

Virtual Reality–Based Training in Mental Health Literacy and Health- Literate Communication: A Mixed-Methods Pilot Study with VR AI-Driven Dialogue

Thomas Herbert Kokholm

CoFlow Visuals ApS

Britta Jensen

Aalborg University

Henrik Bøggild

Aalborg University

Kristine Sørensen

krisoe@hst.aau.dk

Aalborg University

Research Article

Keywords: virtual reality, mental health literacy, health-literate communication, simulation-based education, health professions education, artificial intelligence assessment

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10876926/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Mental health literacy involves recognizing, understanding, and responding to mental health conditions, yet health professions students often lack structured training in health-literate communication. Virtual reality (VR) offers immersive simulation, but its use in this field remains largely unexplored.

Objective

This pilot study examined the feasibility and preliminary outcomes of a VR-based training intervention targeting mental health literacy and health-literate communication among pre- and postgraduate health students.

Methods

Twenty-seven graduate students at two institutions completed a single-session VR intervention with two clinical scenarios; 16 completed the post-intervention evaluation and 17 had rubric-scored sessions. Data collection followed a sequential mixed-methods design at three time points: a pre-intervention questionnaire (Mental Health Literacy Scale, MHLS; HLS19-Q12), a structured group debrief, and a post-intervention Technology Acceptance Model questionnaire. Transcripts were scored by an artificial intelligence (AI)-driven rubric of 14 criteria in two domains (Content and Communication).

Results

Baseline mental health literacy was high (MHLS mean 134.5/160); 56% of participants reached a sufficient or excellent general health literacy level. Post-session ratings showed high perceived usefulness (4.11/5), ease of use (4.08/5), and behavioral intention (4.05/5). Rubric performance was low (mean 32.5%; range 5.1%–59.5%), with near-absent teach-back (1 of 17 sessions) and supportive framing (5 of 17). Empathy-related criteria (48.2%–59.1%) clearly outperformed health literacy criteria (11.8%–14.1%). Participants experienced VR as more engaging than role-play but noted limited avatar expressiveness and conversational latency.

Conclusions

The pilot demonstrates a functional workflow for VR-based mental health literacy training with AI-driven assessment. Floor effects in health literacy techniques against adequate empathic performance identify the competence gap targeted, and acceptance indicators support larger-scale study.

1 Introduction

Mental health literacy, defined as knowledge and beliefs about mental disorders that aid their recognition, management, and prevention [1], is increasingly recognized as a core competency for health professionals beyond psychiatric specialization [2]. The concept extends beyond diagnostic knowledge to attitudes, stigma, and the communicative practices through which professionals help patients understand and act on health information [3]. The relational approach aligns with the broader health literacy framework articulated by Sørensen et al. [4], which frames health literacy not merely as a patient attribute but as a professional and organizational competency that shapes whether information becomes actionable.

Mental health literacy and general health literacy are essential for recognizing symptoms, navigating services, making informed decisions and seeking timely care. This is particularly important because one in eight people worldwide lives with a mental disorder, and limited health literacy is recognized as a barrier to help-seeking [5]. Moreover, European health literacy surveys from 2011 and 2019 showed that almost 50% of the participants faced health literacy challenges. These findings underline the need for health-literate communication, including clear language, culturally responsive information and systematic confirmation of understanding [6, 7] notably, a systematic review found positive outcomes from teach-back in 19 of 20 studies, although study quality and outcomes varied [8].

Despite this recognition, structured training in mental health literacy, health literacy and health-literate communication remains sparse in health professions curricula [9]. Traditional approaches to communication training in healthcare face well-documented limitations: lectures lack interactivity and emotional engagement [10]. Role-play is often experienced as awkward and cognitively burdensome, particularly with familiar peers [11]. Simulated Patients (SP) offer realism but are resource-intensive, difficult to standardize across sites, and impractical for repeated practice. Common to all three modalities is that communicative performance is itself difficult to assess formally and consistently, as evaluation depends on trained human raters and largely unstandardized observation [12].

1.1 Virtual Reality in Health Professions' Health Literacy and Communication Training

Virtual reality (VR) responds to these constraints by offering immersive, standardized and repeatable clinical encounters in a psychologically safe environment, and evidence from nursing and medical education indicates that VR-based communication training is usable, feasible and effective [13–17].

Artificial intelligence (AI) extends VR by enabling real-time conversational interaction through natural language processing and voice recognition, turning static scenarios into responsive clinical encounters [18, 19].

AI-driven evaluation further enables automated, standardized assessment of learner performance from session transcripts [20]. This addresses a key scalability constraint in simulation-based education:

reliance on trained human evaluators for formative feedback [21].

With regards to general health literacy, Skidmore et al examined VR for promoting health literacy in patients with persistent pain [22–24], finding acceptable feasibility from the professional perspective but a need for further development. Gao et al [25] demonstrated improvements in infectious disease–specific health literacy among nursing students following virtual simulation training. Tay et al [26] found that VR enhanced health literacy–related outcomes in agitation management training. However, to our knowledge no published studies have examined VR-based training specifically targeting health-literate communication as a professional competency in mental health contexts.

1.2 Study Rationale and Objectives

This pilot study evaluated the feasibility and preliminary outcomes of a VR-based training intervention targeting mental health literacy and health-literate communication among pre- and postgraduate health students. The objectives were to (1) characterize participants’ baseline mental health literacy, stigma, self-efficacy, and social distance; (2) evaluate the VR intervention’s perceived usefulness, ease of use, and attitudinal acceptance using Technology Acceptance Model constructs; (3) assess clinical communication performance using an AI-driven rubric grounded in Jefferson Scale of Empathy, health literacy, and VR-CoDES frameworks; and (4) identify specific competence gaps.

2 Materials and Methods

2.1 Study Design

A single-arm, sequential mixed-methods pilot study was conducted at two Danish educational sites between April and June 2026. The study employed three measurement time points: T0 (pre-intervention questionnaire, completed before the individual VR session), T1 (structured group debrief immediately following the VR session), and T2 (post-intervention questionnaire, completed after the debrief). VR session performance was assessed independently through AI-driven transcript evaluation.

2.2 Participants

Participation required current enrollment or lecturing in a health professional, health sciences or public health program and informed consent. Exclusion criteria were applied on grounds of lack of Danish language skills, but most important, participant wellbeing: where prior experience of discomfort in VR indicated that the immersive format might be unsuitable, participants could decline the VR session or withdraw at any point without giving a reason; 10 participants chose not to complete a VR session (Table 1).

2.3 Intervention

The VR intervention was delivered on Meta Quest 3 standalone headsets (Meta Platforms, Inc., Menlo Park, CA, USA), with participant speech captured through the built-in headset microphone, using the HealthSim platform (CoFlow Visuals ApS, Horsens, Denmark), which integrates immersive VR with AI-

driven conversational avatars. Participants engaged individually in one of two clinical communication scenarios presented in VR:

Scenario A

“It is difficult at work” (“Det er svært på arbejdet”). A workplace stress scenario in which the patient presents with psychosomatic symptoms, emotional distress, and functional impairment. The scenario requires exploration of psychosocial context, empathic response to emotional cues, and health-literate communication about next steps (n = 10 participants).

Scenario B

“Where is my husband?” (“Hvor er min mand?”). A dementia care scenario in which the patient is confused, distressed, and seeking orientation. The scenario requires perspective-taking, emotional validation, and adapted communication for a cognitively impaired individual (n = 7 participants).

Both scenarios were conducted in Danish. Speech-to-text transcription was performed using OpenAI Whisper [27], and the resulting transcripts were evaluated by the AI-driven evaluation system.

2.4 Instruments

2.4.1 T0: Pre-intervention Questionnaire

The T0 instrument combined two validated frameworks. General health literacy was measured using the 12-item HLS19-Q12 health literacy instrument [4], scored on a 4-point scale (Very easy = 4 to Very difficult = 1). Mental health literacy was measured using the Mental Health Literacy Scale (MHLS) [3], comprising disorder recognition (10 items on a 4-point likelihood scale), helpfulness judgment (2 items), clinical knowledge (3 items on cognitive behavioral therapy and confidentiality), self-efficacy (4 items, 5-point agreement scale), attitudes/stigma (9 items, 5-point agreement scale, reverse-scored), and social distance (7 items, 5-point willingness scale). Background variables included age, gender, educational level, prior profession, prior health/social sector experience, and study program.

2.4.2 T1: Structured Group Debrief

Qualitative data were collected through a facilitated group discussion immediately following the VR session. The debrief was conducted at one of the two institutions with eight participants; at the other institution participants left immediately after their session, so a joint debrief was not logistically feasible. Two researchers (KS and THK) prompted participants to reflect on their experiences with the T0 questionnaire, the VR interaction, and comparisons with other training modalities. Statements were recorded as field notes and thematically organized.

2.4.3 T2: Post-intervention Questionnaire

The T2 instrument was based on Technology Acceptance Model (TAM) constructs [28] and comprised three subscales: Perceived Usefulness (PU, 5 items on a 5-point scale from Slet ikke [not at all] = 1 to I

høj grad [to a great extent] = 5), Perceived Ease of Use (PEOU, 4 items on the same scale), and Attitude Toward Use (ATU, 4 items on a 5-point scale from Meget negativ [very negative] = 1 to Meget positiv [very positive] = 5), of which three items concern recommendation to peers, regular use, and curriculum integration and one is a global item on overall attitude toward VR as a training form. T2 also collected prior VR experience, prior OSCE experience, and an evaluation identification field linking to the VR platform.

2.4.4 Rubric-Based Evaluation

The study adopted the Objective Structured Clinical Examinations (OSCE) taxonomy as the organizing framework for AI-driven scoring, applied as a simplified case-session [29] OSCE style assessments represent a standardized form of performance-based assessment used to evaluate clinical competence [30, 31]. Each transcript was evaluated by the HealthSim AI evaluation system using a deterministic prompt. The rubric comprised 14 criteria across two domains:

- Content domain (weighted 53%): perspective exploration (weight 4), everyday impact (3), cue detection (5), emotion identification (5), information needs (4), plain language explanation (5), and next steps (4).
- Health-Literate Communication domain (weighted 47%): patient involvement (3), empathic response (5), space-giving sequence (4), active listening (4), question invitation (3), teach-back (5), and supportive framing (3).

Seven criteria were derived from the Jefferson Scale of Empathy [32] and VR-CoDES [33] (perspective exploration, everyday impact, cue detection, emotion identification, empathic response, space-giving sequence, active listening). Six criteria were derived from health literacy related frameworks focusing on health literacy and information needs, plain language explanation, teach-back, question invitation, next steps, and supportive framing [4, 8, 34]. Patient involvement bridges both traditions. Each criterion was scored 0% to 100%, and domain-level and total scores were computed using the specified weighting. In practice the system returned scores in increments of 10 for 235 of the 238 criterion ratings; three ratings within a single session fell between increments, and the available data do not record whether this reflects a different scoring granularity or a post-hoc adjustment. Recoding those three ratings either to the nearest or to the lower increment leaves that session's rank and all groupings reported below unchanged.

2.5 The Verona Coding Definitions of Emotional Sequences

The evaluation framework draws on the Verona Coding Definitions of Emotional Sequences (VR-CoDES) [33], a validated system for coding emotional sequences in clinical conversations. VR-CoDES distinguishes between patient cues (implicit emotional signals) and concerns (explicit expressions of distress) and classifies clinician responses from providing space (acknowledgement, exploration, normalization) to reducing space (redirection to clinical facts, unsolicited advice, topic change). Research consistently shows that reducing-space responses predominate in clinical practice [35], and that training can shift this pattern [36].

VR-CoDES suits transcript-based evaluation, as the turn-taking structure of VR dialogues makes emotional sequences explicitly identifiable. In this study, its principles informed the criteria for cue detection, empathic response, and space-giving sequences.

2.6 Data Analysis

Descriptive statistics (means, SDs, medians, inter-quartile range (IQR) and ranges) were computed for continuous variables. Internal consistency was assessed using Cronbach α . Between-group comparisons were performed using Mann–Whitney U tests, and cross-instrument associations were explored using Spearman rank correlations. The 95% confidence interval for Cronbach α was obtained with Feldt's method. All calculations were performed in Microsoft Excel with formulas retained in the analysis workbook to allow verification.

2.7 Ethical Considerations

Participation was voluntary with informed consent collected after written and oral introduction of the study. All data were collected anonymously using running identification numbers not linked to participants. The study was conducted in accordance with the Declaration of Helsinki.

3 Results

The group consisted of 27 participants (25 students and 2 professionals) whose prior professional backgrounds were heterogeneous: the most common were nursing ($n = 3$), midwifery ($n = 2$), social and healthcare assistant ($n = 4$), and physiotherapy ($n = 1$), with four participants reporting no prior profession. A substantial minority had backgrounds outside health and social care entirely (educator, bookkeeper, firefighter, and others; $n = 7$). Prior experience in health or social care was rated as experienced ($n = 8$), limited ($n = 10$), or none ($n = 5$), with four providing other descriptions.

Those completing the postintervention evaluation ($n = 16$) had little or no previous experience with VR assessment.

Table 1

Participant flow and instrument completion across the three measurement time points (T0 pre-intervention, T1 structured group debrief, T2 post-intervention). n is the number of participants contributing data at each stage, and denominators differ across analyses: 27 completed T0, 17 a VR session, 8 the debrief and 16 the T2 questionnaire. Counts are derived from the study dataset

Stage	Instrument / activity	n	Notes
Enrollment	Recruitment at two educational institutions	27	25 students, 2 lecturers
T0	Pre-intervention questionnaire (MHLS, HLS19-Q12)	27	Completed by all enrolled
VR session	VR intervention, one scenario per participant	17	Scenario A (workplace stress) n = 10; Scenario B (dementia care) n = 7 10 participants did not complete a VR session (voluntary opt-out, see 2.2)
T1	Structured group debrief	8	Held at one institution only; at the other, participants left immediately after the session
T2	Post-intervention questionnaire (TAM)	16	Not completed: n = 11, of whom 1 had completed a VR session, but skipped T2

3.1 Baseline Mental Health Literacy (T0)

3.1.1 Mental Health Literacy (MHLS)

The 35-item Mental Health Literacy Scale was completed by all 27 baseline participants. The mean total score was 134.5 out of a possible 160 (median 127.5; SD 10.4; range 114–150; IQR 127.5–142.5; corresponding to 84.2% of the maximum attainable score, SD 8.2 percentage points), indicating high, homogeneous mental health literacy consistent with the sample's professional background.

At the subscale level (normalized to mean score per item), participants scored highest on the attitude/stigma items (items 20–35, mean 4.31 per item) and on information-seeking (items 16–19, mean 4.06), and lowest on knowledge of risk and causes (items 9–10, mean 2.93), locating the weakest area in epidemiological knowledge rather than in attitudes. Attitudinal dimensions were captured within the MHLS attitude subscale reported above and are not reported as separate scores.

3.1.2 General Health Literacy (HLS19-Q12)

General health literacy was assessed with the HLS19-Q12 instrument, which showed high internal consistency (Cronbach $\alpha = .88$). Across the 27 participants, the mean index score was 81.2 on the 0–100 scale (proportion of responses rated "easy" or "very easy"; the corresponding share of the scale maximum was 76.6%, SD 11.0 percentage points; median 79.2, IQR 68.8–82.3; range 54.2–97.9), and the dichotomised health-literacy levels distributed as inadequate (n = 3, 11.1%), problematic (n = 9, 33.3%), sufficient (n = 10, 37.0%), and excellent (n = 5, 18.5%). Just over half the sample (15 of 27) thus placed in

the two highest levels, a distribution close to the 17-country average reported in the HLS19-Q12 validation study (Fig. 1).

3.2 Post-intervention Evaluation (T2)

Across the 16 participants who completed the post-intervention evaluation, the VR training was rated positively overall (mean 4.03 out of 5 across all 13 items). Perceived Usefulness scored 4.11 out of 5, and Perceived Ease of Use 4.08, with a clear split between ease of learning the equipment (mean 4.56) and the realism/technical items (avatar realism mean 3.44; technical disruption mean 3.50), which scored lowest and showed the widest spread.

Behavioral Intention, covering recommendation to peers, regular use, and curriculum integration, scored 4.05, with overall attitude toward VR as a training form at 4.19. The pattern indicates that the principal acceptance barrier is realism and immersion rather than basic usability (Fig. 2).

Internal consistency was acceptable for the 12 TAM subscale items, Cronbach $\alpha = .85$ (95% CI .72–.94; $n = 16$); the single overall attitude item was treated as a separate global rating and excluded from the reliability estimate ($\alpha = .87$ when included). Given the sample size, the estimate is imprecise and should be read as indicative rather than definitive.

3.3 Rubric-Based Communication Performance

Fourteen of the 17 sessions scored below the 50% pass threshold, one scored exactly 50%, and two exceeded it (mean 32.5%; range 5.1%–59.5%) (Fig. 3). Content scores (mean 36.7%, SD 18.2) consistently exceeded Communication scores (mean 27.8%, SD 15.5), with a strong positive correlation between domains (Spearman $\rho = .92$, $P < .001$).

The 14 evaluation criteria revealed a striking bifurcation between empathy-derived and health literacy-derived competencies. Empathy and VR-CoDES criteria (Jefferson Scale–derived) scored substantially higher (criterion means 37.1%–59.1%; Fig. 4 and Table 2); health literacy criteria scored markedly lower (criterion means 0.6%–37.6%).

Floor effects were pronounced but not absolute: teach-back was employed in only 1 of 17 sessions (mean 0.6%), while supportive framing was nonzero in 5 of 17 sessions (mean 5.3%) and exceeded 10% in only one. Information needs assessment and question invitation likewise scored at or near zero in the majority of sessions.

Table 2

Rubric criteria: descriptive statistics across all 17 VR sessions. Each criterion is listed with its domain (Content or Health-Literate Communication) and its weight in the total score. Mean and SD are expressed as percentages of the maximum attainable criterion score. Floor indicates the number of sessions in which the criterion scored zero

Criterion	Domain	Weight	Mean (%)	SD	Floor
Cue detection	Content	5	59.1	23.2	0/17
Perspective exploration	Content	4	50.0	21.2	0/17
Everyday impact	Content	3	47.4	22.6	0/17
Empathic response	Communication	5	48.2	24.6	1/17
Space-giving sequence	Communication	4	43.8	25.8	1/17
Patient involvement	Communication	3	37.1	20.2	0/17
Active listening	Communication	4	42.4	21.7	0/17
Next steps	Content	4	28.2	23.5	3/17
Emotion identification	Content	5	37.6	23.6	1/17
Plain language explanation	Content	5	21.2	20.9	6/17
Question invitation	Communication	3	11.8	16.3	8/17
Information needs	Content	4	14.1	16.6	7/17
Teach-back	Communication	5	0.6	2.4	16/17
Supportive framing	Communication	3	5.3	12.3	12/17
<i>Note 1: Scenario A (workplace stress, n = 10) yielded slightly higher mean total scores than Scenario B (dementia care, n = 7): 33.6% versus 29.7%. The difference was not statistically significant.</i>					

3.4 Cross-instrument Associations

None of the Spearman correlations between T0 subscales and rubric total scores were statistically significant, consistent with the small sample. The largest positive association was between MHLS total score and rubric total ($\rho = .21$, $P = .43$, $n = 17$). Attitude toward VR showed a small, nonsignificant positive association with rubric performance ($\rho = .10$, $P = .70$, $n = 16$); the weak negative trend observed in the preliminary subsample ($n = 8$) did not replicate once all 16 evaluations were included. No significant differences were observed between teaching groups on any measure.

3.5 Qualitative Findings (T1)

The structured debrief produced insights across three domains, integrated with the quantitative acceptance data in Table 3. Regarding the T0 instrument, participants found the general health literacy

items accessible and reflective. The mental health literacy items were perceived as substantially more difficult:

As one participant noted, *"It was difficult to understand what the individual questions were actually measuring."* Another observed that *"Some questions and response categories are reversed or contain negations, so if you read too quickly, you answer incorrectly"*. As such participants noted high linguistic complexity, unfamiliar diagnostic terms, negation in response categories, and ambiguity about what was being measured.

Regarding the VR experience, participants described the interaction as surprisingly human. Participants repeatedly expressed surprise at the interaction quality. One participant remarked, *"It felt more human than I had expected,"* while another explained, *"You can abstract from the avatar and focus on the conversation."*

VR was consistently framed as superior to traditional role-play. Traditional role-play was described in markedly less favorable terms. One participant summarized it simply as *"Role-play is something you just endure,"* while another commented, *"It feels awkward when it is someone you know."* Specific VR limitations included absence of facial expression and nonverbal feedback, conversational latency, sensitivity to ambient noise, and interference from active listening vocalisations.

Regarding future applications, participants endorsed VR integration in both teaching and examination settings, emphasising the standardization advantage for fair assessment. They suggested collaborative VR scenarios, AI-generated cases adapted to specific disciplines, and the possibility of incorporating physiological monitoring features.

Table 3
Joint display integrating quantitative Technology Acceptance Model outcomes with representative qualitative quotations from the structured debrief

<i>Quantitative finding</i>	<i>Representative participant quotation</i>
Easy to learn (mean 4.56)	<i>"It felt more human than I had expected."</i>
Avatar realism (mean 3.44)	<i>"You miss facial expressions."</i>
Technical issues did not disturb concentration (mean 3.50)	<i>"The conversation is a little jerky."</i>
Recommend for teaching (mean 4.19)	<i>"We are very open to integrating it into teaching."</i>
Relevant as a supplement to traditional teaching (mean 4.06)	<i>"Role-play is something you just endure."</i>

3.6 Speech-to-Text Artifacts

Several of the reports contained non-Danish utterances flagged as critical issues (Japanese, Korean, Polish, German, and nonsense strings). These are established artifacts of Whisper-based transcription

for Danish speech and are not attributable to participant behavior. Because the affected passages were flagged rather than scored as participant content, they are unlikely to have materially affected the rubric-based assessment.

4 Discussion

4.1 Principal Findings

This pilot study demonstrates a functional end-to-end workflow for VR-based mental health literacy training with AI-driven rubric assessment, from baseline assessment through immersive simulation to automated performance scoring and structured debrief. The central finding is a stark asymmetry between anticipated knowledge and empathic and health literacy communication competencies performed in practice.

Participants demonstrated moderate capacity for cue detection, perspective exploration, and empathic response, while health literacy-specific techniques were marginal or near-absent.

This pattern supports the intervention's central premise: graduate students bring relational and empathic dispositions from prior professional training and score high on knowledge-based measures of mental health literacy (T0 MHLS mean 134.5/160), yet this knowledge does not spontaneously translate into health-literate communicative practice. Ceiling-level T0 scores against floor-level rubric performance constitute a knowledge-to-practice gap not attributable to insufficient knowledge.

The near-complete floor effect on teach-back is particularly notable, as it is among the best-evidenced strategies for ensuring patient understanding [8]. The rarity of these techniques despite demonstrable baseline knowledge provides empirical justification for structured didactic bridging as a prerequisite component of the full intervention. The gap may equally reflect that these techniques have simply not been taught: health-literate communication is rarely an explicit curriculum component, and participants may never have been trained in teach-back or in assessing information needs.

Critically, didactic instruction prior to VR sessions is not a confound but the core mechanism of change: the intervention activates existing knowledge into practiced communicative behavior.

The scenario-level asymmetry (Scenario A 33.6% vs Scenario B 29.7%) merits separate attention, though the subsamples were small. Two non-exclusive explanations are plausible. First, Scenario B is clinically less familiar for students whose exposure is weighted toward acute mental-health and somatic encounters; diagnostic cognitive load may have crowded out communicative technique. Second, the dementia scenario's narrative structure (a confused, searching patient) provides fewer overt verbal cues than the workplace-stress scenario, reducing opportunities for cue-detection scoring. Both interpretations point to scenario complexity calibration as a design parameter for the full study rather than a confound to be eliminated.

Attitude toward VR was unrelated to rubric performance, providing no evidence that enthusiasm for the medium either supported or detracted from communicative task performance. Explicit instructional framing remains warranted in a full study, orienting learners to the simulation as a deliberate-practice environment rather than a demonstration of immersive technology, but this recommendation now rests on pedagogical principles rather than on an observed performance-goal distinction.

4.2 Comparison With Prior Work

The central finding, high baseline knowledge coexisting with floor-level performance on health-literate communication techniques, reflects the well-documented knowledge-to-practice gap in health professions education. Issenberg et al [10] identified deliberate practice with feedback as the active ingredient of high-fidelity simulation precisely because exposure to didactic content does not reliably produce behavioral change at the bedside. Talevski et al's systematic review [8] similarly found that teach-back, despite robust evidence of effectiveness, is inconsistently and incompletely implemented in routine practice. Moreover, plain language is essential for supporting patient understanding and informed and shared decision-making; however, its use in medical encounters remains inconsistent, and the lack of standardized measurement and feedback mechanisms represents an important gap in clinical communication practice [37]. The present pilot's floor effects, despite high baseline general health literacy, are therefore not anomalous but exemplary: declarative knowledge of communicative techniques does not transfer to procedural competence without structured rehearsal.

Virtual reality (VR) addresses the limitations of lectures, role-play and simulated patients described in the Introduction, as it provides immersive, standardized, and repeatable clinical encounters in a psychologically safe environment [13, 14]. Saab et al. [15] demonstrated high usability (78.5/100) and feasibility (86.4/100) for VR communication training across 268 nursing and medical students in five European countries. Chen [16] found that AI-enhanced VR simulation significantly improved empathic communication in nursing students (Cohen $d = 0.96$ at T1), with strong concordance between AI and educator ratings (intraclass correlation coefficient = 0.81). Birdsall et al [17] reported significant knowledge gains and increased confidence following VR patient simulation in maternal care communication.

Several studies have specifically examined VR for mental health communication. Kim et al [38] found that VR simulation improved nursing students' psychiatric communication competence. Furuheim et al [39] reported that students perceived VR simulation in mental health practice as a meaningful learning activity. Pina-Thomas et al [40] found comparable empathy outcomes but higher engagement for VR versus traditional hearing voices simulations. Mattsson et al [41] described empathy in VR simulation as authentic and emotionally engaging. VR thus supports affective learning in mental health education, but its intersection with health literacy competencies has received minimal attention.

The Technology Acceptance Model results are comparable to those reported by Saab et al [15] and Birdsall et al [17]. The preference for VR over role-play echoes Lauritzen et al [42], who found that VR supported active exploration and translation of theoretical knowledge into practice.

The scenario-level difference mirrors findings by Weisshardt et al [43], who noted that immersive patient-perspective VR experiences in neurological conditions require adaptation of complexity to learner level. The AI evaluation approach shares conceptual ground with Mergen et al [19], who demonstrated AI-driven virtual patient platforms for clinical decision-making training. The application of VR-CoDES principles to automated transcript evaluation represents a novel methodological contribution. Ortwein et al [36] previously applied VR-CoDES to code medical students' interactions with virtual patients but relied on human coders. The present study's AI-driven approach operationalises VR-CoDES constructs within an automated scoring pipeline, enabling scalable evaluation.

Two methodological considerations follow. First, automated scoring inherits the validity of its source coding scheme: VR-CoDES was developed and validated for human coding of cues and concerns in real consultations [33, 35], and its application to VR-mediated encounters with synthetic patients requires explicit validation against human raters before scores can be treated as substitutable. Chen [16] reports strong AI-educator concordance in a related context, but inter-rater reliability between human VR-CoDES coders and the present pipeline remains to be established and is a defined deliverable of the powered study. Second, transcript-based scoring is bounded by transcription fidelity: the Whisper artifacts observed here reduce the lexical surface available to a language-model evaluator.

Scalability of AI evaluation is therefore not unconditional: it depends on language coverage of the underlying speech-recognition layer, which remains weaker for smaller languages such as Danish than for English.

4.3 Limitations

The study has several limitations inherent to its pilot design. The sample of 17 rubric-evaluated sessions (27 at T0) precludes inferential generalization; correlation estimates are presented to inform hypothesis generation, as none of the correlations or group comparisons were statistically significant, and effect size estimates would be unreliable. The qualitative component is similarly limited: the structured debrief was held at only one of the two institutions and with eight participants, so the T1 findings reflect a subset of the sample rather than the cohort as a whole.

The single-arm design without pre-post rubric comparison means that the rubric scores represent a cross-sectional snapshot of performance without an intervention effect estimate. The ceiling effects observed in T0 general health literacy and self-efficacy subscales limit the discriminative power of these measures in a health professional sample.

The speech-to-text artifacts produced by Whisper transcription for Danish speech represent a systematic technical limitation. Non-Danish utterances were present across the transcripts but were correctly flagged in the evaluation reports and therefore treated as transcription issues rather than participant content; they may have introduced unidentified elements in the transcripts, although probably not to an extent that influenced the criterion scores.

4.4 Future Work

The pilot suggests several concrete directions for the powered study and beyond. Most immediately, the powered pre–post design should pair structured didactic priming, microlearning modules targeting health literacy such as teach-back, supportive framing, and information-needs assessment, with repeated VR-based rubric measurement to estimate intervention effects on the competencies that scored at floor. Sample size estimation should target a clinically meaningful improvement on these health literacy criteria rather than on overall totals, given the bifurcated profile observed here. A two-arm design with VR-only versus VR-plus-didactic-bridge would isolate the contribution of explicit instructional support.

Several methodological enhancements follow directly from pilot observations. Language filtering at the transcription layer is a precondition for valid Content scoring. Inter-rater reliability between human coders and the AI pipeline should be established, extending the human-coded VR-CoDES protocol of Ortwein et al [36] to AI-evaluated transcripts. Prompt-stability and hallucination audits of the language-model evaluator are also indicated, given the documented sensitivity of large language models to small input perturbations.

Debrief suggestions point to platform extensions worth piloting separately: collaborative multi-learner scenarios for peer observation and shared reflection; AI-generated cases adapted to specific clinical disciplines and patient populations, including further mental health scenarios such as depression and anxiety, and conditions that impede communication itself, such as aphasia or hearing impairment; and integration of physiological monitoring (heart rate variability, electrodermal activity) as objective indicators of engagement and cognitive load.

Finally, cross-linguistic and cross-cultural generalization remains an open question: the pilot was conducted in Danish, and replication in English-language and multi-site settings would test both the MHL instrument's translation properties and the speech-recognition pipeline's language coverage. It would also clarify whether the observed bifurcation is a feature of the intervention or partly an artifact of the Danish training context.

5 Conclusions

This pilot study explored and established the feasibility of a VR-based training platform for mental health literacy and health-literate communication, combining immersive AI-driven simulation with automated, OSCE-style rubric evaluation grounded in validated health literate-communication frameworks. The near absence of use of health literacy techniques, contrasted with moderate empathic and relational performance, provides a robust baseline characterization of the competence gap the intervention targets. A powered pre-post study with structured didactic priming and repeated rubric measurement is the indicated next step.

Declarations

Acknowledgments

The authors thank the participating students and lecturers for their time and engagement, the Department of Health Science and Technology at Aalborg University for facilitating recruitment and access to facilities, and UCN (University College of Northern Denmark) for supporting participant recruitment and data collection at the second study site.

Ethics Approval

The study was conducted in accordance with the Declaration of Helsinki. Under Danish law, studies that collect anonymous questionnaire, interview, and educational evaluation data from adult volunteers, and that involve no biological material and no clinical intervention, are not subject to mandatory review by the regional committees on health research ethics. Formal ethics approval was therefore not required for this project.

Consent to Participate

All participants received written and oral information about the study and gave informed consent prior to participation. Participation was voluntary, and participants could withdraw at any time without consequence. All data were collected anonymously using running identification numbers that cannot be linked to individual participants.

Consent to Publish

No individually identifiable data, images, or audio recordings are reported. All participants consented to the anonymized use and publication of their aggregated data and of anonymized quotations from the group debrief.

Author Contribution

THK and KS conceived and conducted the study with support from BJ and HB from onset to finalization. THK developed the VR intervention and the AI-driven rubric evaluation pipeline and collected the data. KS contributed with the health-literate communication framework and qualitative reports. KS, BJ and HB chose and translated the measurement instruments in the initial study phase and made substantial contributions to the statistical analysis. KS and BJ recruited participants for the study. THK and KS drafted the first versions of the manuscript with feedback from BJ and HB. All authors read, revised, and approved the final manuscript.

Competing Interests

THK is Director, Business & Innovation at CoFlow Visuals ApS, which develops the HealthSim platform used in this study. KS is the Founder of the Global Health Literacy Academy and Adjunct Professor at Aalborg University.

Funding

This study was funded by Erhvervsfyrtoårn Life Science (co-financed by the European Union), covering the period January to June 2026. Grant number: RFPRO-23-0061. The funder had no influence on the design, intervention, analysis, or reporting of the project.

Data Availability

Anonymized pilot data are available from the corresponding author upon reasonable request.

Disclosure of Use of AI Tools

Speech-to-text transcription of the VR dialogues was performed with OpenAI Whisper [27] and the transcripts were scored by the AI-driven rubric evaluation system described in 2.4.4. Both are components of the intervention, not writing aids. All AI-generated output was reviewed by the authors, who take full responsibility for the content of the manuscript.

References

1. Jorm AF. Mental health literacy: Empowering the community to take action for better mental health. *Am Psychol*. 2012;67:231–43. <https://doi.org/10.1037/a0025957>.
2. Kutcher S, Wei Y, Coniglio C. Mental Health Literacy: Past, Present, and Future. *Can J Psychiatry*. 2016;61:154–8. <https://doi.org/10.1177/0706743715616609>.
3. O'Connor M, Casey L. The Mental Health Literacy Scale (MHLS): A new scale-based measure of mental health literacy. *Psychiatry Res*. 2015;229:511–6. <https://doi.org/10.1016/j.psychres.2015.05.064>.
4. Sørensen K, Van den Broucke S, Fullam J, et al. Health Literacy and Public Health: A Systematic Review and Integration of Definitions and Models. *BMC Public Health*. 2012;12:80. <https://doi.org/10.1186/1471-2458-12-80>.
5. World Health Organization. World mental health report: transforming mental health for all. Geneva: World Health Organization; 2022.
6. Sørensen K, Pelikan JM, Röthlin F, et al. Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). *Eur J Public Health*. 2015;25:1053–8. <https://doi.org/10.1093/eurpub/ckv043>.
7. Pelikan JM, Link T, Straßmayr C, et al. Measuring comprehensive, general health literacy in the general adult population: the development and validation of the HLS19-Q12 instrument in seventeen countries. *Int J Environ Res Public Health*. 2022;19:14129. <https://doi.org/10.3390/ijerph192114129>.
8. Talevski J, Wong Shee A, Rasmussen B, et al. Teach-back: A systematic review of implementation and impacts. *PLoS ONE*. 2020;15:e0231350. <https://doi.org/10.1371/journal.pone.0231350>.
9. Nutbeam D. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. *Health Promot Int*. 2000;15:259–67.

<https://doi.org/10.1093/heapro/15.3.259>.

10. Barry Issenberg S, Mcgaghie WC, Petrusa ER, et al. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teach*. 2005;27:10–28. <https://doi.org/10.1080/01421590500046924>.
11. Lateef F. Simulation-based learning: Just like the real thing. *J Emerg Trauma Shock*. 2010;3:348. <https://doi.org/10.4103/0974-2700.70743>.
12. Kron FW, Fettes MD, Scerbo MW, et al. Using a computer simulation for teaching communication skills: A blinded multisite mixed methods randomized controlled trial. *Patient Educ Couns*. 2017;100:748–59. <https://doi.org/10.1016/j.pec.2016.10.024>.
13. Kyaw BM, Saxena N, Posadzki P, et al. Virtual Reality for Health Professions Education: Systematic Review and Meta-Analysis by the Digital Health Education Collaboration. *J Med Internet Res*. 2019;21:e12959. <https://doi.org/10.2196/12959>.
14. Dyer E, Swartzlander BJ, Gugliucci MR. Using virtual reality in medical education to teach empathy. *J Med Libr Assoc*. 2018;106. <https://doi.org/10.5195/jmla.2018.518>.
15. Saab MM, O'Donovan M, Koç İ, et al. Evaluating virtual reality communication training in nursing and medical education: A multi-country cross-sectional study. *Nurse Educ Today*. 2026;160:106985. <https://doi.org/10.1016/j.nedt.2026.106985>.
16. Chen PJ. AI-enhanced virtual reality simulation for nursing students' empathy: Automated scoring and inter-rater reliability in a randomised controlled study. *Nurse Educ Today*. 2026;159:106968. <https://doi.org/10.1016/j.nedt.2025.106968>.
17. Birdsall S, McAlpin E, Grandhi U. In Situ Evaluation of Virtual Reality Simulation Usability and Learning in Nursing. *J Nurs Educ*. 2026;1–8. <https://doi.org/10.3928/01484834-20260112-01>.
18. Chan MMK, Sze Ki Cheung D, Ho KHM, et al. AI-Powered Virtual Reality Simulation for Clinical Handover Training: A Development Framework. *Nurse Educ*. 2026;51:E24–8. <https://doi.org/10.1097/NNE.0000000000002018>.
19. Mergen M, Junga A, Risse B, et al. Immersive training of clinical decision making with AI driven virtual patients - a new VR platform called medical tr.AI.ning. *GMS J Med Educ*. 2023;40:Doc18. <https://doi.org/10.3205/zma001600>.
20. Hamilton A. Artificial Intelligence and Healthcare Simulation: The Shifting Landscape of Medical Education. *Cureus*. 2024;16:e59747. <https://doi.org/10.7759/cureus.59747>.
21. Liaw SY, Rusli KDB, Tan JZ, et al. Artificial intelligence-enabled virtual reality simulation for clinical deterioration training: An effectiveness-implementation hybrid study. *Nurse Educ Pract*. 2025;87:104462. <https://doi.org/10.1016/j.nepr.2025.104462>.
22. Skidmore N, Ryan CG, Mankelow J, Martin D. Acceptability and feasibility of virtual reality to promote health literacy in primary care from the health professional's view: A qualitative study. *Patient Educ Couns*. 2024;123:108179. <https://doi.org/10.1016/j.pec.2024.108179>.
23. Skidmore N, Ryan CG, Mankelow J, Martin D. Expectations for Adopting Virtual Reality to Promote Health Literacy in Patients With Persistent Pain: Qualitative Analysis of UK-Based Physiotherapists.

- Pain Res Manag. 2025;2025:5547227. <https://doi.org/10.1155/prm/5547227>.
24. Skidmore N, Ryan C, Mankelow J, et al. Exploring the potential of virtual reality for the self-management of chronic pain: A scoping review of its use to address health literacy. *Musculoskelet Sci Pract*. 2024;72:102962. <https://doi.org/10.1016/j.msksp.2024.102962>.
 25. Gao J, Lu X, Xu R, et al. The effectiveness of virtual simulation in improving knowledge, skill, affective outcomes, and satisfaction among nursing students during infectious disease nursing: A single-group pretest-posttest design. *Nurse Educ Today*. 2026;161:107018. <https://doi.org/10.1016/j.nedt.2026.107018>.
 26. Tay GWN, Tong MM, Yap J, et al. Virtual reality for experiential learning: enhancing agitation management skills, confidence, and empathy in healthcare students. *Med Educ Online*. 2025;30:2542809. <https://doi.org/10.1080/10872981.2025.2542809>.
 27. Radford A, Kim JW, Xu T et al. (2022) Robust speech recognition via large-scale weak supervision.
 28. Davis FD. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Q*. 1989;13:319–40. <https://doi.org/10.2307/249008>.
 29. Vandamme K, Wulfman C. Formative training for objective structured clinical examinations (OSCEs). In: Rahman MT, Kassim NLA, editors. *Handbook of dental education technology*. Singapore: Springer Nature Singapore; 2025. pp. 1–26.
 30. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. *BMJ*. 1975;1:447–51. <https://doi.org/10.1136/bmj.1.5955.447>.
 31. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Med Educ*. 1979;13:39–54. <https://doi.org/10.1111/j.1365-2923.1979.tb00918.x>.
 32. Hojat M, Gonnella JS, Nasca TJ, et al. Physician Empathy: Definition, Components, Measurement, and Relationship to Gender and Specialty. *Am J Psychiatry*. 2002;159:1563–9. <https://doi.org/10.1176/appi.ajp.159.9.1563>.
 33. Del Piccolo L, De Haes H, Heaven C, et al. Development of the Verona coding definitions of emotional sequences to code health providers' responses (VR-CoDES-P) to patient cues and concerns. *Patient Educ Couns*. 2011;82:149–55. <https://doi.org/10.1016/j.pec.2010.02.024>.
 34. Peter M, Maddocks S, Tang C, Camp PG. Simplicity: using the power of plain language to encourage patient-centered communication. *Phys Ther*. 2024;104:pzad103. <https://doi.org/10.1093/ptj/pzad103>.
 35. Zimmermann C, Del Piccolo L, Finset A. Cues and concerns by patients in medical consultations: A literature review. *Psychol Bull*. 2007;133:438–63. <https://doi.org/10.1037/0033-2909.133.3.438>.
 36. Ortwein H, Benz A, Carl P, et al. Applying the Verona coding definitions of emotional sequences (VR-CoDES) to code medical students' written responses to written case scenarios: Some methodological and practical considerations. *Patient Educ Couns*. 2017;100:305–12. <https://doi.org/10.1016/j.pec.2016.08.026>.

37. Yen RW, Singh A, Chen D, et al. Measuring the use of spoken plain language by clinicians in healthcare encounters: a scoping review. *Patient Educ Couns*. 2026;147:109531. <https://doi.org/10.1016/j.pec.2026.109531>.
38. Kim GM, Lim JY, Kim EJ, Yeom M. Impact of Virtual Reality Mental Health Nursing Simulation on Nursing Students' Competence. *J Multidiscip Healthc*. 2024;17:191–202. <https://doi.org/10.2147/JMDH.S435986>.
39. Furuheim ACH, Lindenskov L. Virtual reality simulation as a learning activity for nursing students in mental health practice. *Nurse Educ Pract*. 2025;83:104272. <https://doi.org/10.1016/j.nepr.2025.104272>.
40. Pina-Thomas D, Zohn JH, Phillips L. A Comparison of Virtual Reality and Traditional Audio-Recorded Hearing Voices Simulations: Their Impact on Empathy in Health Care Students. *Nurs Educ Perspect*. 2025;46:116–8. <https://doi.org/10.1097/01.NEP.0000000000001271>.
41. Mattsson K, Haavisto E, Jumisko-Pyykkö S, Koivisto JM. Nursing Students' Experiences of Empathy in a Virtual Reality Simulation Game: A Descriptive Qualitative Study. *Comput Inf Nurs*. 2024;42:537–45. <https://doi.org/10.1097/CIN.0000000000001132>.
42. Lauritzen C, Reedtz C, Kristensen KB, et al. Development of Therapeutic Competencies in Health Care Students: Qualitative Focus Group Study Using 360-Degree Video and Virtual Reality Technology. *JMIR Med Educ*. 2026;12:e75776. <https://doi.org/10.2196/75776>.
43. Weisshardt I, Reyes Acosta C, Chauhan T, et al. ENGAGE: analyzing the value of virtual reality in a patient-centric immersive learning program in myasthenia gravis for healthcare professionals. *Front Neurol*. 2026;16:1655351. <https://doi.org/10.3389/fneur.2025.1655351>.

Figures

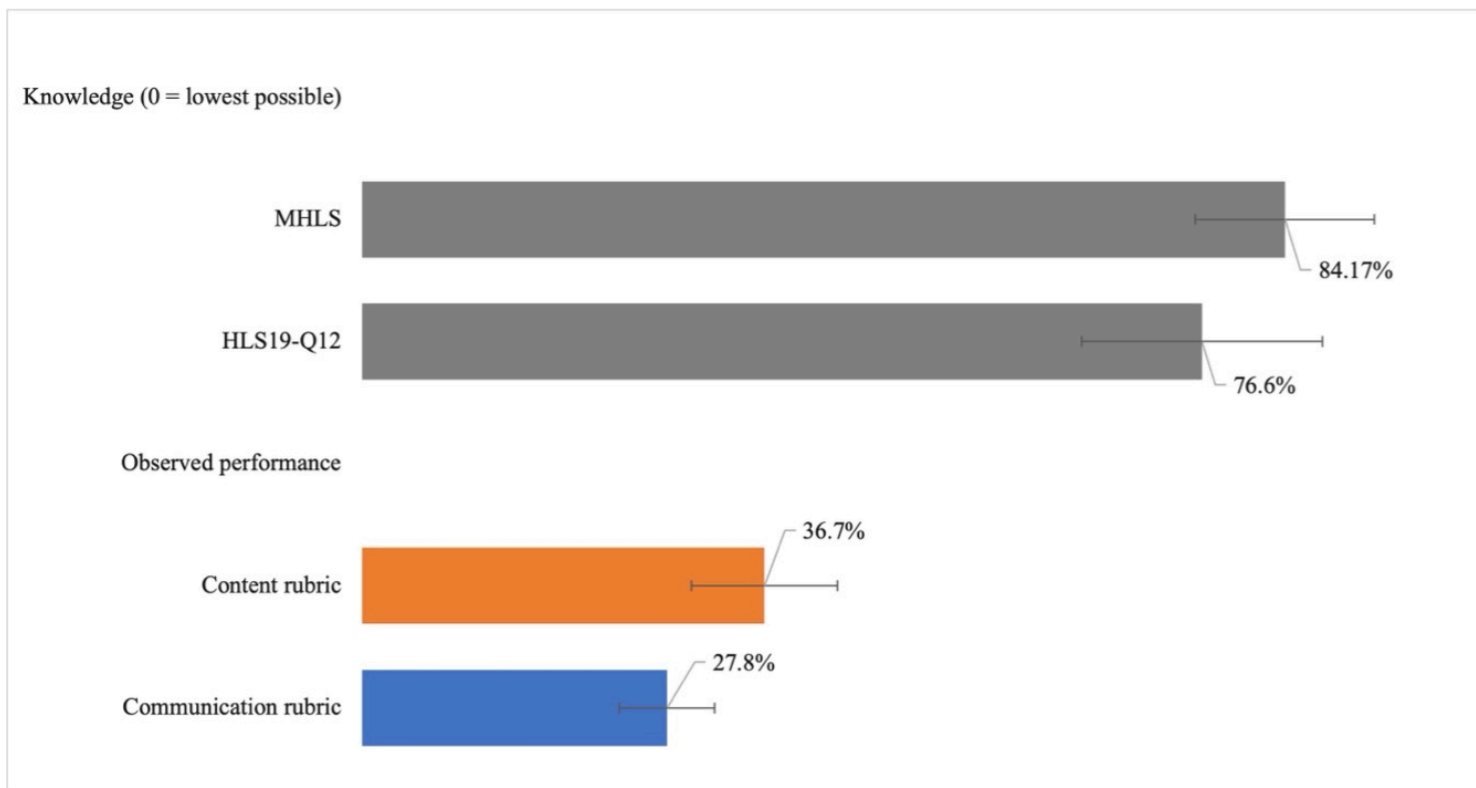


Figure 1

Baseline knowledge contrasted with observed communication performance. Mental health literacy (MHLS) and general health literacy (HLS19-Q12) are expressed as percentages of their maximum attainable scores (n = 27) and compared with AI-derived rubric performance (n = 17 VR sessions); error bars show ± 1 SD

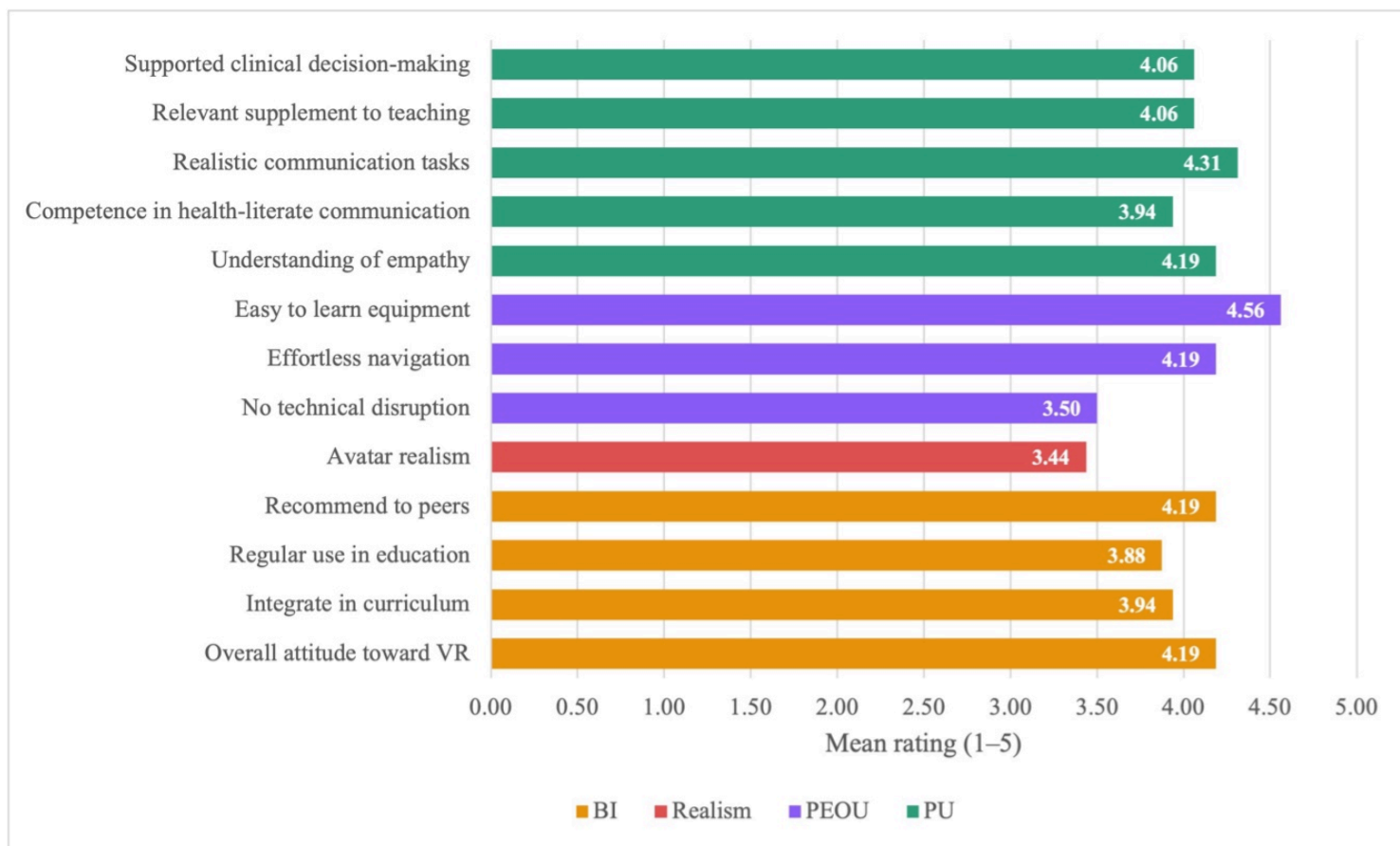


Figure 2

Technology Acceptance Model (TAM) item-level ratings post-VR (N = 16). Bars show item means (1–5 Likert scale) grouped by TAM construct (PU = perceived usefulness; PEOU = perceived ease of use; BI = behavioral intention; the global attitude item is shown separately); error bars show ± 1 SD

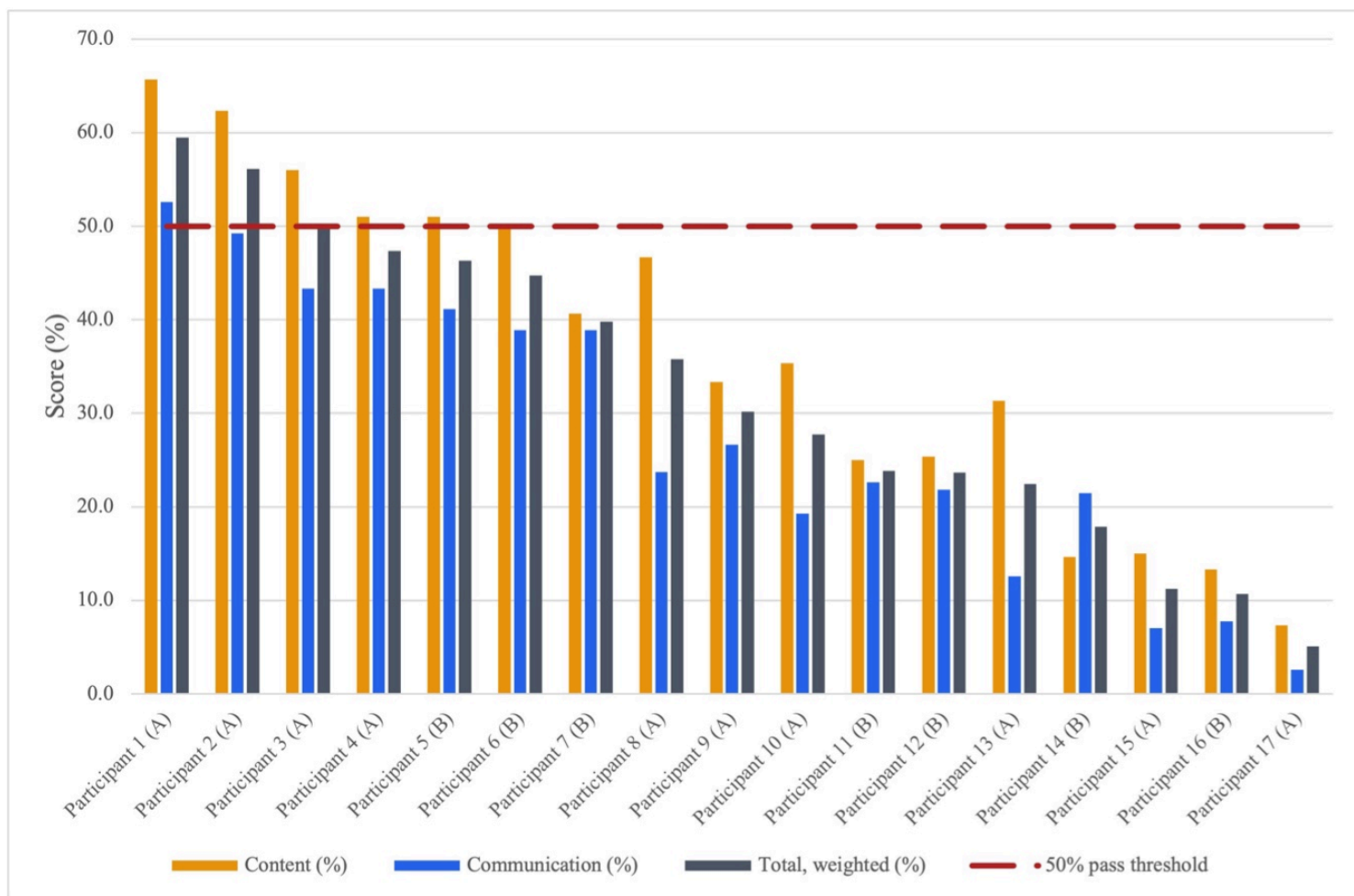


Figure 3

Rubric Content, Communication and weighted total scores by participant (n = 17), ordered by total score. Case assignment is given in brackets (A = workplace stress; B = dementia). The dashed line marks the 50% pass threshold

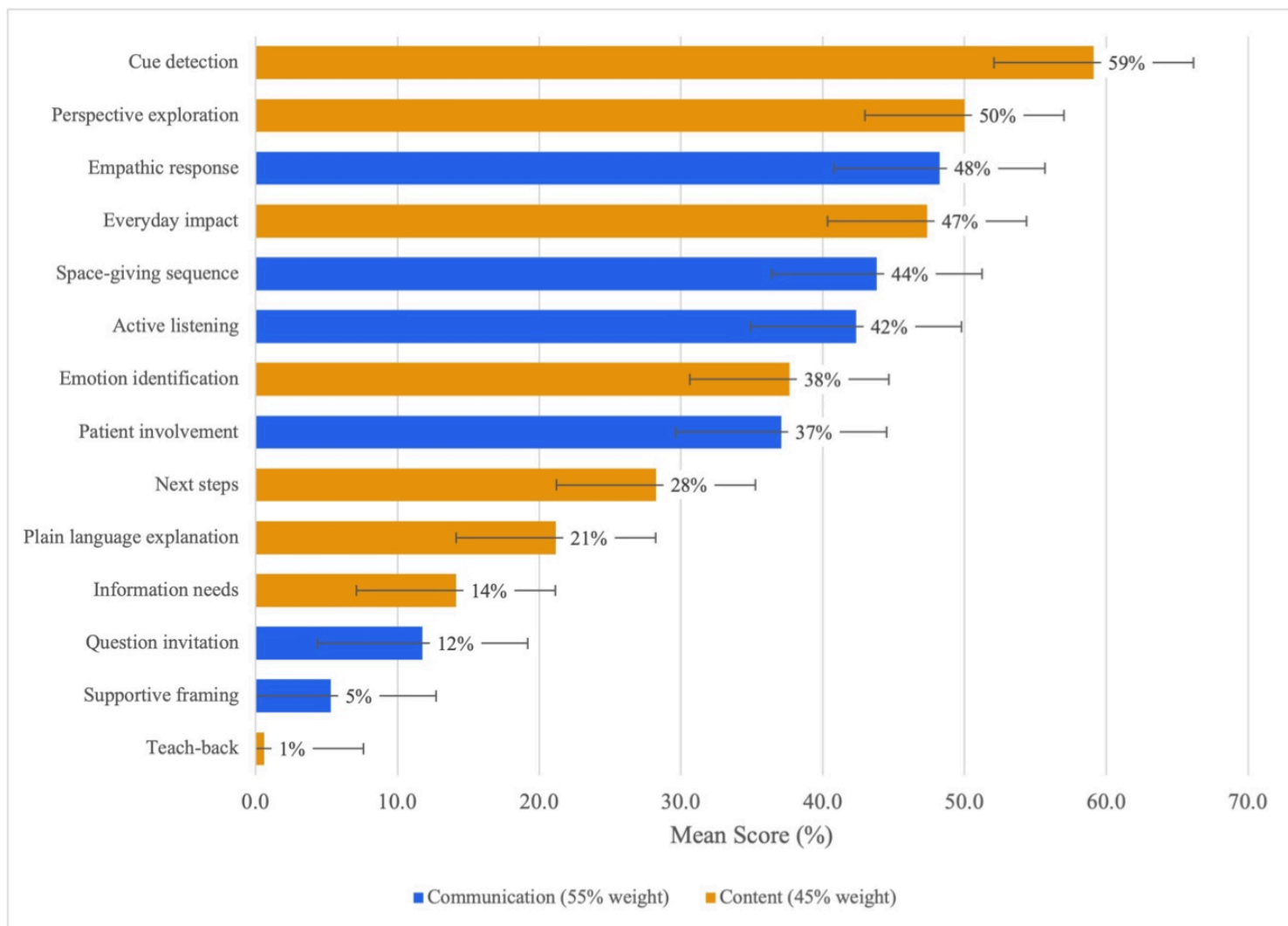


Figure 4

Rubric criterion scores across all 17 VR sessions, ordered by mean score and color-coded by domain (orange = Content; blue = Communication); error bars show ± 1 SD