



EVALUATING LANGUAGE TRANSFER ACROSS VIRTUAL REALITY SCENARIOS

Mekhribon Ismatova

Abstract. This extended abstract proposes a comparative study of beginner foreign language learning across four VR scenario families: a transport hub, a cafe, a school garden and a welcome desk. Adaptive VR would be compared with fixed support VR and matched classroom activities. The primary planned outcome is delayed performance on fresh non-VR communicative tasks. A common task specification would permit contextual localisation while preserving learning objectives. Illustrations and the pie chart depict proposed materials and curriculum allocations, not participant data. The contribution is a prospective scenario-transfer evaluation framework with an international scope.

Keywords: virtual reality; language learning; transfer; scenario design; comparative evaluation.

Rationale and research question. Successful communication in a rehearsed virtual setting may not transfer to an unfamiliar situation. Educational VR reviews identify limitations in task design, outcome measurement and retention assessment [1,2]. Language-oriented work on virtual immersion supports investigation of contextual activities [3], but findings from virtual worlds cannot automatically be transferred to a particular headset intervention.



The main research question is whether adaptive VR would improve delayed communication on unfamiliar tasks compared with fixed support VR. A secondary contrast would compare fixed support VR with matched classroom instruction. The prospective hypothesis is that adaptive support would benefit delayed transfer; no effect has yet been measured. The study would initially target school learners approximately 12 to 14 years old, subject to hardware eligibility and institutional approval.

Task design and contextual scope. Each scenario would connect a practical outcome to listening, reading, interaction and a short message. Learners would ask about a route, request an item, coordinate object placement or explain a notice. Information gaps and rotating roles would provide reasons to communicate. Figure 1 shows original phrase examples available during practice; displayed models would not count as independently produced language.

Educators could adapt familiar foods, routes, place names and social conventions while retaining the objective, information gap and assessment construct. Localisation could still change difficulty, so materials would require review and pilot calibration. An initial implementation in one setting could support later replication elsewhere; international relevance would not establish universal or cross-national effectiveness.

Illustrative scenarios and curriculum allocation



Figure 1. Proposed transport, cafe, garden and welcome-desk tasks. AI-generated conceptual illustration produced with OpenAI image generation; it does not depict study participants or implemented software.

Eight weekly sessions would provide four focal communicative opportunities per learner per session. Two sessions would cover each scenario family, giving eight planned opportunities per family and 32 overall. Retries, missed sessions and actual speaking turns would be logged separately. Equal scenario coverage is a provisional design choice rather than a demonstrated optimal allocation.

Proposed allocation of 32 focal task opportunities

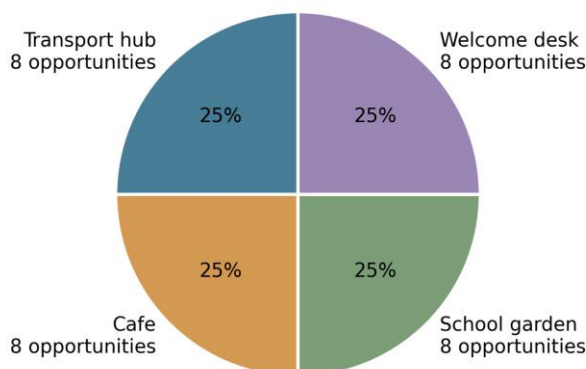


Figure 2. Proposed allocation of 32 focal opportunities: 8 per scenario family, or 25 percent each. These values describe the curriculum plan, not observed outcomes.

Proposed evaluation and conclusion. Feasibility work would first examine comprehension of instructions, headset tolerance, recruitment, scoring and teacher workload. The main study would then use three parallel conditions: adaptive VR, the same VR scenes with fixed support, and matched classroom role plays with fixed support. Where contamination was likely, classes would be allocated within sites after baseline assessment. Content, partner roles, scheduled practice time and focal opportunities would be matched. The VR-versus-classroom contrast would estimate a delivery-package difference; it would not isolate immersion perfectly.

Baseline, immediate post-intervention and four-week delayed assessments would use parallel non-VR role plays with unfamiliar maps, menus, plans and notices. The primary outcome would be the delayed mean score across four tasks. A provisional rubric would assess comprehension, message clarity, interaction and task fulfilment. Expert review and pilot scoring would precede use of a composite. Consistent with argument-based validation,

the inference would concern performance across sampled tasks, rather than overall certified proficiency [4].

Raters would be masked to condition where feasible, with a prespecified subset independently double-rated. Sample-size planning would account for the primary contrast, class clustering, expected attrition and an educationally meaningful effect [5]. No participant number is claimed sufficient. The analysis would adjust for baseline performance and the allocation structure, report effect estimates with confidence intervals and document missing-data assumptions. Technical interruptions would not automatically receive zero language scores.

Practice logs would distinguish participation, support, technical failures and teacher intervention. Interviews would explore task relevance and implementation experiences. Guardian consent, learner assent, minimal recording and an equivalent classroom option would precede recruitment. Task forms and speech-derived feedback would require validation for the intended age and language groups.

The proposed framework makes delayed transfer and contextual adaptation explicit. It could support a more interpretable test of adaptive VR than repeated assessment within a familiar training scene. Its effectiveness, measurement equivalence and wider portability remain open questions. The next step is expert review and feasibility testing, followed by a prospectively specified comparative study.

References:

[1] Radianti, J., Majchrzak, T. A., Fromm, J., and Wohlgenannt, I. (2020). A systematic review of immersive virtual reality applications for higher



education: Design elements, lessons learned, and research agenda. *Computers & Education*, 147, 103778. <https://doi.org/10.1016/j.compedu.2019.103778>

[2] Hamilton, D., McKechnie, J., Edgerton, E., and Wilson, C. (2021). Immersive virtual reality as a pedagogical tool in education: A systematic literature review of quantitative learning outcomes and experimental design. *Journal of Computers in Education*, 8, 1–32. <https://doi.org/10.1007/s40692-020-00169-2>

[3] Lan, Y.-J. (2020). Immersion into virtual reality for language learning. *Psychology of Learning and Motivation*, 72, 1–26. <https://doi.org/10.1016/bs.plm.2020.03.001>

[4] Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>

[5] Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, 8(1), 33267. <https://doi.org/10.1525/collabra.33267>

