded input had been a live input instead. The avatar also responded to the participants' gaze direction and activities as outlined above, as the avatar's behaviour was not pre-recorded, but determined on the fly. In a live presentation scenario, we expect that Tutor In-sight would not only support the presentation itself but also the question-and-answer section, owing to the integrated workspace and enhanced grounding process as well as the teacher dashboard with its Q & A panel. However, additional research is needed to verify our expectations and Tutor In-sight's impact on learning outcomes in a real-world setting.

Fifth, while teacher participants considered eye-tracking as a better alternative to web cameras and none of the student participants had privacy concerns about eye-tracking data, eye-gaze tracking may be considered obtrusive and raising privacy concerns. However, opting out of all monitoring tools would impede a teacher's ability to detect and assist students needing help and prohibit the use of personalized features, such as mutual gaze, that enhance student engagement. Consequently, an alternative monitoring tool that strikes a balance between monitoring critical data and protecting privacy is necessary. The "black box" characteristic of machine learning monitoring tools (as shown by Asish et al. [6]) has the potential to meet this requirement, but further investigation is necessary for its practical usage.

7 Conclusion

In this paper, we proposed Tutor In-sight, a two-part set of MR presentation tools for real-time remote presentation. For students, an MR avatar is augmented in-situ with the presentation material in their workspace. Tutor In-sight was created to take advantage of integrated virtual and physical workspace and miniature avatars capable of coordinating gaze, performing gestures, and repositioning themselves to be inside the field of vision of the students. For teachers, Tutor In-sight's dashboard visualises students' eye gaze data, allowing the teachers to monitor students. In addition, the Tutor In-sight teacher tool is compatible with existing presentation tools such as Microsoft PowerPoint, enabling the teacher to produce an MR presentation with minimal additional workload using their familiar presentation software. To gain a better understanding of the teachers' and students' usage of remote presentation tools, Tutor In-sight was compared to a videoconference and VR remote presentation tool in two user studies, one for teachers and one for students, using a mocked presentation of "The seven natural wonders of the world". The teachers' study revealed that Tutor In-sight's dashboard was useful for the teachers, allowing them to monitor students while existing remote presentation tools cannot. In the students' study, the Tutor In-sight MR avatar was shown to provide a higher degree of co-presence, improve attention allocation, and increase viewer engagement when compared to the videoconferencing system while maintaining note-taking ability, which was diminished in the VR system. Finally, teachers' and students' feedback from both user studies is summarised into design requirements for future remote presentation systems and to further improve the Tutor In-sight system.

Acknowledgments

The publication has been supported by European Union's Horizon 2020 research and innovation program under grant agreement No 856533, project ARETE.

References

- [1] Muhammad Adnan and Kainat Anwar. 2020. Online Learning amid the COVID-19 Pandemic: Students' Perspectives. *Online Submission* 2, 1 (2020), 45–51.
- [2] Hosam Al-Samarraie. 2019. A scoping review of videoconferencing systems in higher education: Learning paradigms, opportunities, and challenges. *International Review of Research in Open and Distributed Learning* 20, 3 (2019).
- [3] Judith Amores, Xavier Benavides, and Pattie Maes. 2015. ShowMe: A Remote Collaboration System that Supports Immersive Gestural Communication. *Extended Abstracts of the ACM CHI'15 Conf. on Human Factors in Computing Systems* (2015), 1343–1348. <https://doi.org/10.1145/2702613.2732927>
- [4] Kartik Anand, Siddhaling Urolagin, and Ram Krishn Mishra. 2021. How does hand gestures in videos impact social media engagement - Insights based on deep learning. *International Journal of Information Management Data Insights* 1, 2 (2021), 100036. <https://doi.org/10.1016/j.jjime.2021.100036>
- [5] Sean Andrist, Michael Gleicher, and Bilge Mutlu. 2017. Looking Coordinated: Bidirectional Gaze Mechanisms for Collaborative Interaction with Virtual Characters. In *Proceedings of the 2017 CHI Conf. on Human Factors in Computing Systems* (Denver, Colorado, USA) (CHI '17). ACM, NY, USA, 2571–2582. <https://doi.org/10.1145/3025453.3026033>
- [6] Sarker Monojit Asish, Arun K Kulshreshth, and Christoph W Borst. 2022. Detecting distracted students in educational VR environments using machine learning on eye gaze data. *Computers & Graphics* 109 (2022), 75–87.
- [7] Hani Atwa, Mohamed Hany Shehata, Ahmed Al-Ansari, Archana Kumar, Ahmed Jaradat, Jamil Ahmed, and Abdelhalim Deifalla. 2022. Online, Face-to-Face, or Blended Learning? Faculty and Medical Students' Perceptions During the COVID-19 Pandemic: A Mixed-Method Study. *Frontiers in Medicine* 9 (2022). <https://doi.org/10.3389/fmed.2022.791352>
- [8] Huidong Bai, Prasanth Sasikumar, Jing Yang, and Mark Billinghurst. 2020. A user study on mixed reality remote collaboration with eye gaze and hand gesture sharing. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
- [9] Jeremy N Bailenson, Nick Yee, Dan Merget, and Ralph Schroeder. 2006. The effect of behavioral realism and form realism of real-time avatar faces on verbal disclosure, nonverbal disclosure, emotion recognition, and copresence in dyadic interaction. *Presence: Teleoperators and Virtual Environments* 15, 4 (2006), 359–372.
- [10] Frank Biocca, Chad Harms, and Judee K Burgoon. 2003. Toward a more robust theory and measure of social presence: Review and suggested criteria. *Presence: Teleoperators & virtual environments* 12, 5 (2003), 456–480.
- [11] Robert Bixler and Sidney D'Mello. 2016. Automatic gaze-based user-independent detection of mind wandering during computerized reading. *User Modeling and User-Adapted Interaction* 26, 1 (2016), 33–68.
- [12] Leanne S Bohannon, Andrew M Herbert, Jeff B Pelz, and Esa M Rantanen. 2013. Eye contact and video-mediated communication: A review. *Displays* 34, 2 (2013), 177–185.
- [13] Mark Bowden. 2015. *Winning body language: Control the conversation, command attention, and convey the right message without saying a word*.
- [14] John Brooke. 2013. SUS: a retrospective. *Journal of usability studies* 8, 2 (2013), 29–40.
- [15] David M Broussard, Yitoshee Rahman, Arun K Kulshreshth, and Christoph W Borst. 2021. An interface for enhanced teacher awareness of student actions and attention in a vr classroom. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 284–290.
- [16] Jerome Bruner. 1999. Folk pedagogies. *Learners and pedagogy* 1 (1999), 4–20.
- [17] Hancheng Cao, Chia-Jung Lee, Shamsi Iqbal, Mary Czerwinski, Priscilla NY Wong, Sean Rintel, Brent Hecht, Jaime Teevan, and Longqi Yang. 2021. Large scale analysis of multitasking behavior during remote meetings. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [18] Yuanzhi Cao, Xun Qian, Tianyi Wang, Rachel Lee, Ke Huo, and Karthik Ramani. 2020. An Exploratory Study of Augmented Reality Presence for Tutoring Machine Tasks. In *Proceedings of the 2020 CHI Conf. on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). ACM, NY, USA, 1–13. <https://doi.org/10.1145/3313831.3376688>
- [19] Jacky C. P. Chan, Howard Leung, Jeff K. T. Tang, and Taku Komura. 2011. A Virtual Reality Dance Training System Using Motion Capture Technology. *IEEE Trans. Learn. Technol.* 2 (April 2011), 187–195. <https://doi.org/10.1109/TLT.2010.27>
- [20] Alan Cheng, Lei Yang, and Erik Andersen. 2017. Teaching Language and Culture with a Virtual Reality Game. In *Proceedings of the 2017 CHI Conf. on Human Factors in Computing Systems* (Denver, Colorado, USA) (CHI '17). ACM, NY, USA, 541–549. <https://doi.org/10.1145/3025453.3025857>
- [21] Michelene TH Chi and Ruth Wylie. 2014. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist* 49, 4 (2014), 219–243.
- [22] Carolyn Chisadza, Matthew Clance, Thulani Mthembu, Nicky Nicholls, and Eleni Yitbarek. 2021. Online and face-to-face learning: Evidence from students' performance during the Covid-19 pandemic. *African Development Review* 33 (2021), S114–S125.
- [23] Luca Chittaro and Fabio Buttussi. 2015. Assessing Knowledge Retention of an Immersive Serious Game vs. a Traditional Education Method in Aviation Safety. *Visualization and Computer Graphics, IEEE Transactions on* (04 2015), 529–538. <https://doi.org/10.1109/TVCG.2015.2391853>
- [24] Vanessa Echeverría, Allan Avendaño, Katherine Chiluiza, Aníbal Vásquez, and Xavier Ochoa. 2014. Presentation Skills Estimation Based on Video and Kinect Data Analysis. In *Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge* (Istanbul, Turkey) (MLA '14). Association for Computing Machinery, New York, NY, USA, 53–60. <https://doi.org/10.1145/2666633.2666641>
- [25] Gov.UK Explore education statistics. 2022. *Class sizes - state-funded primary and secondary schools' from 'Schools, pupils and their characteristics'*. <https://explore-education-statistics.service.gov.uk/data-tables/permalink/38627779-bbb1-4517-82fc-31d5a1648b19>
- [26] Zhenan Feng, Vicente A González, Robert Amor, Ruggiero Lovreglio, and Guillermo Cabrera-Guerrero. 2018. Immersive virtual reality serious games for evacuation training and research: A systematic literature review. *Computers & Education* (2018), 252–266.
- [27] Chi-chung Foo, Billy Cheung, and Kent-man Chu. 2021. A comparative study regarding distance learning and the conventional face-to-face approach conducted problem-based learning tutorial during the COVID-19 pandemic. *BMC medical education* 21, 1 (2021), 1–6.
- [28] Susan R Fussell and Leslie D Setlock. 2014. Computer-mediated communication. (2014).
- [29] Miharuru Fuyuno, Takeshi Saitoh, Yuko Yamashita, and Daisuke Yokomori. 2020. Gaze-Point Analysis of EFL Learners while Watching English Presentations: Toward Effective Teaching. *International Journal* 14, 1 (2020), 17–28.
- [30] Aaron M Genest, Carl Gutwin, Anthony Tang, Michael Kalyn, and Jenya Ivkovic. 2013. KinectArms: a toolkit for capturing and displaying arm embeddings in distributed tabletop groupware. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 157–166.
- [31] Professor Chad Harms and Professor Frank Biocca. 2004. Internal Consistency and Reliability of the Networked Minds Measure of Social Presence. <http://cogprints.org/7026/>
- [32] Matthias Heintz, Effie Lai-Chong Law, and Pamela Andrade. 2021. Augmented Reality as Educational Tool: Perceptions, Challenges, and Requirements from Teachers. In *Technology-Enhanced Learning for a Free, Safe, and Sustainable World*. Tinne De Laet, Roland Klemke, Carlos Alario-Hoyos, Isabel Hilliger, and Alejandro Ortega-Arranz (Eds.). Springer International Publishing, Cham, 315–319.
- [33] Kenneth Holstein, Bruce M McLaren, and Vincent Alevan. 2018. Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms. In *International conference on artificial intelligence in education*. Springer, 154–168.
- [34] Lasse Jensen and Flemming Konradsen. 2018. A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies* 4 (2018), 1515–1529.
- [35] Allison Jing, Kieran May, Gun Lee, and Mark Billinghurst. 2021. Eye See What You See: Exploring How Bi-Directional Augmented Reality Gaze Visualisation Influences Co-Located Symmetric Collaboration. *Frontiers in Virtual Reality* (2021), 79. <https://doi.org/10.3389/frvir.2021.697367>
- [36] Brennan Jones, Yaying Zhang, Priscilla NY Wong, and Sean Rintel. 2021. Belonging there: VROOM-ing into the uncanny valley of XR telepresence. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–31.
- [37] John F Kihlstrom and Lillian Park. 2016. *Cognitive psychology: overview*.
- [38] Taeheon Kim, Ashwin Kachhara, and Blair MacIntyre. 2016. Redirected head gaze to support ar meetings distributed over heterogeneous environments. In *2016 IEEE Virtual Reality (VR)*. IEEE, 207–208.
- [39] Konstantinos Koumaditis, Francesco Chinello, Panagiotis Mitkidis, and Simon Karg. 2020. Effectiveness of Virtual Versus Physical Training: The Case of Assembly Tasks, Trainer's Verbal Assistance, and Task Complexity. *IEEE Computer Graphics and Applications* 40, 5 (2020), 41–56.
- [40] Jie Li, Guo Chen, Huib de Ridder, and Pablo Cesar. 2020. Designing a Social VR Clinic for Medical Consultations. In *Extended Abstracts of the 2020 CHI Conf. on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI EA '20). ACM, NY, USA, 1–9. <https://doi.org/10.1145/3334480.3382836>
- [41] Manolis Mavrikis, Sergio Gutierrez-Santos, and Alex Poulouvasilis. 2016. Design and evaluation of teacher assistance tools for exploratory learning environments. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. 168–172.
- [42] Riccardo Mazza and Vania Dimitrova. 2007. CourseVis: A graphical student monitoring tool for supporting instructors in web-based distance courses. *International Journal of Human-Computer Studies* 65, 2 (2007), 125–139.

- [43] Fariba Mostajeran, Frank Steinicke, Oscar Javier Ariza Nunez, Dimitrios Gatsios, and Dimitrios Fotiadis. 2020. Augmented Reality for Older Adults: Exploring Acceptability of Virtual Coaches for Home-Based Balance Training in an Aging Population. In *Proceedings of the 2020 CHI Conf. on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). ACM, NY, USA, 1–12. <https://doi.org/10.1145/3313831.3376565>
- [44] Cornelia Müller, Alan Cienki, Ellen Fricke, Silva Ladewig, David McNeill, and Sedinha Teßendorf. 2013. Body-language-communication. *An international handbook on multimodality in human interaction* 1, 1 (2013), 131–232.
- [45] Michael Nebeling, Shwetha Rajaram, Leiwei Wu, Yifei Cheng, and Jaylin Herskovitz. 2021. XRStudio: A Virtual Production and Live Streaming System for Immersive Instructional Experiences. In *Proceedings of the 2021 CHI Conf. on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). ACM, NY, USA, 1–12. <https://doi.org/10.1145/3411764.3445323>
- [46] Dinna Nina Mohd Nizam and Effie Lai-Chong Law. 2021. Derivation of young children's interaction strategies with digital educational games from gaze sequences analysis. *International Journal of Human-Computer Studies* 146 (2021), 102558.
- [47] Catherine S Oh, Jeremy N Bailenson, and Gregory F Welch. 2018. A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI* (2018), 114.
- [48] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Ming-song Dou, et al. 2016. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th annual symposium on user interface software and technology*. 741–754.
- [49] Niklas Osmers, Michael Prilla, Oliver Blunk, Gordon George Brown, Marc Janßen, and Nicolas Kahrl. 2021. The role of social presence for cooperation in augmented reality on head mounted devices: A literature review. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [50] Michael B. O'Connor, Simon J. Bennie, Helen M. Deeks, Alexander Jamieson-Binnie, Alex J. Jones, Robin J. Shannon, Rebecca Walters, Thomas J. Mitchell, Adrian J. Mulholland, and David R. Glowacki. 2019. Interactive molecular dynamics in virtual reality from quantum chemistry to drug binding: An open-source multi-person framework. *The Journal of Chemical Physics* 22 (2019), 220901. <https://doi.org/10.1063/1.5092590>
- [51] Tomislav Pejisa, Julian Kantor, Hrvoje Benko, Eyal Ofek, and Andrew Wilson. 2016. Room2room: Enabling life-size telepresence in a projected augmented reality environment. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 1716–1725.
- [52] Thammathip Piumsomboon, Arindam Dey, Barrett Ens, Gun Lee, and Mark Billinghurst. 2019. The effects of sharing awareness cues in collaborative mixed reality. *Frontiers in Robotics and AI* (2019), 5.
- [53] Thammathip Piumsomboon, Gun A. Lee, Jonathon D. Hart, Barrett Ens, Robert W. Lindeman, Bruce H. Thomas, and Mark Billinghurst. 2018. Mini-Me: An Adaptive Avatar for Mixed Reality Remote Collaboration. In *Proceedings of the 2018 CHI Conf. on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI '18). ACM, NY, USA, Article 46, 13 pages. <https://doi.org/10.1145/3173574.3173620>
- [54] Thammathip Piumsomboon, Gun A Lee, Andrew Irlitti, Barrett Ens, Bruce H Thomas, and Mark Billinghurst. 2019. On the shoulder of the giant: A multi-scale mixed reality collaboration with 360 video sharing and tangible interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–17.
- [55] Thammathip Piumsomboon, Youngho Lee, Gun A Lee, Arindam Dey, and Mark Billinghurst. 2017. Empathic mixed reality: Sharing what you feel and interacting with what you see. In *2017 International Symposium on Ubiquitous Virtual Reality (ISUVR)*. IEEE, 38–41.
- [56] Jaziar Radianti, Tim A. Majchrzak, Jennifer Fromm, and Isabell Wohlgenannt. 2020. A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education* (2020), 103778. <https://doi.org/10.1016/j.compedu.2019.103778>
- [57] Iulian Radu. 2014. Augmented reality in education: a meta-review and cross-media analysis. *Personal and Ubiquitous Computing* 6 (2014), 1533–1543.
- [58] Vikram Ramanarayanan, Chee Wee Leong, Lei Chen, Gary Feng, and David Suendermann-Oeft. 2015. Evaluating Speech, Face, Emotion and Body Movement Time-Series Features for Automated Multimodal Presentation Scoring. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (Seattle, Washington, USA) (ICMI '15). Association for Computing Machinery, New York, NY, USA, 23–30. <https://doi.org/10.1145/2818346.2820765>
- [59] ReadyPlayerMe. 2022. *Cross-game Avatar Platform for Metaverse*. <https://readyplayer.me/>
- [60] Giuseppe Riva, Brenda K Wiederhold, and Fabrizia Mantovani. 2021. Surviving COVID-19: the neuroscience of smart working and distance learning. *Cyberpsychology, Behavior, and Social Networking* 24, 2 (2021), 79–85.
- [61] Samiha Samrose, Daniel McDuff, Robert Sim, Jina Suh, Kael Rowan, Javier Hernandez, Sean Rintel, Kevin Moynihan, and Mary Czerwinski. 2021. Meetingcoach: An intelligent dashboard for supporting effective & inclusive meetings. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [62] Jeff Sauro and Joseph S Dumas. 2009. Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the SIGCHI Conf. on human factors in computing systems*. ACM, 1599–1608.
- [63] John Short, Ederyn Williams, and Bruce Christie. 1976. *The social psychology of telecommunications*. Toronto; London; New York: Wiley.
- [64] Harrison Jesse Smith and Michael Neff. 2018. Communication behavior in embodied virtual reality. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.
- [65] Rajinder S. Sodhi, Brett R. Jones, David Forsyth, Brian P. Bailey, and Giuliano Maciocci. 2013. BeThere: 3D Mobile Collaboration with Spatial Input. In *Proceedings of the SIGCHI Conf. on Human Factors in Computing Systems* (Paris, France) (CHI '13). ACM, NY, USA, 179–188. <https://doi.org/10.1145/2470654.2470679>
- [66] National Teacher and Principal Survey. 2018. *Average class size in public schools, by class type and state: 2017–18*. https://nces.ed.gov/surveys/ntps/tables/ntps1718_ftable06_t1s.asp
- [67] Santawat Thanyadit, Parinya Pungpongson, Thammathip Piumsomboon, and Ting-Chuen Pong. 2022. XR-LIVE: Enhancing Asynchronous Shared-Space Demonstrations with Spatial-temporal Assistive Toolsets for Effective Learning in Immersive Virtual Laboratories. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–23.
- [68] Santawat Thanyadit, Parinya Pungpongson, and Ting-Chuen Pong. 2019. ObserverVAR: Visualization system for observing virtual reality users using augmented reality. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 258–268.
- [69] Ana M Villanueva, Ziyi Liu, Zhengzhe Zhu, Xin Du, Joey Huang, Kylie A Peppler, and Karthik Ramani. 2021. Robotar: An augmented reality compatible teleconsulting robotics toolkit for augmented makerspace experiences. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [70] Vuforia. 2022. *Vuforia Developer Library: VuMarks*. <https://library.vuforia.com/objects/vumarks>
- [71] Frederik Winther, Linoj Ravindran, Kasper Paabel Svendsen, and Tiare Feuchtnier. 2020. Design and Evaluation of a VR Training Simulation for Pump Maintenance. In *Extended Abstracts of the 2020 CHI Conf. on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI EA '20). ACM, NY, USA, 1–8. <https://doi.org/10.1145/3334480.3375213>
- [72] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conf. on Human Factors in Computing Systems* (Vancouver, BC, Canada) (CHI '11). ACM, NY, USA, 143–146. <https://doi.org/10.1145/1978942.1978963>
- [73] Jason Wolfgang Woodworth, David Michael Broussard, and Cristoph W. Borst. 2022. Redirecting Desktop Interface Input to Animate Cross Reality Avatars. In *2022 IEEE Virtual Reality (VR)*. IEEE.
- [74] Franceska Xhakaj, Vincent Alevan, and Bruce M McLaren. 2017. Effects of a teacher dashboard for an intelligent tutoring system on teacher knowledge, lesson planning, lessons and student learning. In *European conference on technology enhanced learning*. Springer, 315–329.
- [75] Kexin Yang, Xiaofei Zhou, and Iulian Radu. 2020. XR-Ed Framework: Designing Instruction-driven and Learner-centered Extended Reality Systems for Education. *arXiv preprint arXiv:2010.13779* (2020).
- [76] Boram Yoon, Hyung-il Kim, Gun A Lee, Mark Billinghurst, and Woontack Woo. 2019. The effect of avatar appearance on social presence in an augmented reality remote collaboration. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 547–556.
- [77] Haipeng Zeng, Xinhuan Shu, Yanbang Wang, Yong Wang, Liguang Zhang, Ting-Chuen Pong, and Huamin Qu. 2020. Emotioncues: Emotion-oriented visual summarization of classroom videos. *IEEE transactions on visualization and computer graphics* 27, 7 (2020), 3168–3181.
- [78] Ferdinand Rudolf Hendrikus Zijlstra and L Van Doorn. 1985. The construction of a scale to measure subjective effort. *Delft, Netherlands* (1985).