From Individual QoE to Shared Mental Models: A Novel Evaluation Paradigm for Collaborative XR

Sam Van Damme*, Jack Jansen†, Silvia Rossi†, Pablo Cesar†‡
* IDLab, Department of Information Technology (INTEC), Ghent University - imec, Ghent, Belgium
† Distributed & Interactive Systems Group (DIS), Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands
‡ TU Delft, Delft, The Netherlands
Email: Sam.VanDamme@UGent.be, Jack.Jansen@CWI.nl, Silvia.Rossi@CWI.nl, Pablo.Cesar@CWI.nl

Abstract—**Extended Reality (XR) systems are rapidly shifting from isolated, single-user applications towards collaborative and social multi-user experiences. To evaluate the quality and effectiveness of such interactions, it is therefore required to move beyond traditional individual metrics such as Quality-of-Experience (QoE) or Sense of Presence (SoP). Instead, group-level dynamics such as effective communication, coordination etc. need to be encompassed to assess the shared understanding of goals and procedures. In psychology, this is referred to as a Shared Mental Model (SMM). The strength and congruence of such an SMM are known to be key for effective team collaboration and performance. In an immersive XR setting, though, novel Influence Factors (IFs) emerge that are not considered in a setting of physical co-location. Evaluations on the impact of these novel factors on SMM formation in XR, however, are close to non-existent. Therefore, this work proposes SMMs as a novel evaluation tool for collaborative and social XR experiences. To better understand how to explore this construct, we ran a prototypical experiment based on ITU recommendations in which the influence of asymmetric end-to-end latency is evaluated through a collaborative, two-user block building task. The results show how also in an XR context strong SMM formation can take place even when collaborators have fundamentally different responsibilities and behavior. Moreover, the study confirms previous findings by showing in an XR context that a teams' SMM strength is positively associated with its performance.**

Index Terms—**Collaborative Extended Reality (XR), Shared Mental Model (SMM), end-to-end latency, block building**

I. Introduction

Extended Reality (XR) systems are shifting from isolated, single-user applications to collaborative and social experiences [1]–[4]. To evaluate the quality and effectiveness of such interactions, it is therefore required to move beyond traditional individual metrics such as Quality-of-Experience (QoE) or Sense of Presence (SoP). While these offer valuable insights into a user's personal experience, they fall short in capturing the nuances of collaboration that underpin successful teamwork such as effective communication, coordination, and a shared understanding of goals and procedures. In psychology, this is referred to as a Shared Mental Model (SMM), i.e., a shared understanding and interpretation of the environment and tasks among team members [2], [5], [6]. The strength and congruence of this SMM are known to be key for effective team collaboration and performance. In addition, it is recognized as a mediator for other outcomes such as effectiveness in collaborative learning [7], [8]. As such, SMM assessment provides valuable insights that traditional individual measurements often miss therefore holding potential to become a standard assessment approach for social and collaborative XR.

In an immersive setting, though, novel Influence Factors (IFs) emerge that are not considered in case of physical co-location, such as the virtual environment [9], avatar representation [10], virtual task [11], network [2], etc. As studies evaluating such influences on SMM formation in XR are close to non-existent, this work proposes SMMs as a novel evaluation tool for collaborative and social XR experiences. To better understand how to explore this construct, we ran a prototypical experiment based on ITU recommendations in which the influence of asymmetric end-to-end latency is evaluated through a collaborative, two-user block building task [12], [13]. The results show how also in an XR context strong SMM formation can take place even when collaborators have fundamentally different responsibilities and behavior, e.g., conversational speaking time. Moreover, the study confirms previous findings by showing in an XR context that a teams' SMM strength is positively associated with its performance.

The remainder of this paper is organized as follows. Section II gives a brief overview of the state of the art related to the measurement, IFs, and application of SMMs in collaborative and social XR. Next, Section III describes the technical setup, the specific collaborative task and the evaluation methodology. In Section IV, the most prominent findings of this work are presented. Section V, at last, summarizes this work by listing the most important conclusions.

II. Related Work

This Section first provides an overview of SMMs and their evaluation methods in literature, followed by their most important IFs. Next, a number of studies that explore SMM evaluation in an XR context are outlined.

A. Defining and measuring SMMs

SMMs can be defined as collective understandings or representations held by members of a group that help them interact effectively and coordinate actions in a collaborative

Sam Van Damme was funded by Research Foundation Flanders (FWO) during the conduction of this research (grant nrs. 1SB1824N, V463424N). This work was supported by the European Union as part of the Horizon Europe Framework Program under grant agreement No 101070109 (TRANSMIXR).

TABLE I: Overview of related studies

Authors	Focus	SMM measurment	Key findings
Kiss et al. [14]	VR vs. face-to-face collaborative learning	Banks and Millward [15])	Teams work efficiently in both VR and F2F with similar SMM development.
Slezak et al. [16]	SMM in multi-user VR model of ISS via design and risk assessment task	Observational pilot study	Students struggled to form a robust SMM, limiting success in risk assessment.
Milef et al. [2]	Impact of QoS (latency, packet bursts) on SMM and task performance in collaborative VR firefighting simulator	Subjective QoE, responsiveness, interaction, and performance through Likert scales	Large latencies impact QoE, while packet bursts impact both QoE and performance. Network role (client/server) also relevant.
Bröring [17]	SMM formation in group work using immersive technology	Observational case study, sequential analysis of student work sessions	Students did not prioritize SMM, working individually and dividing subtasks. Communication was sparse, used to stay updated about the other students' status.

setting. These models encompass shared knowledge about roles, tasks, tools, and processes, enabling team members to anticipate each other's actions and adjust their own behavior accordingly. As such, they are critical for teamwork, decision-making, and maintaining common ground. In this sense, three key characteristics of SMMs can be distinguished: **(i)** *Knowledge structures*, i.e. SMMs organize knowledge about tasks, team roles, equipment, and communication patterns; **(ii)** *Predictive utility*, as they allow team members to predict others' behaviors and adapt to changes in goals or processes and **(iii)** *Collaboration*, as SMMs enhance coordination by fostering mutual understanding and reducing conflicts. From this perspective, SMMs could be regarded as the degree to which individual perceptions or QoE align or complement each other to foster effective collaboration.

In (psychological) literature, SMM strength is typically subdivided in the *SMM of the task* (i.e. shared knowledge and interpretation about work objectives, team resources, task procedures, and practices) and the *SMM of the team* (i.e. understanding interpersonal interactions, team members' roles, responsibilities, and role interdependencies) [5]. In XR, SMMs can therefore be related to (i) the *task* itself (e.g. complexity, engagement), (ii) *social presence* to synchronize individual mental models through (non-)verbal cues and (iii) *spatial presence*, to anchor mental models to the shared visualization. To evaluate SMMs, multiple techniques are put forward such as *multidimensional scaling*, *interactive or text-based mapping techniques* [18], [19], *cognitive task analysis* [18], and *similarity rating scales* [20]. For the latter, the *SMM Rating Scale (SMMRS)* [6] is worth pointing out as it is one of few methods for SMM assessment in a technology-mediated collaborative setting. Furthermore, the *Five Factor Perceived Shared Mental Model Scale (5-PSMMS)* [5] is worth mentioning given its fine-grained SMM evaluation via subscales on equipment, execution, interaction, composition and temporal aspects.

B. SMM Influence Factors

Multiple IFs of SMM formation can be identified, including **(i)** task complexity [21], **(ii)** team size [6], **(iii)** time pressure [21], **(iv)** communication quality [6], **(v)** the clarity of each member's role [22] and **(vi)** familiarity and individual differences between team members [22], [23]. Regarding XR, there is limited evidence that **(1)** SoP, **(2)** cybersickness, **(3)** avatar and embodiment quality, **(4)** intuitivity of spatial movement and reasoning and **(5)** network distortions such as latency and desynchronization contribute to the SMM [24]. Especially the latter is worth pointing out as poor and asymmetric QoS between collaborators can lead to a divergence of the individual interpretations of the shared environment and task, ultimately disrupting the aforementioned SMM [2]. Furthermore, a *transfer* of the SMM can take place between subsequent tasks in case of sufficient task similarity, a limited time interval in between, no or few intervening experiences and the ability to learn and repeat communication patterns/strategies between team members [24].

C. SMMs in Social and Collaborative XR

Table I shows four related studies. Although all of these explore SMM evaluation in XR, there is limited focus on the specific IFs intrinsic to *remote* collaboration. Only *Milef et al.* [2] evaluate the impact of QoS. Moreover, there is no clear assessment methodology with only one study using a well-established method [14] while others rely on observations or self-constructed questions. Furthermore, each of these studies is focusing on very specific use cases such that their generalizability and ecological validity can be questioned.

As such, this work aims to address these gaps by focusing on an XR-specific IF, i.e. (asymmetric) latency, of which the impact on SMM formation is validated through two validated questionnaires (SMMRS [6] and 5-PSMMS [5]). By adopting a standardized block-based task based on ITU Recommendations [12], [13], the relevance, generalizability and ecological validity of this study are further strengthened.

III. Experimental methodology

This Section presents the followed evaluation methodology. First, Section III-A provides a description of the technical setup. Next, Section III-B introduces the task to be performed. This is followed by a description of the procedure for the subjective and objective user study in Section III-C.

A. Technical setup

To limit the influence of artificial avatar representations on SMM forming [24], we used the open-source *VR2Gather*-framework [25], [26]. This is a customizable, end-to-end system to transmit volumetric contents in multiparty, real-time communication. To this end, two identical setups (Fig. 1) were made in two physically distant rooms. Each setup consists of four *Microsoft Azure Kinect* depth cameras **(1 in Fig. 1)**, each on top of a tripod **(2 in Fig. 1)**. These are put in a

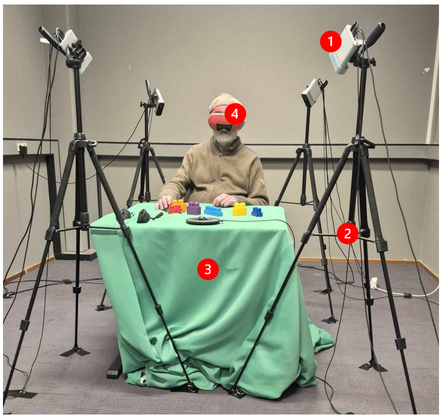


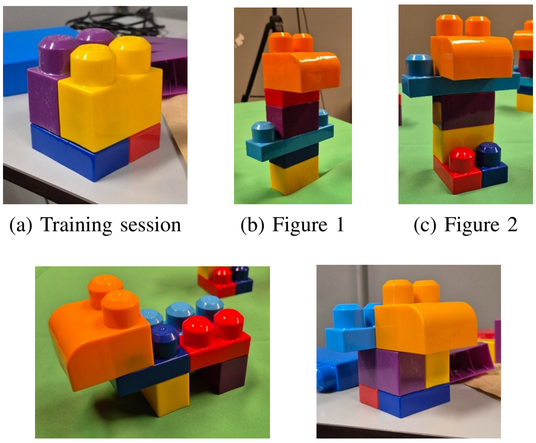
Fig. 1: A participant conducting the experiment.

square constellation around a physical table and chair **(3 in Fig. 1)**. The table is decorated with a green cloth to remove its surface from the virtual environment via green screen removal. Two *Meta Quest Pro* headsets **(4 in Fig. 1)** are used to immerse participants which are wirelessly (*Meta Air Link*) connected to a *Windows 10 Enterprise* gaming PC. In the first room, this is a *four-core i7-770K* running at *4.2 GHz*, with *32 GB* of memory, and an *Nvidia GeForce GTX 1080 Ti* graphics card. In the second room, a *12-core i9-7920X* running at *2.9 GHz*, with *128 GB* of memory, and *dual-Nvidia GeForce GTX 1080 Ti* graphics cards is used. Both machines are directly connected via a dedicated Gigabit Ethernet link. *Unity* version *2022.3.21f1* was used allowing to plug our custom scene into the existing *VR2Gather* framework. In addition, we enabled the open-source *VRTstatistics*-software for tracking task duration and extended its code to track the in- and output signal power of microphone and audio channel. The later allows to monitor whether or not a participant is talking at a given point in time.

B. *Task description*

To evaluate SMM construction and the influence of (asymmetric) end-to-end delay, we picked the *block building task* from the *ITU-T P.920* and *ITU-T P.1305* standards for interaction assessment in videoconferencing [12], [13]. One of both participants takes the role of the *instructor* while the other becomes the *builder*. Using a pre-build figure (Fig. 2), the goal for the *instructor* is to guide the *builder* in recreating the identical figure from its individual building blocks. As such, it covers the crucial aspects of collaboration as defined by Pérez et al. [27]: *deliberation* (consultation between collaborators), *exploration* (identification of objects), and *manipulation* (interaction with system elements/objects).

To adapt this task to XR, as it was originally intended for 2D video conferencing, the open-source *cwipc* [28] and *VR2Gather*-software [25], [26] are used for multi-camera volumetric capturing of both the building blocks and the remotely connected users. Afterwards, both users and their respective blocks are brought together in a shared, Unity-based virtual environment, which consists of a virtual table. The point cloud representations of both users are placed on opposite sides of this table, facing each other (Fig. 3). The quality of the blocks' representations was artificially limited to avoid participants completely relying on the visual channel and to encourage audiovisual communication and interaction. Both channels are always delayed simultaneously, i.e. no desynchronization between audio and video takes place. An open-source implementation of the above is publicly available at https://github.com/samdamme/CWI-VR-SMM.



(a) Training session (b) Figure 1 (c) Figure 2

(d) Figure 3 (e) Figure 4

Fig. 2: Overview of the 5 different figures.

Since the task requires little skill or prior knowledge [21]–[23], involves two collaborators with clearly defined roles [22] and gives no communication constraints [6], SMM formation is expected to happen relatively fast (i.e. order of minute(s)). The realistic point cloud representation [6] and limited required movement (intuitive + minimal motion sickness) [4] add to this assumption [6]. A transfer of the SMM is expected between subsequent, similar playthroughs limited time in between, as pairs can learn and repeat communication strategies [24]. Therefore, similar to learning and order effects, counterbalancing strategies are put in place such as a prior training session, Greco-Latin square ordering, intervening in-session questionnaires, and sufficiently different figures.

C. *Evaluation methodology*

Table II shows the steps of a single experimental session. Note that two users participate concurrently and collabora-

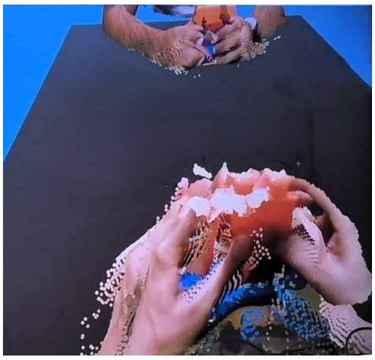


Fig. 3: The block building application as seen by the builder.

TABLE II: Experimental flow of a session with timing and content of the subjective evaluations and objective metrics. Particular subscales are indicated between parentheses.

Welcome & introduction	
Informed consent	
Pre-session questionnaire	• Ishihara tests [29] • Demographics (age, gender, nationality) • Visual, auditive or motor impairments • Self-assessed sensitivity to motion-sickness, familiarity with XR, technological proficiency • Baseline VRSQ [30]
Training session	
Playthrough n (max. 10 min., 4x)	
In-session questionnaire n (4x)	• Subjective performance (Global QoE, System Annoyance, Delay Perception, Interruptions) [31] • Presence (Involvement, Adaptation, Accomplishment) [31] • Social Factors (Social Presence, Social Annoyance, Social Adaptation, Collaboration) [31] • SMMRS (Task, Team) [6]
	• Task duration • Talking/no talking states
Post-session questionnaire	• VRSQ [30] • NASA TLX [32] • 5-PSMMS (Execution, Interaction, Composition, Temporal) [5]
Thanks & closing	

TABLE III: Overview of the four experimental conditions.

Condition	Latency builder (ms)	Latency instructor (ms)
A	0	0
B	1200	0
C	0	1200
D	1200	1200

tively, and that they are assigned randomly to the pairs. First, participants are welcomed and given a brief introduction on the purpose and organization of the experiment, including oral instructions on the use of XR (headset, boundaries...) and the actual block building task. Next, user roles (*builder* vs. *instructor*) are assigned randomly and each of the roles is further clarified. Participants can ask questions in case anything is unclear, which are answered accordingly.

Next, they are given an informed consent form asking for permission to collect and process their anonymized data and for their personal information, such as demographics, to be used in the study. Participants are made aware that the signal power of microphone and audio channel are monitored and that the possibility of experiencing cybersickness exists. Moreover, they are informed on their rights regarding the withdrawal from the study and data access, correction, deletion rights. Next, participants conduct *Ishihara* tests [29] to check for possible color-blindness, followed by the completion of a pre-session questionnaire on age, gender, and nationality. They are asked to report any visual, auditive or motor impairments and to assess their own sensitivity to motion sickness, familiarity with XR and technological proficiency in general. In addition, a prior baseline VRSQ [30] is filled.

Following this, both participants are accompanied to their respective room, chair and table and are requested to put on the Head-Mounted Display (HMD). They are instructed that no limitations are posed on the allowed communication and interaction, apart from remaining seated, keeping the HMD on and holding the conversation in English, i.e. not to favor pairs that have a common language. Afterwards, they are connected remotely for the training session, which is conducted without any artificial delay being induced and with a straightforward block figure (Fig. 2a). No data is recorded at this point, and participants are allowed to ask questions which are addressed accordingly. In parallel, the researcher confirms that all technical requirements are fulfilled.

Next, the actual experiment takes place in a within-subjects design of four consecutive rounds of 10 minutes max. each. Each round consists of a different network condition (A to D, Table III) and figure (1 to 4, Fig. 2). Latencies are added on top of the intrinsic system delay (+-*300 ms*). The value of *1200 ms* is chosen to be sufficiently above the noticeability threshold of *900 ms* as determined by *Cortès et al.* [31]. The order of each session's testing rounds is determined using a balanced *Greco-Latin* square design [33] to mitigate ordering and learning effects. Participants are intentionally kept unaware of these conditions. After each round, participants remove the HMD and fill the in-session questionnaire. For this, we extended the 11 questions determined and validated by *Cortès et al.* [31] (subjective performance, presence, social factors) with the *SMMRS* [6]. The latter consists of 4 questions in the *SMM Task* subscale and 5 in the *Team Process Satisfaction* subscale (7-point Likert). Questions expressing negative opinions are inverted (i.e., higher is better) before averaging.

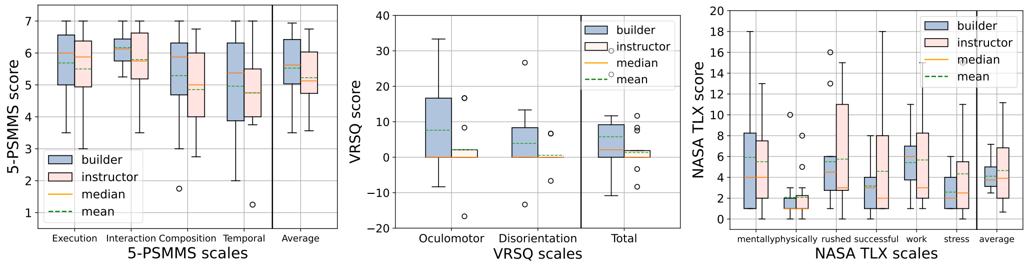
After these four rounds, participants fill a post-session questionnaire consisting of a second *VRSQ* [30], a *NASA TLX* [32] to assess cognitive load, and the *5-PSMMS* [5] as a second measurement of SMM strength. The 5-PSMMS consists of 5 subscales (*Equipment, Execution, Interaction, Composition, Temporal*) with four 7-point Likert questions each. All items share the item stem "Team members have a similar understanding about..." (i.e., higher is better) which are averaged into per-scale and global scores. Note that the *Equipment*-scale was dropped as not relevant for this task. Once the post-session questionnaire was finished, participants were thanked and the session was closed. Each session did not take longer than 1 hour maximum.

IV. Results

This Section first provides an overview of the participants' demographics. Next, the strength and agreement of SMM construction are discussed, followed by the objective measurements related to task duration and conversational activity time. Note that the obtained dataset is publicly available at https://cloud.ilabt.imec.be/index.php/s/tL9NRRsLqttNHCd.

A. Participants

Participants were gathered on a voluntary basis from academia through online and ad valvas announcement and were offered a €10 voucher as an incentive. Volunteers suffering from color-blindness (*Ishihara tests*) or non-correctable visual, auditive, or motor impairments that interfere with the

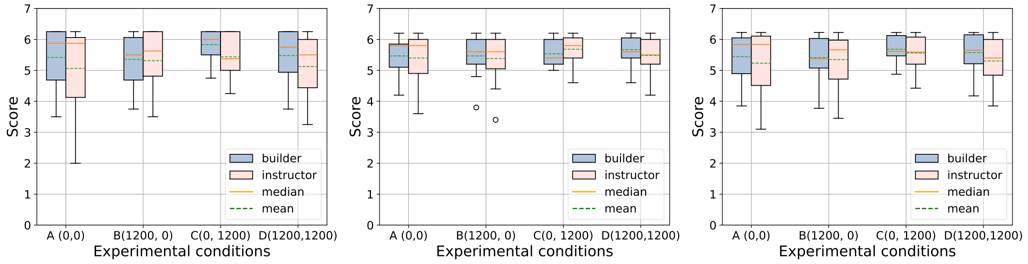


(a) 5-PSMMS (1-7 scale)

(b) VRSQ (Difference, 0-100 scale)

(c) NASA TLX (0-20 scale)

Fig. 4: Score distribution of the 5-PSMMS, VRSQ, and NASA TLX subscales and average for both builder and instructor.



(a) Task (0-7 scale)

(b) Team (0-7 scale)

(c) Total (0-7 scale)

Fig. 5: Boxplots with the distribution of scores on the different SMMRS subscales: task (a), team (b), and total (c). A distinction is made between the user roles of builder and instructor as well as between the different experimental delay conditions (A-D).

experience are excluded from participation. Furthermore, volunteers having insufficient English proficiency to understand instructions and/or hold a fluent conversation are withheld. A total of 24 participants conducted the experiment, which were randomly assigned in pairs. 13 of them (54.4%) identified as female, 10 (41.7%) as male and 1 (4.2%) as non-binary. Ages vary between 18 and 72 with an average of 32.7 and a median of 27. 8 participants had the Dutch nationality (33.3%), 7 (29.2%) the Indian and 4 (16.7%) the Italian. The remaining nationalities (21%) include Chinese, German, American, Belgian, and Kazakh. 16 participants (66.7%) indicated not to be prone to motion sickness, 6 (25%) to be sensitive to this and 2 of them (8.3%) were unsure. 2 participants (8.3%) are very familiar with XR, 11 (45.8%) somewhat and 9 (37.5%) a little, while 2 participants (8.3%) indicated never to have used XR. 6 participants (25%) assess their own technological proficiency as very strong, 10 (41.7%) as strong, 7 (29.7%) as moderate and 1 (4.2%) as poor.

B. SMM strength and agreement

Fig. 4a shows the distributions of the post-session 5-PSMMS-scores on a 1-7 scale, for each role and subscale as well as on average. Relatively high scores are observed with respective averages of *(5.69, 5.5)*, *(6.17, 5.79)*, *(5.29, 4.85)*, *(4.96, 4.75)*, and *(5.53, 5.22)* for each (*builder*, *instructor*)-pair. *Wilcoxon Signed Rank Tests* for ordinal data did not show any significant difference between user roles for each of the respective scales [*(W=32.5, p=0.63, RBC=0.17)*, *(W=16.0, p=0.27, RBC=0.42)*, *(W=20.5, p=0.52, RBC=0.25)*, *(W=20.0, p=0.82, RBC=0.11) and (W=27.0, p=0.37, RBC=0.31)*]. As such, this shows a prior indication on how participants also in a collaborative XR context **(a)** perceive a strong SMM and **(b)** show a strong level of agreement on this SMM strength between collaborators. Therefore, this poses an interesting finding as it opens the door towards SMMs as a potential evaluation tool for collaborative and social XR.

This is partly induced by the limited occurrence of cybersickness (Fig. 4b) while at the same time maintaining a sufficiently high cognitive load to establish such an SMM (Fig. 4c). More important, however, is that there are no significant differences between user roles to be observed. As such, there are no fundamental differences in individual perception that could lead to desynchronized interpretations or understanding of task and team, eventually leading to a disruption of the SMM. Nevertheless, a somewhat wider spread of VRSQ-scores for the builder role can be observed, which can be attributed to a higher degree of required exploration and interaction to build the figure.

This strong SMM forming is confirmed by the in-session *SMMRS* (Fig. 5). For each subscale, condition, and user role, high scores with medians and averages between 5 and 6 on a 0-7 scale are observed. Once again, no significant differences between user roles are to be reported. Notably, none of the subscales or roles shows a significant difference between conditions as calculated through *Friedman tests* (Table IV). This is supported by the in-session question on *delay perception* (*Did you perceive any reduction in your ability to interact during the conversation due to delay?* | $\chi^2_{\text{builder}} = 0.64$, $p_{\text{builder}} = 0.89$, $W_{\text{builder}} = 0.02$, $\chi^2_{\text{instructor}} = 0.62$, $p_{\text{instructor}} = 0.89$, $W_{\text{instructor}} = 0.02$). This is remarkable as the similar study by *Cortès et al.* [31], although not focusing on SMMs, concluded

TABLE IV: χ^2, p and Kendall's W values of the Friedman tests as conducted for each SMMRS subscale and user role over the four experimental conditions.

Scale	Role					
	Builder			Instructor		
	χ^2	p	W	χ^2	p	W
Task	4.65	0.20	0.13	1.34	0.72	0.04
Team	1.31	0.73	0.04	2.25	0.52	0.06
Total	1.82	0.61	0.05	0.52	0.91	0.01

TABLE V: Mean, median, and std of the task durations for each condition as well as the PLCC (r_{PLCC}) with the total SMMRS (as averaged between builder and instructor).

Condition	Mean (s)	Median (s)	Std. (s)	r_{PLCC}
A	154.0	133.0	71.45	-0.841 ($p < 0.01$)
B	156.45	131.0	107.5	-0.784 ($p < 0.01$)
C	144.18	122.0	58.07	-0.251 ($p = 0.43$)
D	154.82	121.0	75.36	-0.543 ($p = 0.07$)
Overall	158.17	131.0	82.23	-0.662 ($p < 0.001$)

a strong noticeability threshold of 900 ms. A possible explanation is that the blocks' point cloud representations require a higher cognitive load (Fig. 4c) compared to the photorealistic representation of *Cortès et al.*, therefore leaving less working memory to perceive such differences. As *Cortès et al.* did not measure cognitive load, further research is required to confirm this assumption.

C. Objective task duration

Table V shows the mean, median, and standard deviation of the task durations (in s) for each condition and overall. As the assumptions for Repeated Measures Analysis of Variance (RMANOVA) on sphericity and normality were violated, a *Friedman test* was conducted between conditions. No significant differences were observed ($\chi^2 = 0.47$, $p = 0.93$, $W = 0.01$). Pearson Linear Correlation Coefficients (PLCCs) were also calculated between the per-session duration and the total *SMMRS* as averaged between *builder* and *instructor*. As shown in Table V, significant correlations of -0.841 and -0.784 are observed between task duration and SMM strength for conditions A and B as well as a -0.662 correlation overall. As such, the above results offer prior indications that the relation between a team's SMM strength and its according performance also holds in collaborative XR.

D. Conversational interactions

Fig. 6 shows the activity time per user and condition as a percentage of the total session duration. This is calculated by first normalizing the audio signals to -26dBov following ITU-T P.56 [31], [34] and then calculating the squared mean amplitude of each 200 ms audio segment. Any segment with a dBFS exceeding a threshold value of -16dBFS [31], [34] is classified as active. The percentage is then calculated as the fraction of active segments relative to the total number of 200 ms segments within a session.

RMANOVA comparisons between conditions do not indicate any significant difference in activity time for neither the builder ($F = 0.22$, $p = 0.88$, $\eta^2_G = 0.008$) nor the instructor ($F = 0.20$, $p = 0.90$, $\eta^2_G = 0.003$). The difference between user roles, however, shows to be strongly significant for all experimental conditions (*paired t-test*| $p_i < 0.01$, $d_i > 0.8$, $i = A, B, C, D$). As such, this indicates that also in social or collaborative XR, individual differences in user behavior do not necessarily lead to a disagreement on, or disruption of, the SMM which further strengthens the idea that in order to accurately evaluate experience in collaborative XR, we need to dare step beyond individual perception and user behavior and look at the interaction from a collective viewpoint.

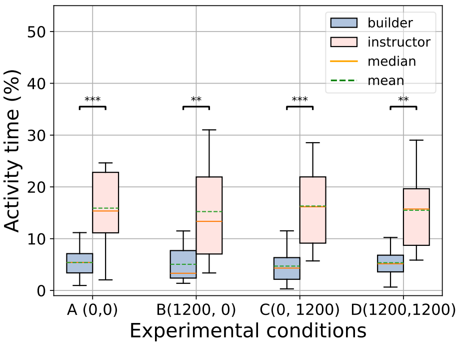


Fig. 6: Activity time per user and experimental condition as a percentage of the total duration of the session. ** (p<0.01) and *** (p<0.001) indicate statistically significant differences.

V. Conclusion

This work illustrated the potential of SMMs as a novel evaluation tool for collaborative and social XR. Through a prototypical ITU-based experiment including a two-user block building task in different (asymmetric) latency conditions, it was shown how also in an XR context strong SMM formation can take place even when collaborators have fundamentally different responsibilities and behavior. Moreover, the study confirms previous findings by showing in an XR context that a teams' SMM strength is positively associated with its performance. As such, the introduction of SMMs in XR opens up possibilities to assess and model collaboration beyond individual QoE where the strength of the SMM can be seen as a common mediator of multiple outcomes of interest such as team performance or collaborative learning effectiveness. Further extensions include the adaptation of SMM evaluation methods beyond questionnaires and communication analysis, i.e. by means of concept maps, card sorting tasks, observational study, or analyzing paralinguistic content from the speech signals. One could even explore objectively measurable concepts such as posture mirroring or physiological synchrony as indicators of alignment between collaborators, potentially correlating to SMM strength. Furthermore, given the limited generalizability from 12 user pairs, a repetition of the above experimentation on a larger sample size to validate results is recommended. In addition, it would be of interest to investigate whether the above conclusions hold for other task types as well as to extent from pairs to triads or even larger user groups to evaluate how feelings of being left out affect SMM formation.

REFERENCES

[1] Shannon M Moore and Michael N Geuss. Familiarity with teammate's attitudes improves team performance in virtual reality. *PloS one*, 15(10):e0241011, 2020.

[2] Nicholas Milef, Adam Ryason, Di Qi, Samuel O Alfred, Cullen D Jackson, and Suvranu De. Disruptions to shared mental models from poor quality of service in collaborative virtual environments. *Scientific reports*, 11(1):23556, 2021.

[3] Anne Massey, Mitzi Montoya, Binny M Samuel, and Jaime Windeler. Presence and team performance in synchronous collaborative virtual environments. *Small Group Research*, 55(2):290–329, 2024.

[4] David Roberts, Robin Wolff, Oliver Otto, and Anthony Steed. Constructing a gazebo: supporting teamwork in a tightly coupled, distributed task in virtual reality. *Presence*, 12(6):644–657, 2003.

[5] Jandre J. van Rensburg, Catarina M. Santos, Simon B. de Jong, and Sjir Uitdewilligen. The five-factor perceived shared mental model scale: A consolidation of items across the contemporary literature. *Frontiers in Psychology*, 12, 2022.

[6] Hayward P Andres. Technology-mediated collaboration, shared mental model and task performance. *Journal of Organizational and End User Computing (JOEUC)*, 24(1):64–81, 2012.

[7] Catarina Marques dos Santos. *Shared mental models and shared temporal cognitions: contributions to team processes and team effectiveness.* ISCTE-Instituto Universitario de Lisboa (Portugal), 2016.

[8] Piet Van den Bossche, Wim Gijselaers, Mien Segers, Geert Woltjer, and Paul Kirschner. Team learning: building shared mental models. *Instructional science*, 39:283–301, 2011.

[9] David Roberts, Robin Wolff, Oliver Otto, and Anthony Steed. Constructing a gazebo: supporting teamwork in a tightly coupled, distributed task in virtual reality. *Presence*, 12(6):644–657, 2003.

[10] Daniel Roth, Jean-Luc Lugrin, Dmitri Galakhov, Arvid Hofmann, Gary Bente, Marc Erich Latoschik, and Arnulph Fuhrmann. Avatar realism and social interaction quality in virtual reality. In *2016 IEEE virtual reality (VR)*, pages 277–278. IEEE, 2016.

[11] Curtiss Murphy. Why games work and the science of learning. In *Selected Papers Presented at MODSIM World 2011 Conference and Expo*, 2012.

[12] Recommendation P ITU-T. 920, interactive test methods for audiovisual communications. *International Telecommunications Union Radiocommunication Assembly*, 1996.

[13] Recommendation P ITU-T. 1305, effect of delays on telemeeting quality. *International Telecommunications Union Radiocommunication Assembly*, 2016.

[14] Laura Kiss, Balázs Péter Hámornik, and Máté Köles. Development of shared knowledge in a virtual reality environment for collaborative learning. *Re-Imagining Learning Scenarios*, 2016.

[15] Adrian P Banks and Lynne J Millward. Differentiating knowledge in teams: The effect of shared declarative and procedural knowledge on team performance. *Group Dynamics: Theory, Research, and Practice*, 11(2):95, 2007.

[16] Robert J Slezak, Nir Keren, and Tor Finseth. Virtual reality application for enhancing risk assessment skills. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC), Orlando, FL, USA*, volume 26, 2018.

[17] Tabea Bröring. Shared situation awareness in student group work when using immersive technology, 2023.

[18] Nor'ain Mohd Yusoff and Siti Salwah Salim. Shared mental model processing in visualization technologies: A review of fundamental concepts and a guide to future research in human-computer interaction. In *Engineering Psychology and Cognitive Ergonomics. Mental Workload, Human Physiology, and Human Energy: 17th International Conference, EPCE 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I 22*, pages 238–256. Springer, 2020.

[19] Susan Mohammed, Lori Ferzandi, and Katherine Hamilton. Metaphor no more: A 15-year review of the team mental model construct. *Journal of management*, 36(4):876–910, 2010.

[20] L. A. DeChurch and J. R. Mesmer-Magnus. Measuring shared team mental models: A meta-analysis. *Group Dynamics: Theory, Research, and Practice*, 14(1):1–14, 2010.

[21] Andrea Gurtner Thomas Ellwart, Christian Happ and Oliver Rack. Managing information overload in virtual teams: Effects of a structured online team adaptation on cognition and performance. *European Journal of Work and Organizational Psychology*, 24(5):812–826, 2015.

[22] Tristan E Johnson, Ercan Top, and Erman Yukselturk. Team shared mental model as a contributing factor to team performance and students' course satisfaction in blended courses. *Computers in Human Behavior*, 27(6):2330–2338, 2011.

[23] Wu Jing. Study on the relationship between mental model of designer team and factors in remote collaboration. In Constantine Stephanidis, editor, *HCI International 2016 – Posters' Extended Abstracts*, pages 46–51, Cham, 2016. Springer International Publishing.

[24] Andra F. Toader and Thomas Kessler. Task variation and mental models divergence influencing the transfer of team learning. *Small Group Research*, 49(5):545–575, 2018.

[25] Jack Jansen, Thomas Röggla, Silvia Rossi, Irene Viola, and Pablo Cesar. Open-sourcing vr2gather: A collaborative social vr system for adaptive multi-party real time communication. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 11210–11213, New York, NY, USA, 2024. Association for Computing Machinery.

[26] CWI DIS Group. Vr2gather: Github repository, 2025. Accessed: 2025-03-25.

[27] Pablo Pérez, Ester Gonzalez-Sosa, Jesús Gutiérrez, and Narciso García. Emerging immersive communication systems: Overview, taxonomy, and good practices for qoe assessment. *Frontiers in Signal Processing*, 2, 2022.

[28] CWI DIS Group. cwipc: Github repository, 2025. Accessed: 2025-03-25.

[29] John H. Clark. The ishihara test for color blindness. *American Journal of Physiological Optics*, 5:269–276, 1924.

[30] Hyun K. Kim, Jaehyun Park, Yeongcheol Choi, and Mungyeong Choe. Virtual reality sickness questionnaire (vrsq): Motion sickness measurement index in a virtual reality environment. *Applied Ergonomics*, 69:66–73, 2018.

[31] Carlos Cortés, Irene Viola, Jesús Gutiérrez, Jack Jansen, Shishir Subramanyam, Evangelos Alexiou, Pablo Pérez, Narciso García, and Pablo César. Delay threshold for social interaction in volumetric extended reality communication. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(7), April 2024.

[32] SG Hart. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Human mental workload/Elsevier*, 1988.

[33] James R Lewis. Pairs of latin squares to counterbalance sequential effects and pairing of conditions and stimuli. In *Proceedings of the Human Factors Society Annual Meeting*, volume 33, pages 1223–1227. SAGE Publications Sage CA: Los Angeles, CA, 1989.

[34] ITU. Objective measurement of active speech level. Recommendation ITU-T P.56, International Telecommunication Union (ITU), 1993.