

Beyond Reaction and Learning: Evaluating XR Training Against the Actual Demands of Challenging Environments

Raimund Schatz^{1,*}, Helmut Schrom-Feiertag¹, Lea Skorin-Kapov², Jan-Niklas Voigt-Antons³ and Leif Oppermann⁴

¹*AIT Austrian Institute of Technology, Giefinggasse 4, 1210 Vienna, Austria*

²*University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia*

³*Hamm-Lippstadt University of Applied Sciences, Lippstadt, Germany*

⁴*Fraunhofer FIT, Sankt Augustin, Germany*

Abstract

Extended Reality (XR) offers opportunities to prepare professionals for challenging environments (CEs) characterized by stress, uncertainty and high consequences of error. These characteristics require evaluation constructs and criteria that capture how trainees experience such demands and acquire the capabilities to handle them. Within the field of XR for challenging environments (XR4CE), we argue for a dedicated training evaluation framework that builds on the Cognitive Affective Model of Immersive Learning (CAMIL) and Kirkpatrick's evaluation model. It connects CE demands with constructs for experience, task-related processes and training outcomes to assess learning, competence acquisition and transfer alongside the training experience. Demand profiles describe the conditions and support available in the target task, training and assessment. Validating the framework requires evidence from a broad range of CE studies to establish its coverage, the suitability of selected measures and its practical value across settings. We therefore invite cross-lab collaboration to test the framework across settings and build cumulative evidence about how XR training prepares professionals for CE demands.

Keywords

XR training, evaluation, challenging environments, transfer of training, demand profile

1. A key challenge for XR trainings in challenging environments: generating cumulative scientific evidence

Extended Reality (XR) offers opportunities to train professionals for challenging environments (CEs), ranging from emergency response and healthcare to aviation and industrial operations. XR for challenging environments (XR4CE) denotes the broader research and application area; this paper focuses on training evaluation within that area. Effective action in CEs may require handling uncertainty, time pressure and unexpected events while avoiding errors with serious consequences. Evaluation therefore needs to establish what trainees learn and under which conditions they can apply it. A positive training experience or successful execution of a familiar procedure addresses only part of this question.

Existing evidence illustrates the difficulty of drawing broader conclusions. Across 330 studies of virtual training, Strojny and Dużmańska-Misiarczyk found that evaluation concentrated on Kirkpatrick levels I (Reaction) and II (Learning); 44 studies (13%) assessed level III (Behaviour), and none assessed level IV (Results) [1]. While this review extends beyond CEs, a review of XR training for mass casualty incidents similarly identified heterogeneous outcomes and measurement tools across 18 studies, only four of which included follow-up [2].

Different instruments and study designs can answer complementary questions. Cumulative evidence nevertheless requires a shared basis for identifying relevant constructs and interpreting findings across conditions. We argue that the specific characteristics of CEs call for a dedicated framework connecting these demands with constructs, measures and criteria for training experience and impact. Its validation requires contributions from many studies across CE tasks and populations. A common evaluation

XR4CE 2026: 2nd International Workshop on XR for Challenging Environments, November 3, 2026, Valencia, Spain

*Corresponding author.

✉ raimund.schatz@ait.ac.at (R. Schatz)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

basis could, in turn, help these studies contribute to a broader understanding of learning, competence acquisition and transfer under CE conditions.

2. Motivation: why XR4CE needs a dedicated evaluation framework

We select two complementary foundations because evaluation needs to address both the training experience and its learning and operational outcomes. The Cognitive Affective Model of Immersive Learning (CAMIL) relates properties of immersive virtual reality to psychological experience and learning [3], informing the choice of experience and learning-process constructs. Kirkpatrick's model distinguishes reaction, learning, behaviour and results [4], helping evaluators distinguish the kinds of outcomes their studies can address. These are selected starting points, not an exhaustive account of relevant approaches. The proposed XR4CE training evaluation framework builds on them by linking construct selection and outcome interpretation to explicit CE demands: exposure to stressors calls for attention to appraisal and regulation, uncertainty to situation awareness and decision-making, and interdependence to coordination and team adaptation. A dedicated framework should specify when these constructs matter, how they can be assessed, and how they inform conclusions about learning and operational capability.

Table 1 proposes six candidate demands as a synthesis informed by selected CE training literature. The selection draws on stress in firefighter training [5], uncertainty, timely decisions and coordination in mass casualty response [2], and hazardous work, changing conditions and safety decisions in industrial training [6]. We distinguish these demands because each can change what counts as adequate performance or which processes need assessment. The list is provisional and open to extension; it is not an exhaustive or validated taxonomy.

The demands can overlap and occur together: time pressure, for example, can constrain how a team resolves conflicting information. They are not individually exclusive to CEs. The framework should help evaluators select constructs and criteria for the demands of their target task. Stress-inducing conditions, such as threat cues, must also be distinguished from trainees' responses. A demand profile records the conditions provided; measures of appraisal, experienced stress or physiology capture responses to those conditions. Both can be assessed in the same study and related to each other. This distinction matters because the same conditions may elicit different responses across trainees.

This approach builds on work that already connects evaluation to application requirements. Harris et al. emphasize task-relevant psychological, affective and ergonomic fidelity [7]. Cordeiro et al. provide preliminary guidelines for evaluating immersive safety training, covering scenario design, recruitment, equipment and feasibility; the paper identifies empirical evaluation of the guidelines as future work [6]. An earlier Kirkpatrick-based proposal by Lutz et al. addresses industrial onboarding; its practical evaluation was ongoing at the time of publication [8]. More recently, Zamora Alvarez et al. distinguish transfer distance, claim reach, durability, fidelity and methodological strength, including changes between training and assessment [9]. The XR4CE training evaluation framework draws on these perspectives to make CE-specific choices of criteria and their interpretation explicit and usable.

3. Our position: evaluation must be rooted in the demands posed by CEs

Challenging environments are demanding environments. Consequently, our central position is that *a human performance claim can only be interpreted and validated when knowing the actual demands under which it was obtained*. Correct action under routine conditions and effective action under uncertainty support different conclusions. We use *demand-referenced evaluation* to describe this connection between a criterion, the conditions under which it was met and the claim it supports. The analogy to criterion-referenced assessment [10] serves to clarify the term. The rationale also connects to transfer-appropriate processing [11], argument-based validation of score interpretations [12], and the role of context in quality-of-experience research [13].

Table 1

Six candidate demands of challenging environments and illustrative evaluation questions. The categories can overlap and are open to extension.

CE Demand	Related Evaluation Question
Threat and other stress-inducing conditions	How do trainees respond to stressors, and can they maintain or regain adequate performance?
Uncertainty, incomplete or conflicting information	Are relevant cues detected and decisions taken appropriately despite uncertainty?
Time pressure and dynamics	Are actions timely and consistent with task-specific quality and safety requirements?
Unexpected events and changed conditions	Is a learned routine sensibly adapted once its preconditions no longer hold?
High consequences and low error tolerance	Are critical errors avoided, detected and managed, including justified decisions not to start or to stop an unsafe action?
Interdependence and distributed information	Do shared situation awareness and coordination remain effective under disruption?

A *demand profile* concisely records relevant demands and available information or support, separately for the target task, training and assessment. These profiles can differ for justified reasons: training may introduce instructional prompts, and preparation for hazardous work can take place without exposing trainees to the actual danger. Reporting such differences helps readers judge the scope of the evidence. The profile provides the context for selecting and interpreting evaluation constructs. Directly comparing or pooling scores across populations or instruments requires evidence that the scores have a sufficiently comparable meaning and scale. Studies can also contribute through qualitative synthesis or analysis of contextual differences without treating their scores as interchangeable.

Two evaluation perspectives guide this interpretation. *Experience* covers how training is experienced and judged, including perceived workload and experienced disruption from technical problems. Objective records of latency or tracking failures describe technical conditions and can help explain these experiences. *Impact* covers learning, competence acquisition, retention, transfer and operational outcomes. The latter concern effects in actual work or deployment, such as safety-critical errors or the quality of task execution. Attributing outcomes to training requires an appropriate study design. For example, the review by Cardós-Alonso et al. classifies some self-efficacy and perceived preparedness outcomes under Kirkpatrick level III [2]. Such self-reports need to remain distinguishable from measured competence and workplace behaviour. Likewise, higher training load is not necessarily beneficial. Mehta et al. found lower accuracy during stressful VR training of firefighters, without a demonstrated advantage in subsequent retrieval of the same valve sequence within the session [5]. Long-term retention and operational transfer remained untested.

The framework should turn these distinctions into evaluation choices. A claim about decision-making under uncertainty, for example, calls for documenting which information is available, selecting constructs such as situation awareness, and assessing decision quality under those information constraints. Measures of perceived uncertainty describe experience; evidence of improved decision quality addresses impact. The following section outlines how the framework organizes these aspects of evaluation while keeping the scope of each claim explicit.

4. Towards an XR4CE Training Evaluation Framework

Demand profiles describe the requirements against which training is evaluated. Five levels organize what is evaluated, from design conditions to outcomes. Experience and impact are perspectives on the aspects of training being evaluated, rather than additional levels: experience is addressed primarily at

L2, impact at L4 and L5, while L1 and L3 help contextualize and interpret both. The levels distinguish evaluation purposes, not mutually exclusive psychological categories:

- **L1: XR and training design.** Documents the conditions provided, including task-relevant cues, interaction, stressors and instructional support.
- **L2: Training experience and immediate learner responses.** Addresses how trainees experience and respond to the training conditions, including presence, agency, cognitive load and appraisal.
- **L3: Task-related cognitive, regulatory and team processes.** Examines how trainees handle task demands through situation awareness, decision-making, stress regulation and coordination.
- **L4: Learning, competence acquisition and performance.** Distinguishes performance during practice from evidence of learning and acquired capability in the trained domain. Relevant outcomes include accuracy, critical errors and performance under load, interpreted in relation to the assessment design.
- **L5: Transfer and operational impact.** Examines the application of acquired capabilities to tasks or contexts that differ from training in specified, transfer-relevant ways and, where assessed, to workplace behaviour and operational outcomes.

For L4, successful practice alone does not establish learning or competence acquisition. Such claims require an assessment distinct from guided practice and an appropriate basis for judging change, such as pre-training performance or a comparison condition. Available assistance should be reported: an aid intended for operational use may legitimately remain available, whereas claims of independent capability require assessment without instructional prompts. For L5, the assessment must specify what differs from training and why that difference is relevant to the intended transfer claim. A difficulty increase within a rehearsed exercise does not, by itself, demonstrate transfer. Reporting the timing of assessment distinguishes immediate performance from retention.

These levels distinguish the aspects of training being evaluated, without implying an automatic progression between them. Retention can be examined at both L4 and L5. A separate companion impact model is to be developed to formulate testable hypotheses about causal relationships among training conditions, learner responses, task-related processes and outcomes.

We propose three CE-specific evaluation pathways through the framework, selected to foreground uncertainty, stress and adaptation to disruption in interdependent teams. Each links a combination of CE demands in Table 1 to relevant design conditions, learner responses, task-related processes and outcomes across the five levels. These pathways guide the selection and interpretation of measures; they do not imply established causal relationships or require every study to assess all five levels. *Operational cognition under uncertainty* connects incomplete or conflicting information and time pressure to situation awareness and decision quality. *Performance under stress* relates threat and other stressors to appraisal, regulation and task performance. *Team adaptation under disruption* connects interdependence and unexpected changes to coordination, adaptation and recovery. High consequences of error inform the performance and safety criteria across all three. The pathways can overlap and do not constitute an exhaustive classification. Studies select relevant constructs according to their intended claims, considering expertise, XR experience, stress reactivity and team familiarity where these affect interpretation.

Where access to professionals or operational settings is limited, assessments in new simulated tasks or contexts can provide evidence of transfer within the scope of those conditions; actual workplace application remains an open question. Evaluators need only assess the constructs and outcomes relevant to their claims.

Validation across diverse CE studies must examine demand coverage, the suitability of selected constructs and measures, and the framework's contribution to evaluation decisions. Findings across tasks, populations and demand profiles can establish its applicability and guide refinement.

5. Call for Action

The diversity of CE tasks, populations and demand combinations makes validation a collective undertaking. A single study can examine only part of the framework's scope. We therefore invite complementary studies through cross-lab collaboration, bringing together expertise and settings beyond individual groups' reach. Together, these studies could reveal which constructs and measures are useful across settings and where task-specific adaptations are needed.

We envisage the following *levels of involvement*, depending on interests, available material and opportunities for empirical work:

Discussion and feedback. Practitioners and researchers can examine whether the framework captures their CE characteristics and corresponding evaluation needs. Emergency response trainers might highlight decisions made with incomplete information, while industrial practitioners might emphasize procedural constraints. Such contributions can reveal missing constructs and limits to applicability.

Sharing demand profiles and study experiences. Existing studies could contribute descriptions of demands, measures and outcomes: for example, how time pressure was represented, which prompts trainees received, and what the performance test required. Undocumented conditions would remain identified as unknown. Relating these accounts could reveal recurring evaluation needs and evidence gaps, while helping groups position their findings within a broader research context.

Exploring shared evaluation elements. Groups planning studies could explore overlapping measures or reporting elements, such as documenting assistance similarly or including common items on perceived task demands. Such a bridging set could support both local evaluation and the development of shared measures. Its suitability across tasks, populations and instruments would need to be examined before treating scores as equivalent.

Collaborative application and comparison. Groups could apply the framework to their CE studies and compare findings. Healthcare and emergency response studies could compare decision-making under uncertainty across settings. Firefighting and industrial maintenance studies could examine how teams detect communication breakdowns and reorganize work under disruption, comparing shared coordination criteria and task-specific safety requirements. Comparisons with existing approaches could inform the framework's added value. Joint work would develop around participants' questions, access to professionals and practical constraints.

Each study retains its own contribution while helping establish recurring findings, contextual differences and remaining gaps. The workshop offers an opportunity to identify partners as well as feasible next steps for cross-lab activities.

6. Conclusion

The specific characteristics of challenging environments require corresponding evaluation constructs and criteria. We therefore argue for a dedicated XR4CE training evaluation framework that provides guidance for assessing experience, learning, competence acquisition and transfer under CE demands. Establishing its validity and practical value requires evidence from many studies across tasks, populations and settings. Cross-lab collaboration can connect these studies, enabling the XR4CE community to build cumulative evidence and clarify which training approaches support particular capabilities, under which demands, and how far the evidence supports transfer to other tasks and operational settings.

References

- [1] P. Strojny, N. Dużmańska-Misiarczyk, Measuring the effectiveness of virtual training: A systematic review, *Computers & Education: X Reality* 2 (2023) 100006. URL: <https://linkinghub.elsevier.com/retrieve/pii/S294967802200006X>. doi:10.1016/j.cexr.2022.100006.
- [2] M. d. C. Cardós-Alonso, L. Otero-Varela, M. Redondo, M. Uzuriaga, M. González, T. Vazquez, A. Blanco, S. Espinosa, A. M. Cintora-Sanz, Extended reality training for mass casualty incidents:

a systematic review on effectiveness and experience of medical first responders, *International Journal of Emergency Medicine* 17 (2024) 99. doi:10.1186/s12245-024-00685-3.

- [3] G. Makransky, G. B. Petersen, The Cognitive Affective Model of Immersive Learning (CAMIL): a Theoretical Research-Based Model of Learning in Immersive Virtual Reality, *Educational Psychology Review* 33 (2021) 937–958. URL: <https://doi.org/10.1007/s10648-020-09586-2>. doi:10.1007/s10648-020-09586-2.
- [4] D. L. Kirkpatrick, J. D. Kirkpatrick, *Evaluating training programs: the four levels*, 3 ed., Berrett-Koehler, San Francisco, 2006.
- [5] R. K. Mehta, J. Kang, Y. Shi, J. Du, Effectiveness of training under stress in immersive VR: an investigation of firefighter performance, gaze entropy, and pupillometry, *Frontiers in Virtual Reality* 6 (2025) 1542507. doi:10.3389/frvir.2025.1542507.
- [6] A. Cordeiro, R. Leite, L. Almeida, C. Neves, T. Silva, A. Siqueira, M. Catapan, I. Winkler, Preliminary Design Guidelines for Evaluating Immersive Industrial Safety Training, *Informatics* 12 (2025) 88. URL: <https://www.mdpi.com/2227-9709/12/3/88>. doi:10.3390/informatics12030088.
- [7] D. J. Harris, J. M. Bird, P. A. Smart, M. R. Wilson, S. J. Vine, A framework for the testing and validation of simulated environments in experimentation and training, *Frontiers in Psychology* 11 (2020) 605. doi:10.3389/fpsyg.2020.00605.
- [8] M. Lutz, S. R. Sorko, K. Seybold, Evaluation of the learning success in the implementation phase of Extended Reality, *Industry 4.0* 8 (2023) 99–104. URL: <https://stumejournals.com/journals/i4/2023/3/99>.
- [9] D. Zamora Alviarez, E. Drew, R. Alqasemi, Foundations for Interpreting VR Training Transfer: Transfer Distance and Claim Reach, *Electronics* 15 (2026) 3243. URL: <https://www.mdpi.com/2079-9292/15/15/3243>. doi:10.3390/electronics15153243.
- [10] R. Glaser, Instructional technology and the measurement of learning outcomes: Some questions, *American Psychologist* 18 (1963) 519–521. doi:10.1037/h0049294.
- [11] C. D. Morris, J. D. Bransford, J. J. Franks, Levels of processing versus transfer appropriate processing, *Journal of Verbal Learning and Verbal Behavior* 16 (1977) 519–533. doi:10.1016/S0022-5371(77)80016-9.
- [12] M. T. Kane, Validating the interpretations and uses of test scores, *Journal of Educational Measurement* 50 (2013) 1–73. doi:10.1111/jedm.12000.
- [13] A. Perkis, C. Timmerer, S. Baraković, J. Baraković Husić, S. Bech, S. Bosse, J. Botev, K. Brunnström, L. Cruz, K. De Moor, A. De Polo Saibanti, W. Durnez, S. Egger-Lampl, U. Engelke, T. H. Falk, J. Gutiérrez, A. Hameed, A. Hines, T. Kojic, D. Kukolj, E. Liotou, D. Milovanovic, S. Möller, N. Murray, B. Naderi, M. Pereira, S. Perry, A. Pinheiro, A. Pinilla, A. Raake, S. R. Agrawal, U. Reiter, R. Rodrigues, R. Schatz, P. Schelkens, S. Schmidt, S. Shafiee Sabet, A. Singla, L. Skorin-Kapov, M. Suznjevic, S. Uhrig, S. Vlahović, J.-N. Voigt-Antons, S. Zadtootaghaj, *QUALINET White Paper on Definitions of Immersive Media Experience (IMEx)*, arXiv preprint arXiv:2007.07032 (2020). URL: <http://arxiv.org/abs/2007.07032>.