

PAPER

Toward an Integrative Framework for Valid Research in Educational Technology: Insights from Virtual Reality Learning

Miriam Mulders ,
Katja Buntins 

University of Duisburg-Essen,
Essen, Germany

miriam.mulders@uni-due.de

ABSTRACT

This methodological and conceptual contribution develops an integrative framework for research on educational technology that systematically links internal and external validity, different research approaches, and heterogeneous outcome dimensions. Drawing on the recurring disappointment regarding the limited learning benefits of digital technologies—illustrated by the example of virtual reality (VR)—the article argues that technological innovations can only be adequately assessed if their effects can be clearly attributed and transferred to authentic learning contexts. Using central research logics, including media comparison, value-added, and ATI/LTI approaches, it shows how the choice of research approach shapes the type of evidence generated and the associated validity requirements. In addition, the article extends existing outcome models by incorporating a behavioral dimension to reflect the multidimensional nature of teaching and learning processes. The proposed framework illustrates that progress in educational technology research depends on the combination of methodological rigor, context-sensitive evidence, and multidimensional outcome assessment. Finally, practice-oriented recommendations are presented to support cumulative evidence development, transparent reporting standards, and the integration of laboratory and field research.

KEYWORDS

educational technology, virtual reality (VR), internal validity, external validity, research design

1 INTRODUCTION

For several decades, digital technologies have been promoted as a means of improving learning outcomes [1]. Yet, across successive waves of innovation, initial optimism has repeatedly given way to disillusionment, as anticipated benefits have often failed to translate into measurable effects [2]. Meta-analyses consistently show that technology-enhanced learning rarely outperforms traditional instruction in controlled comparisons [3], [4].

Mulders, M., Buntins, K. (2026). Toward an Integrative Framework for Valid Research in Educational Technology: Insights from Virtual Reality Learning. *International Journal of Emerging Technologies in Learning (iJET)*, 21(4), pp. 18–40. <https://doi.org/10.3991/ijet.v21i04.62876>

Article submitted 2026-06-11. Revision uploaded 2026-08-05. Final acceptance 2026-08-07.

© 2026 by the authors of this article. Published under CC-BY.

This recurring pattern has prompted researchers to reconsider how media effects are conceptualized and measured in educational research. Critics argue that positive findings reported in empirical studies are often not directly attributable to the medium itself but instead result from differences in the underlying instructional design [5], [6].

Technology-based interventions are frequently more carefully planned and methodologically refined than their comparison conditions, meaning that observed learning gains may be attributable to the quality of the instructional design rather than to inherent properties of the medium [7]. This issue has fueled a long-standing debate over whether observed effects should be attributed to instructional design rather than to the medium itself.

These recurring debates underscore the need to examine the methodological foundations of educational technology research, particularly when emerging technologies are evaluated under rapidly changing technical, instructional, and institutional conditions. Virtual reality (VR) provides a particularly instructive context because its educational effects cannot be considered independently of the design of the virtual environment, the instructional tasks, learner characteristics, implementation conditions, and the outcomes selected for evaluation.

This paper therefore develops an integrative framework that links internal and external validity, different research approaches, and multidimensional outcome categories. VR-based learning serves as a paradigmatic case through which the framework is specified and applied. The framework is intended to support researchers in positioning individual studies within broader research programs, selecting designs that align with their study questions, and interpreting findings in relation to their methodological and practical scope. In this way, the contribution addresses both the scientific evaluation and the educational implementation of emerging technologies. The framework is not presented as an empirically validated measurement model. Rather, it offers a theoretically grounded orientation structure whose feasibility and utility should be examined through future applications across research projects, technologies, and educational contexts.

2 THEORETICAL BACKGROUND

This section outlines the theoretical foundations of the framework within the field of technology-enhanced learning. It first reviews key developments in VR research as a guiding example before introducing internal and external validity as core principles of rigorous empirical inquiry. It then summarizes major research designs and methodological strategies used to examine effectiveness and generalizability. Together, these elements provide the conceptual basis for evaluating educational technologies and inform the subsequent recommendations for research and practice.

2.1 VR as an emerging educational technology

Virtual reality has become a prominent emerging technology in education because it enables learners to interact with spatial, simulated, and potentially authentic environments that may otherwise be inaccessible, costly, dangerous, or difficult to reproduce. Educational applications include laboratory simulations, vocational trainings, medical procedures, historical reconstructions, and perspective-based learning

scenarios. Across these contexts, VR has been associated with cognitive, affective, and psychomotor outcomes [8], [9], [10], [11].

However, VR does not constitute a uniform instructional intervention. Educational VR applications differ substantially in their technological features, interaction design, instructional guidance, duration, social configuration, and integration into broader instructional arrangements. Consequently, findings cannot be interpreted solely as effects of VR. Instead, they reflect specific combinations of technology, instructional design, learner characteristics, implementation conditions, and outcome measures [12].

These characteristics make VR a particularly relevant context for examining methodological validity in educational technology research. Controlled laboratory studies may isolate specific design features and support causal inference, but they often simplify the institutional and instructional conditions under which VR is used in practice. Field studies, by contrast, capture authentic implementation conditions but may offer less control over confounding variables. Moreover, the educational value of VR may be reflected not only in cognitive performance but also in motivation, action competence, psychomotor performance, transfer behavior, or implementation efficiency [11].

Virtual reality, therefore, serves in this paper not merely as an example but as a paradigmatic case of an emerging educational technology. It illustrates why the evaluation of technological innovation requires the coordinated consideration of research scope, validity, and outcome selection. The proposed framework uses this context to support the design and interpretation of studies seeking to determine whether, how, for whom, and under which instructional conditions emerging technologies contribute to learning.

2.2 Internal and external validity as cornerstones of educational technology research

This paper clarifies the importance of internal and external validity in educational technology research. Although the two concepts are complementary, they are also in methodological tension [12]: High internal validity requires the control of confounding variables under relatively artificial conditions [13], whereas high external validity requires realistic, less controlled contexts [14]. Educational technology research must account for this trade-off. Internal and external validity should therefore be understood as successive elements of a cumulative research process: Internal validity supports the identification of robust effects, whereas external validity supports their generalizability and practical relevance. Accordingly, study designs can be structured—analogue to levels of evidence in medicine—to progressively increase both methodological rigor and applicability [15].

Internal validity. Internal validity refers to the extent to which an observed effect can be confidently attributed to the intended intervention rather than to alternative explanations. A study is internally valid when competing causes are systematically ruled out and the causal relationship between the independent and dependent variables can be established with sufficient precision [16]. Internal validity therefore constitutes the foundation of causal inference. In educational technology research, threats to internal validity are particularly salient. Common threats include confounding variables, such as prior knowledge or teacher quality [17], [18]. If such factors are not adequately measured and controlled, the estimated intervention effect may be either inflated or attenuated [19], [20]. This tension is clearly illustrated

in VR-enhanced learning: laboratory studies on VR, for example [21], often achieve high internal validity through standardized procedures and strict control of confounding variables but are frequently criticized for limited ecological realism and low transferability to everyday educational settings. Conversely, classroom-based VR studies (e.g., [22]) provide valuable practical insights but often lack the methodological rigor required for unequivocal causal attribution.

The following sections outline typical threats to internal validity and illustrate how they may arise in VR-enhanced learning contexts:

- a) Selection effects represent a pervasive challenge: participants often self-select into studies, while marginalized or hard-to-reach groups are systematically underrepresented. This may undermine randomization and bias effect estimates [23], [24]. In educational research, non-response often results from limited interest, time constraints, language barriers, or low levels of trust [25], [26], [27]. In VR-enhanced learning, this risk is particularly pronounced because studies are typically conducted at well-resourced institutions with access to headsets and technical support, potentially skewing participation toward more privileged populations. Accordingly, existing critiques point to a bias toward WEIRD (Western, Educated, Industrialized, Rich, Democratic) samples and suggest that underprivileged groups may be underrepresented or excluded [28].
- b) Further threats include maturation processes, which refer to natural biological, cognitive, or social developments that occur over time independently of the intervention [29], [30]. In VR-enhanced learning, maturation effects may be particularly salient during longer intervention periods: learners may improve because of increasing familiarity with immersive environments, enhanced spatial reasoning, or gradual adaptation to VR controls rather than because of the intervention itself. For example, a longitudinal comparison of immersive VR simulations and video playback in a science course found that performance in the VR group increased across successive interventions and remained stable in delayed post-tests, suggesting that part of the improvement may have reflected cumulative exposure rather than the intervention alone [31].
- c) Testing effects are also relevant: repeated exposure to assessment materials can produce practice gains, greater familiarity, or reduced anxiety, thereby mimicking genuine intervention effects [32], [33]. In VR-enhanced learning environments, this risk is particularly pronounced because learners may become increasingly familiar with navigation, controls, or interaction mechanics. Consequently, improved performance may partly reflect familiarity with the medium rather than actual learning gains. For example, Galor et al. [34] demonstrated this issue in the VR-Color Trails Test, where repeated administrations raised concerns about practice effects and led to the development of alternative test versions.

To mitigate these threats, a combination of design-based and statistical strategies is required. Randomized controlled trials (RCTs) remain the gold standard because random assignment balances confounding factors across groups [35], [36]. When randomization is infeasible, for instance in intact school classes, quasi-experimental designs can be employed and complemented by methods such as matching [37], analysis of covariance (ANCOVA), or regression adjustment [38]. For nested educational data, such as students within classes and classes within schools, hierarchical linear models or mixed-effects models can account for group-level influences [39], [40]. Weighting techniques, including post-stratification and calibration weighting, can further correct imbalances in demographic or socioeconomic variables [41], [42].

Finally, more recent approaches emphasize the use of causal-inference frameworks and directed acyclic graphs (DAGs) to identify confounders systematically [43], [44], while sensitivity analyses and simulation studies can estimate the potential influence of unobserved confounders and assess the robustness of the findings [45], [46].

External validity. External validity refers to the generalizability of findings across contexts, populations, and time [47]. Cronbach and Shapiro [14] organize these dimensions within the UTOS framework. At the unit level, external validity concerns the transferability of findings across different populations, such as learners from other age groups, cultural backgrounds, or socioeconomic contexts [48]. At the treatment level, generalizability depends on whether comparable effects emerge when interventions are operationalized differently, for example, in laboratory-based versus field-based implementations [49]. With regard to outcomes, findings should remain consistent across different measurement instruments or modes of data collection, for instance when self-reports are complemented by behavioral observations or physiological measures [50]. Finally, at the setting level, external validity is established when effects can be replicated across different institutions, teaching cultures, or technological infrastructures [51]. VR research is still dominated by individual studies conducted under highly specific conditions and focused on particular applications. Systematic replication studies, in which the same intervention is reexamined across different units, treatments, outcome measures, or settings, remain rare, thereby limiting the extent to which robust conclusions about effectiveness can be drawn. Nevertheless, there are notable exceptions. For example, Kohnke and Fount [52] replicated an earlier VR-based teacher training paradigm with different participant groups and design variations, providing evidence of how effects converge or diverge across contexts. Similarly, Han, Liu, and Zheng [53] conducted a VR laboratory experiment that explicitly compared learners with high versus low prior knowledge to examine how signaling effects depend on aptitude, thereby demonstrating that VR interventions can be systematically tested across different learner groups. Such studies indicate that generalizability in VR research must be empirically examined through deliberate replication and extension.

However, more comprehensive models of external validity are needed. Extensions such as the M-STOUT framework therefore incorporate additional dimensions, namely mechanisms and time [47]. Mechanisms refer to the causal processes through which interventions produce their effects. Identifying these processes is critical because, without understanding how and why an intervention works, it remains unclear whether its effects will transfer across contexts. External validity requires that relevant covariates can be reproduced across populations and settings [54]. In educational research, such mechanisms are often conceptualized as learning processes that link instructional quality to outcomes [55]. The time dimension addresses both short- and long-term generalizability. Findings should remain stable across different points in time and robust under longer-term implementation [47]. This includes determining whether technology-based interventions remain effective across multiple school years or whether initial novelty effects diminish over time [56]. Longitudinal studies further examine whether gains persist beyond immediate post-tests [57]. In VR-enhanced learning, both dimensions are particularly salient. Mechanistic research increasingly investigates how immersive features influence presence, motivation, and cognitive engagement as mediators of learning [58]. At the same time, time-based evidence remains limited but is gradually emerging: although short-term engagement effects may decline with repeated use, longitudinal studies suggest that well-designed VR training can produce durable improvements in complex skills and transfer [59]. These findings indicate that

external validity in VR research concerns not only who and where, but also how and over what period effects occur.

Building on these conceptual considerations, assessing external validity requires a range of methodological strategies. Replication studies examine whether effects persist under modified conditions, while systematic reviews and meta-analyses estimate both mean effects and heterogeneity across studies [60]. Meta-regressions and subgroup analyses further identify moderators and mediators that help explain variation in effects. In addition, cross-validation techniques [61], [62] can be used to assess robustness by partitioning datasets and examining model stability across subsamples. In educational technology research, such approaches are particularly important because findings often vary substantially across contexts.

2.3 Research approaches in educational technology

Methodological issues cannot be considered in isolation, as research approaches shape the type of evidence generated. According to Buchner and Kerres [63], three central approaches can be distinguished in educational technology research: the media comparison approach, the value-added research approach, and the aptitude-treatment interaction (ATI) or learner-treatment interaction (LTI) approach. Each approach addresses a distinct set of research questions.

The media comparison approach was historically predominant. It compares emerging technologies, such as augmented reality, interactive whiteboards, or AI-based systems, with established media such as books, worksheets, or videos. Typical questions concern whether learners achieve better outcomes with one medium than with another. Its main advantage lies in its simplicity and orienting function, as it can provide initial indications of the potential effectiveness of new technologies. However, it has been criticized for remaining strongly technology-centered [5], [6] and for risking the attribution of effects to the medium rather than to the underlying instructional design [64]. Its contribution to evidence development therefore lies less in identifying causal mechanisms than in initiating research programs and motivating further investigation. Nevertheless, this approach remains highly prevalent in VR research: recent systematic reviews [65], [66] show that most VR studies still adopt media comparison designs, typically contrasting immersive VR with conventional instruction.

The value-added research approach adopts a more differentiated perspective by focusing on the targeted modification of specific design features rather than on comparisons between entire media systems. Typical questions address whether elements such as explanatory feedback or personalized language improve learning outcomes or whether the omission of particular features reduces performance. Its strength lies in its theoretical precision, as it enables the identification of effective design principles and contributes to the development of instructional theories [67], [68]. However, the resulting findings are often context-dependent and not readily generalizable. For evidence development, this approach provides building blocks for design-oriented recommendations that extend beyond the demonstration of overall effects. The number of VR-enhanced learning studies following this approach has increased. For example, Klingenberg et al. [69] showed that generative learning strategies embedded in VR can enhance learning when integrated into already engaging environments. Other studies likewise demonstrate how specific instructional features can add value beyond conventional instruction. Overall, this development reflects a shift toward isolating the contributions of individual design elements in VR research.

The ATI or LTI approach broadens the perspective by systematically incorporating individual learner characteristics. Its aim is to identify interactions between learner characteristics, such as prior knowledge, motivation, or self-regulation, and specific media-based interventions. Typical questions include: “Do learners with high prior knowledge benefit more from simulations than beginners?” or “What role does metacognitive knowledge play in the use of adaptive learning systems?” The advantage of this approach lies in its practical relevance, as it takes learner heterogeneity into account and thereby enables more differentiated design decisions. A frequent criticism is that ATI/LTI studies are methodologically complex and require large samples to detect interaction effects reliably [70]. For evidence development, this approach helps bridge the gap between theory and practice by identifying the conditions under which, and the learners for whom, specific technologies are effective. In VR-enhanced learning, early empirical studies already provide proof of concept for this approach. For example, Han and colleagues [53] conducted an immersive VR experiment that tested signaling tools across learners with different levels of prior knowledge, showing that the effect of signaling depends on learner aptitude, specifically prior knowledge. An earlier example is provided by Chen et al. [71], who compared guided and non-guided exploration modes in VR across learners with different characteristics and found that navigational aids benefit learners differentially.

2.4 Research designs in educational technology

Educational technology research cannot be understood solely by distinguishing among theoretical approaches. It must also be examined through the methodological designs by which these approaches are operationalized. Each research tradition is characterized by recurring design patterns that reflect its epistemological orientation and its emphasis on internal or external validity. Systematic reviews and meta-analyses [72], [73], [74] indicate that these patterns have persisted across decades of research rather than representing arbitrary or isolated methodological choices.

The media comparison approach is rooted in the experimental paradigm and seeks to identify causal relationships between instructional media and learning outcomes. It typically relies on controlled experiments, quasi-experimental pre-post designs, and standardized assessments intended to maximize internal validity by controlling confounding variables and standardizing treatment conditions [5], [6], [75]. This design logic follows the tradition established by Shadish, Cook, and Campbell [13], who argue that randomized and controlled experiments remain the most reliable means of inferring causal relationships. However, the high degree of control that supports internal validity may also reduce ecological realism and, consequently, external validity [76]. Findings from tightly controlled laboratory studies therefore often require replication in naturalistic settings to establish their generalizability [47]. In VR research, this logic is reflected in the predominance of highly controlled laboratory experiments, in which participants are typically invited to university laboratories and exposed to virtual environments under artificial and standardized conditions. Such studies seek to isolate causal effects by minimizing contextual variability and measuring learning-related correlates with high precision. Consequently, VR research has developed empirically validated and widely used instruments for assessing VR-specific constructs such as presence, including the Multimodal Presence Scale developed by Makransky, Lilleholt, and Aaby [77].

The value-added approach shifts the focus from causal attribution to the pedagogical and practical effectiveness of interventions. Rather than comparing media *per se*, it examines the contribution of specific instructional designs or technological affordances to measurable improvements in learning processes. Methodologically, this approach often employs longitudinal or within-subject designs, mixed methods, and design-based research, thereby combining experimental control with contextual realism [78], [79], [80]. These designs correspond to what Slavin [76] terms effectiveness studies, which emphasize sustained impact and external relevance over isolated treatment effects. The value-added tradition therefore occupies an intermediate position along the validity continuum: it combines the rigor of controlled designs with the authenticity of field-based inquiry and seeks to balance internal and external validity within iterative, practice-oriented research cycles. In VR-based learning research, value-added perspectives are increasingly used to determine how instructional design features contribute to meaningful learning. Accordingly, Mulders and colleagues proposed the Meaningful iVR Learning Framework [81], which posits that the added value of VR emerges when immersion, interaction, and imagination, as key features of VR learning, are purposefully aligned with instructional goals such as reducing extraneous load, segmenting complex tasks, and fostering generative learning activities. Numerous empirical studies have investigated which instructional methods effectively support VR-based learning. For instance, segmentation and pretraining have been shown to reduce cognitive load and enhance comprehension in VR-based science learning [82], [83], while generative learning strategies such as self-explanation or summarization significantly improve knowledge transfer and retention [84], [85]. Collectively, these studies exemplify the value-added logic in VR research: they move beyond media comparison and instead identify design and instructional principles that make immersive learning pedagogically effective.

The ATI/LTI approach [70] advances this methodological development by emphasizing context sensitivity and individual variation. Rather than testing universal effects, ATI/LTI studies investigate for whom and under which conditions particular instructional interventions are effective. This requires analytical strategies capable of detecting moderator and mediator effects, such as regression analyses, multilevel modeling, and cluster or interaction analyses, which reveal the mechanisms and boundary conditions underlying observed effects [13], [86]. Because these designs prioritize heterogeneity and contextual variation, they can strengthen external validity by showing whether and how causal mechanisms generalize across populations, settings, and time [47]. In VR-based learning research, ATI/LTI principles are increasingly applied to examine how learner characteristics moderate the effectiveness of immersive experiences. Studies have shown, for instance, that learners with higher spatial ability benefit more from immersive visualization than those with lower spatial ability [87], while novice learners may experience greater cognitive load and therefore require additional instructional support [88], [89]. Such research demonstrates that VR-based learning is not universally effective but depends on the alignment between learner aptitudes and the instructional design of the virtual environment. By systematically investigating these interactions, VR studies contribute to a more differentiated understanding of when and for whom immersive learning environments are most beneficial, thereby extending external validity through contextual sensitivity.

Figure 1 illustrates the relationship between causal inference and validity in educational technology research. The horizontal axis represents increasing strength of causal inference, while the vertical axis reflects a shift from internal

toward external validity. The plotted research designs indicate how methodological approaches are positioned along these dimensions: media comparison studies, particularly RCTs, emphasize strong causal inference and internal validity, value-added approaches occupy an intermediate position, and ATI/LTI approaches place greater emphasis on external validity, contextual variation, and the analysis of differential effects. As Shadish, Cook, and Campbell [13] emphasize, progress in causal understanding depends not on privileging one form of validity but on their systematic integration. Educational technology research can develop cumulatively when evidence from internally and externally focused studies is connected across research programs, while individual studies retain clearly defined methodological priorities.

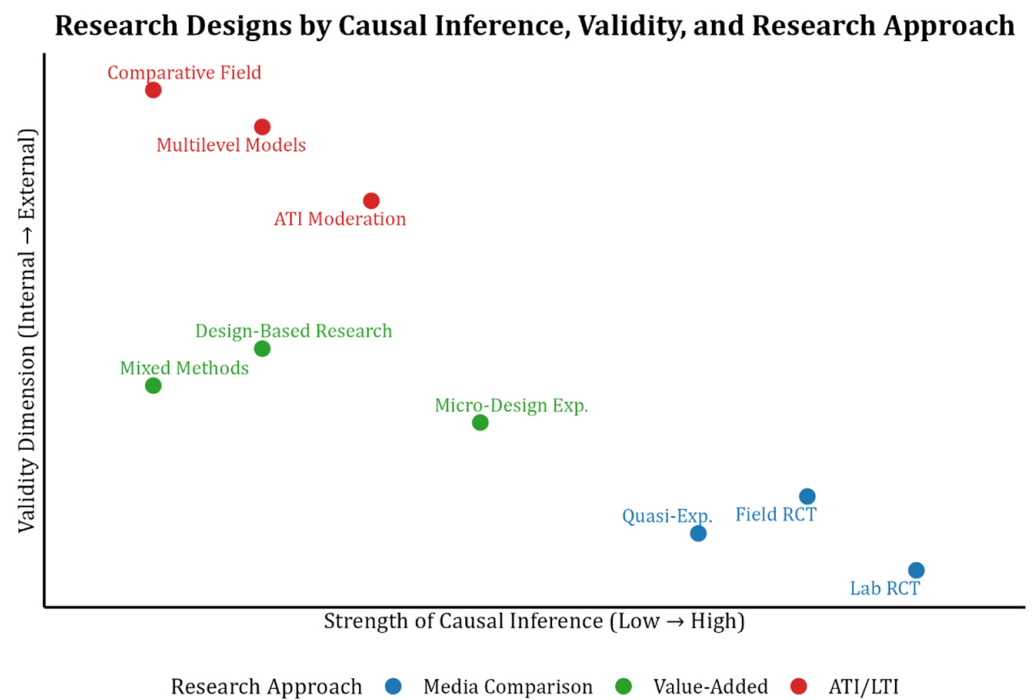


Fig. 1. Research designs by strength of causal inference and validity dimension

2.5 Relevant outcomes in educational technology research

Educational technology research is concerned not only with determining whether an intervention is effective or whether its findings can be generalized to other contexts. Equally important is the question of which learning outcomes are assessed and whether they adequately reflect the complexity of educational processes. The practical relevance of a study therefore depends on whether the selected outcomes align with its specific educational objectives and intended scope. When the impact of technology is evaluated exclusively through cognitive performance measures, the resulting perspective remains narrow and neglects motivational, emotional, and behavioral dimensions [67], [90]. Indeed, meta-analyses indicate that the predominance of cognitive outcome measures has persisted across decades of digital learning research [73], [74]. This pattern is also well documented in VR research [58]. However, not every study needs to assess all possible outcome dimensions. Outcome selection should therefore be theoretically justified and aligned with the mechanisms, instructional goals, and forms of learning addressed by the intervention.

Broadening the conceptualization of outcomes is particularly important in educational research. Unlike pharmacological studies, in which an intervention may target a single measurable parameter, such as blood pressure, learning processes are inherently multidimensional and context-dependent. In addition to the acquisition of knowledge and skills, they encompass affective, motivational, social, and behavioral processes that unfold over time. If these dimensions are overlooked, research risks reducing education to a purely cognitive endeavor and failing to capture its broader transformative potential [64].

Buchner and Mulders propose a framework that distinguishes four key outcome dimensions: cognition, affect, psychomotor performance, and efficiency [91]. This structure enables the systematic categorization of research findings while simultaneously highlighting their relevance to educational practice. For the purposes of this paper, the framework is extended by adding a behavioral dimension, as behavior precedes learning gains in many learning models and may therefore provide an early indication of emerging effects [92], [93].

Table 1 summarizes the five outcome dimensions considered in this paper (i.e., cognition, affect, psychomotor performance, efficiency, and behavior), provides brief definitions, and presents illustrative examples together with possible measurement methods. The classification is intended as a heuristic rather than an exhaustive taxonomy and may be extended to include additional outcome categories. The examples are therefore not prescriptive, rather, they illustrate how different types of outcomes can be operationalized in educational technology research.

Table 1. Outcome variables in educational research and their measurement

Variable	Definition	Exemplary Constructs	Measurements
Cognition	Refers to mental processes such as understanding, remembering, analyzing, or problem-solving.	• Factual knowledge • Conceptual understanding • Ability to transfer • Problem-solving skills • Cognitive load	• Knowledge tests (e.g., multiple choice, open questions) • Transfer tasks • Eye movement measurement • Cognitive load questionnaires (e.g., NASA-TLX [94])
Affect	Includes emotional and motivational aspects of learning.	• Motivation to learn • Interest • Self-efficacy • Flow experience • Frustration tolerance • Empathy • Perspective taking	• Questionnaires (e.g., SELMO [95], IMMS [96]) • Surveys • Emotional scales (e.g., PANAS [97]) • Observation of non-verbal signals • Measurement of heart rate or hormones
Psychomotor skills	Refers to physical-motor learning performance (i.e., learning and performing movements or actions).	• Fine motor skills • Precision of action • Coordination • Dexterity • Test implementation	• Observation and videography • Error rates in the execution of actions • Motion sensors • Time measurements for sequences
Efficiency	Ratio between the effort invested (e.g., time, resources, cognitive load) and the learning success achieved.	• Time spent per learning success • Ratio of input to output • Subjective assessment of the effort-benefit ratio	• Time-learning performance ratio • Effort estimates • Cost-benefit analyses • Questionnaires on perceived efficiency
Behavior	Examines longer-term changes in learning behavior.	• Use of learning strategies • Media selection behavior • Transfer to real situations	• Analysis of log data • Learning diaries • Follow-up interviews • Observation of behaviors

Existing frameworks offer important but partial perspectives on the evaluation of educational technologies. Validity-oriented frameworks such as UTOS distinguish the units, treatments, outcomes, and settings across which findings may generalize [14]. Extensions such as M-STOUT additionally incorporate mechanisms and time, thereby providing a more differentiated account of external validity [47]. These approaches are particularly valuable for assessing the scope of empirical claims, but they do not specify how different research traditions in educational technology generate distinct forms of evidence. Research approaches such as media comparison, value-added, and ATI/LTI studies address this methodological dimension by distinguishing whether a study examines overall media differences, specific instructional features, or interactions between learner characteristics and treatments. However, these approaches do not themselves provide an integrated account of internal and external validity or of the range of educational outcomes that may be relevant to a technological intervention.

Outcome-oriented classifications, in turn, broaden the evaluation of educational technologies beyond cognitive achievement by incorporating affective, psychomotor, efficiency-related, and behavioral dimensions [91]. Their primary contribution lies in clarifying what constitutes an educational effect, but they do not address how such effects should be causally identified, generalized, or related to different research approaches.

The proposed framework does not replace these established perspectives. Rather, it integrates their complementary contributions by linking three questions: What kind of claim can a study support in terms of internal and external validity? Which research approach corresponds to the study's question and stage of evidence development? Which outcomes adequately represent the educational purpose of the technology? This integration is particularly relevant for emerging technologies such as VR, whose effects depend on the interplay among technological features, instructional design, learner characteristics, implementation settings, and selected outcome measures.

However, the research approaches and validity dimensions discussed above should not be interpreted as competing standards that every study must satisfy simultaneously. Instead, they represent directed research scopes corresponding to different questions and stages of evidence development. A controlled media comparison study may appropriately prioritize causal attribution, whereas a field-based implementation study may focus on feasibility, contextual conditions, and transferability. Integration becomes particularly relevant when an instructional design involves multiple interacting components, when findings are intended to inform implementation across contexts, or when learner characteristics are expected to moderate the effectiveness of an intervention. In such cases, combining evidence across validity dimensions, research approaches, and outcome categories can provide a more comprehensive account of whether, how, for whom, and under which conditions an educational technology is effective.

3 AN INTEGRATIVE FRAMEWORK FOR RESEARCH ON EMERGING EDUCATIONAL TECHNOLOGIES

The preceding sections demonstrate that research on emerging educational technologies cannot be evaluated according to a single methodological criterion. Studies differ in the questions they address, the degree of causal control they require, the contexts to which their findings are intended to apply, and the outcomes through

which educational value is assessed. These differences do not necessarily constitute methodological weaknesses. Rather, they reflect directed research scopes that are appropriate for different stages and purposes of technology-related research.

The proposed framework integrates three dimensions: validity, research approach, and educational outcomes. It does not require every individual study to maximize internal and external validity, combine all research approaches, or assess all possible outcomes. Instead, it serves as an orientation tool for positioning individual studies within broader cumulative research programs.

In the context of VR-based learning, for example, a controlled media comparison study may provide initial evidence of differences between an immersive application and a conventional instructional format. A subsequent value-added study may isolate the contribution of specific instructional features, such as guidance, feedback, or learner control. ATI/LTI research may then examine whether these features function differently depending on learners' prior knowledge, spatial ability, motivation, or technological experience. Field and replication studies can further determine whether the identified effects persist across institutions, instructional arrangements, and longer implementation periods.

Through this mapping, the framework supports both research design and evidence interpretation. It helps researchers clarify what type of claim a study can support, which outcomes are relevant to its instructional purpose, and how its findings contribute to the cumulative evaluation of an emerging technology. Throughout the framework, VR is used to demonstrate how these methodological decisions shape the development of practically relevant and transferable evidence.

3.1 Validity dimension

The validity dimension defines the scope of the claims a study can support. Internal validity concerns the credibility of causal attribution [16], whereas external validity concerns the transferability of findings across learners, instructional settings, implementation contexts, and time [14], [47]. Within the proposed framework, these dimensions are not treated as requirements that every study must maximize simultaneously. Rather, they represent distinct methodological priorities that can be coordinated across cumulative research programs.

3.2 Research approach dimension

The research approach determines the type of question addressed and the evidence generated. Media comparison studies examine differences between instructional formats, value-added studies isolate the contribution of specific design features, and ATI/LTI studies investigate interactions between instructional conditions and learner characteristics [70]. The framework conceptualizes these approaches as complementary research scopes rather than as a fixed hierarchy or sequence that every project must follow.

3.3 Outcome dimension

The outcome dimension concerns which educational effects are considered relevant to the purpose of an intervention. Depending on the instructional objectives,

studies may examine cognitive, affective, psychomotor, efficiency-related, or behavioral outcomes [90]. The framework does not require all categories to be assessed. Rather, it promotes a theoretically justified alignment among the intended educational function of the technology, the selected outcome measures, and the claims derived from them.

3.4 Illustrative application to VR-based second-language learning

The preceding sections have outlined the proposed framework, which integrates research approaches, validity dimensions, and outcome categories into a coherent perspective on educational technology research. The framework is intended not as a prescriptive model but as an orientation tool for understanding how methodological choices, validity considerations, and outcome selection interact. Accordingly, it does not assume that every individual study must simultaneously maximize internal and external validity, combine all research approaches, or assess all outcome dimensions. Rather, each study adopts a directed research scope that reflects its specific question, theoretical assumptions, methodological priorities, and instructional context. The framework is intended to make these choices explicit and to demonstrate how studies with different scopes can contribute complementarily to cumulative evidence development.

To illustrate the practical use of the framework, consider the development of a VR (or AR) application for second-language acquisition [98], [99]. An initial media comparison study could examine whether learners using a VR-based language learning environment achieve different outcomes from those receiving conventional classroom- or textbook-based instruction. Such a study may prioritize internal validity through standardized learning materials, comparable instructional time, and control of prior language proficiency. Depending on the instructional objective, relevant outcomes may include vocabulary acquisition, listening comprehension, speaking performance, motivation, or language-learning anxiety. The selected outcomes should therefore reflect the specific educational function attributed to the VR environment rather than serve as a general test of whether VR “works.” A subsequent value-added study could investigate the contribution of individual instructional features within the VR environment, such as subtitles, corrective feedback, conversational prompts, scaffolding, or interaction with virtual agents. Here, internal validity remains important for isolating the contribution of the selected feature, while external validity becomes increasingly relevant when the feature is intended for use across different instructional settings. Outcome selection may include not only language performance, but also cognitive load, willingness to communicate, perceived competence, or interaction behavior. An ATI/LTI study could then examine whether the effectiveness of these features differs according to learners’ prior proficiency, motivation, anxiety, cultural background, or technological experience. Such analyses extend the scope of inference by identifying for whom and under which conditions a particular instructional design is effective. Finally, field-based and longitudinal studies could examine whether language skills acquired in VR transfer to authentic communication situations, remain stable over time, and can be feasibly integrated into regular language instruction.

More generally, the framework can help position each study according to its research scope, validity priorities, and selected outcomes. Its practical utility, therefore, lies not in prescribing a single comprehensive design but in supporting the coordination and interpretation of complementary studies within a cumulative research program.

The following chapter builds on the framework's conceptual foundation to derive recommendations for future research and practice. These recommendations translate the framework's key insights into actionable guidance by showing, for example, how different research designs can be combined to address both internal and external validity, how outcome measures can be broadened beyond cognition, and how VR-based research designs can serve as a blueprint for investigating other emerging technologies. In doing so, the section aims to support researchers and practitioners in designing studies that are both methodologically rigorous and practically relevant.

4 RECOMMENDATIONS FOR THE FURTHER DEVELOPMENT OF EDUCATIONAL TECHNOLOGY RESEARCH

4.1 Principles for action and future research needs

The preceding discussion shows that, although educational technology research has generated important insights in recent decades, the resulting evidence often remains selective and fragmented. What is currently known? First, meta-analyses of media comparison studies indicate that, on average, emerging technologies rarely produce superior learning outcomes compared with traditional media [4], [100]. Second, value-added studies demonstrate that learning success depends not on the medium itself, but on specific design features, such as feedback, structuring, or the reduction of cognitive load [67]. Third, ATI/LTI studies show that effects are strongly influenced by learner characteristics, including prior knowledge, motivation, and learning strategies [70], [101]. Overall, media effects are contingent on design, context, and learner characteristics [64].

To support the systematic coordination of these different research scopes, we propose a set of guiding principles for the further development of educational technology research. Based on the three dimensions of the framework (i.e., research approaches, validity, and outcomes), these principles emphasize complementary validity, the combination of research approaches, the integration of multiple learning outcomes, cumulative evidence development, practical relevance, and transparency. Table 2 synthesizes these principles, illustrates them with examples, and outlines possible steps for implementation.

Table 2. Principles of action to guide future educational technology research

Principle of Action	Description	VR-Examples	Indicative Implementation
Complementary validity	Internal and external validity must be understood as interrelated rather than competing goals.	Controlled VR-simulation in medical education, later tested in real hospital training.	- Encourage sequential study designs that progress from controlled experiments to field trials. - Promote funding schemes that explicitly require replication across contexts.
Combination of research approaches	Different research approaches (i.e., media comparison, value-added, and ATI/LTI) address complementary questions and can be linked across cumulative research programs where this is appropriate to the research objective.	Media comparison of VR vs. textbook, followed by value-added analysis of VR-specific features, such as the integration of AI-based avatars.	- Design research programs in which different approaches are selected and combined according to the questions, maturity, and implementation context of the technology under investigation. - Foster interdisciplinary collaboration.

(Continued)

Table 2. Principles of action to guide future educational technology research (*Continued*)

Principle of Action	Description	VR-Examples	Indicative Implementation
Multiperspective outcomes	Outcome selection should reflect the educational purpose of the technology and may include cognitive, affective, psychomotor, efficiency-related, or behavioral dimensions.	VR-based motor skills training assessed through action precision (psychomotor) and long-term transfer of strategies (behavior).	- Justify outcome selection in relation to the instructional objectives and consider additional dimensions when they are theoretically or practically relevant. - Develop standardized instruments for under-researched outcome dimensions. - Encourage journals to accept studies that report broader sets of outcomes.
Cumulative evidence development	Evidence must be embedded in research programs and consolidated through replication, synthesis, and review.	Meta-analysis consolidating value-added research on multimedia principles in VR.	Provide incentives for replication studies.
Practical relevance	Research must connect to the needs of teachers, universities, and lifelong learning contexts.	Studies developing classroom-feasible VR interventions that account for limited numbers of headsets, e.g., by rotating small groups through VR modules while others work on complementary tasks.	- Involve practitioners in co-design. - Fund practice-oriented pilot studies. - Ensure dissemination through professional development networks.
Transparency and comparability	Uniform reporting standards enable replication, synthesis, and meta-analyses, while open science practices enhance trust.	Publication of all outcome data from a null-effect study on a VR learning application.	- Establish reporting guidelines (e.g., CONSORT 2025 Statement) for educational research. - Promote open data and open materials policies.

For researchers and practitioners working with emerging technologies, the principles summarized in Table 2 can serve as a planning and reflection tool. In a VR project, for example, the framework can guide decisions about whether an initial study should prioritize causal isolation, authentic classroom implementation, the analysis of instructional features, or differences among learner groups. It can also support decisions about whether cognitive outcomes alone are sufficient or whether psychomotor performance, motivation, transfer behavior, and implementation efficiency are needed to evaluate the educational purpose of the application. The framework therefore translates methodological considerations into concrete decisions regarding study design, technology implementation, and evidence interpretation. More broadly, the principles provide a structured orientation for advancing educational technology research by showing how methodological rigor, diverse research approaches, comprehensive outcome perspectives, cumulative evidence, practical relevance, and transparency can be integrated into coherent research programs.

5 DISCUSSION

The integrative framework provides a structured lens for examining research on emerging educational technologies by linking internal and external validity, research approaches, and outcome dimensions. Using VR as a paradigmatic application context, the framework demonstrates how methodological choices shape the types of claims that can be made about technological effectiveness, instructional

value, learner-specific conditions, and practical implementation. Its contribution, therefore, lies not in prescribing a single ideal study design but in supporting the cumulative development of evidence across complementary research scopes.

5.1 Strengths of the approach

A major strength of the proposed framework lies in its systematic integration of previously separate research traditions. Rather than treating media comparison, value-added, and ATI/LTI approaches as isolated paradigms, the framework conceptualizes them as complementary components of a cumulative research process. This perspective makes explicit how validity considerations and outcome selection interact, thereby providing researchers with a structured orientation. By extending the outcome spectrum (i.e., cognition, affect, psychomotor skills, and efficiency) to include a behavioral dimension, the framework highlights long-term and practically relevant learning effects that have often been overlooked in previous research.

5.2 Limitations and simplifications

Several limitations should be acknowledged. First, the framework constitutes a conceptual and methodological synthesis rather than an empirically validated measurement or causal model. The present article therefore cannot establish its feasibility or utility through empirical data. Future research should examine how reliably researchers can use the framework to classify studies, align research questions with designs and outcomes, and identify gaps within cumulative research programs. Its applicability should also be evaluated across different emerging technologies, educational sectors, and implementation contexts. Second, the focus on validity, research approaches, and outcomes necessarily abstracts from other influential dimensions, such as policy, technological innovation cycles, and organizational constraints [64]. Third, VR is used as an illustrative case, which may limit the generalizability of the framework. Although the framework is intended to be transferable to other technologies, such as AI-enhanced learning systems, this transferability remains to be demonstrated empirically. Finally, the proposed principles imply institutional changes, including incentives for replication studies and the systematic adoption of open science practices, that are not yet widely established in the field.

5.3 Implications for research and practice

The framework has several implications for research and practice. The principle of complementary validity highlights that rigorous laboratory experiments and authentic field trials should be conceived as sequential rather than competing stages. Across cumulative research programs, media comparison, value-added, and ATI/LTI studies can generate complementary evidence on whether, how, under which conditions, and for whom a technology is educationally effective. Individual studies may legitimately focus on only one of these questions, provided that their scope and evidential claims are clearly defined. Equally important is the broadening of outcome measures. Cognitive, affective, psychomotor, efficiency-related,

and behavioral indicators are needed to capture the complex effects of educational technologies.

Moreover, the framework underscores the importance of cumulative evidence development. Meta-analyses and replication studies are essential for building a robust evidence base. It also emphasizes practical relevance, transparency, and comparability. Uniform reporting standards and open science practices, including open data, open materials, and the reporting of null findings, are crucial for replicability and cumulative knowledge building.

5.4 Future directions

The framework serves as a starting point for a more collaborative and cumulative research culture in educational technology. Rather than producing an increasing number of isolated studies on individual tools or effects, future research should coordinate study designs, share data and instruments, and test theoretically grounded interventions across diverse contexts and learner groups. This would enable not only stronger causal claims but also richer explanations of the mechanisms and conditions underlying effectiveness. For VR and other emerging technologies, this requires moving beyond novelty-driven evaluations toward longitudinal and cross-institutional research programs that combine laboratory precision with classroom authenticity. By establishing such research infrastructures, the field can generate evidence that is rigorous, transferable, and responsive to rapid technological change. In this way, educational technology research can progress from fragmented experimentation toward a genuinely cumulative science capable of informing innovation and practice.

6 REFERENCES

- [1] S. Timotheou *et al.*, “Impacts of digital technologies on education,” *Education and Information Technologies*, vol. 28, pp. 6695–6726, 2023. <https://doi.org/10.1007/s10639-022-11431-8>
- [2] K. Facer and N. Selwyn, *Digital Technology and the Futures of Education: Towards ‘Non-Stupid’ Optimism*. UNESCO, 2021. Available: <https://unesdoc.unesco.org/ark:/48223/pf0000377071>
- [3] R. M. Tamim, R. M. Bernard, E. Borokhovski, P. C. Abrami, and R. F. Schmid, “What forty years of research says about the impact of technology on learning: A second-order meta-analysis and validation study,” *Review of Educational Research*, vol. 81, no. 1, pp. 4–28, 2011. <https://doi.org/10.3102/0034654310393361>
- [4] A. C. Cheung and R. E. Slavin, “How features of educational technology applications affect student reading outcomes: A meta-analysis,” *Educational Research Review*, vol. 7, no. 3, pp. 198–215, 2012. <https://doi.org/10.1016/j.edurev.2012.05.002>
- [5] R. E. Clark, “Reconsidering research on learning from media,” *Review of Educational Research*, vol. 53, no. 4, pp. 445–459, 1983. <https://doi.org/10.3102/00346543053004445>
- [6] R. B. Kozma, “Will media influence learning?” *Educational Technology Research and Development*, vol. 42, no. 2, pp. 7–19, 1994. <https://doi.org/10.1007/BF02299087>
- [7] C. Robinson and M. Nathan, “Considerations of learning and learning research,” *Journal of Interactive Learning Research*, vol. 12, no. 1, pp. 69–88, 2001. [Online]. Available: <https://www.learntechlib.org/p/8458>

- [8] J. Radianti, T. A. Majchrzak, J. Fromm, and I. Wohlgenannt, "A systematic review of immersive virtual reality applications for higher education," *Computers & Education*, vol. 147, art. no. 103778, 2020. <https://doi.org/10.1016/j.compedu.2019.103778>
- [9] L. Jensen and F. Konradsen, "A review of VR head-mounted displays in education," *Education and Information Technologies*, vol. 23, no. 4, pp. 1515–1529, 2018. <https://doi.org/10.1007/s10639-017-9676-0>
- [10] D. Hamilton, J. McKechnie, E. Edgerton, and C. Wilson, "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*, vol. 8, pp. 1–32, 2021. <https://doi.org/10.1007/s40692-020-00169-2>
- [11] B. Wu, X. Yu, and X. Gu, "Effectiveness of immersive virtual reality using head-mounted displays," *British Journal of Educational Technology*, vol. 51, no. 6, pp. 1991–2005, 2020. <https://doi.org/10.1111/bjet.13023>
- [12] C. Andrade, "Internal, external, and ecological validity in research design, conduct, and evaluation," *Indian Journal of Psychological Medicine*, vol. 40, no. 5, pp. 498–499, 2018. https://doi.org/10.4103/IJPSYM.IJPSYM_334_18
- [13] W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin, 2002.
- [14] L. J. Cronbach and K. Shapiro, *Designing Evaluations of Educational and Social Programs*. San Francisco, CA: Jossey-Bass, 1982.
- [15] S. S. Wallace, G. Barak, G. Truong, and M. W. Parker, "Hierarchy of evidence within medical literature," *Hospital Pediatrics*, vol. 12, no. 8, pp. 745–750, 2022. <https://doi.org/10.1542/hpeds.2022-006690>
- [16] D. T. Campbell and J. C. Stanley, *Experimental and Quasi-Experimental Designs for Research*. Boston, MA: Ravenio Books, 2015.
- [17] H. S. Bloom, C. J. Hill, A. R. Black, and M. W. Lipsey, "Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions," *Journal of Research on Educational Effectiveness*, vol. 1, no. 4, pp. 289–328, 2008. <https://doi.org/10.1080/19345740802400072>
- [18] J. E. Rockoff, "The impact of individual teachers on student achievement," *American Economic Review*, vol. 94, no. 2, pp. 247–252, 2004. <https://doi.org/10.1257/0002828041302244>
- [19] S. Greenland, J. Pearl, and J. M. Robins, "Confounding and collapsibility in causal inference," *Statistical Science*, vol. 14, no. 1, pp. 29–46, 1999. <https://doi.org/10.1214/ss/1009211805>
- [20] T. J. VanderWeele and I. Shpitser, "On the definition of a confounder," *Annals of Statistics*, vol. 41, no. 1, pp. 196–220, 2013. <https://doi.org/10.1214/12-AOS1058>
- [21] J. Wolfartsberger, R. Zimmermann, G. Obermeier, and D. Niedermayr, "VR-supported training for industrial assembly tasks," *Computers in Industry*, vol. 147, art. no. 103838, 2023. <https://doi.org/10.1016/j.compind.2022.103838>
- [22] M. Mulders, K. H. Träg, and L. Kirner, "Go green: Evaluating an XR application on bio-diversity in German secondary school classrooms," *Instructional Science*, vol. 53, no. 3, pp. 429–455, 2025. <https://doi.org/10.1007/s11251-024-09697-1>
- [23] J. J. Heckman, "Sample selection bias as a specification error," *Econometrica*, vol. 47, no. 1, pp. 153–161, 1979. <https://doi.org/10.2307/1912352>
- [24] J. A. Cranford et al., "Reasons for nonresponse in a web-based survey of alcohol involvement," *Addictive Behaviors*, vol. 33, pp. 206–210, 2008. <https://doi.org/10.1016/j.addbeh.2007.07.008>
- [25] I. A. Stoop, *The Hunt for the Last Respondent*. SCP, 2010.

- [26] A. K. Yancey, A. N. Ortega, and S. K. Kumanyika, "Effective recruitment and retention of minority research participants," *Annual Review of Public Health*, vol. 27, pp. 1–28, 2006. <https://doi.org/10.1146/annurev.publhealth.27.021405.102113>
- [27] S. E. Morgan *et al.*, "Tailored messages for research participation," *Journal of Cancer Education*, vol. 37, no. 1, pp. 16–22, 2022. <https://doi.org/10.1007/s13187-020-01775-5>
- [28] T. C. Peck, K. McMullen, and J. Quarles, "DiVRsify: Break the cycle and Develop VR for Everyone," *arXiv preprint arXiv:2110.00497*, 2021. <https://doi.org/10.48550/arXiv.2110.00497>
- [29] C. Hertzog and J. R. Nesselroade, "Assessing psychological change in adulthood," *Psychology and Aging*, vol. 18, no. 4, pp. 639–657, 2003. <https://doi.org/10.1037/0882-7974.18.4.639>
- [30] P. B. Baltes, H. W. Reese, and J. R. Nesselroade, *Life-Span Developmental Psychology: Introduction to Research Methods*. New York, NY: Psychology Press, 2016.
- [31] P. Pande, A. Thit, A. E. Sørensen, B. Mojsoska, M. E. Moeller, and P. M. Jepsen, "Long-term effectiveness of immersive VR simulations," *Research in Learning Technology*, vol. 29, art. no. 2482, 2021. <https://doi.org/10.25304/rlt.v29.2482>
- [32] J. P. Hausknecht, J. A. Halpert, N. T. Di Paolo, and M. O. Moriarty Gerrard, "Retesting in selection," *Journal of Applied Psychology*, vol. 92, no. 2, pp. 373–385, 2007. <https://doi.org/10.1037/0021-9010.92.2.373>
- [33] H. L. Roediger and A. C. Butler, "Retrieval practice and long-term retention," *Trends in Cognitive Sciences*, vol. 15, no. 1, pp. 20–27, 2011. <https://doi.org/10.1016/j.tics.2010.09.003>
- [34] N. Galor, M. Wilf, and M. Plotnik, "Developing multiple shortened forms of a VR-based color trails test," *Applied Neuropsychology: Adult*, vol. 31, no. 6, pp. 1383–1390, 2022. <https://doi.org/10.1080/23279095.2022.2130067>
- [35] E. L. Hannan, "Randomized clinical trials and observational studies," *JACC: Cardiovascular Interventions*, vol. 1, no. 3, pp. 211–217, 2008. <https://doi.org/10.1016/j.jcin.2008.01.008>
- [36] N. Döring, *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*, 6th ed. Berlin, Germany: Springer, 2023. <https://doi.org/10.1007/978-3-662-64762-2>
- [37] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983. <https://doi.org/10.1093/biomet/70.1.41>
- [38] M. Eid, M. Gollwitzer, and M. Schmitt, *Statistik und Forschungsmethoden*. Beltz, 2017.
- [39] H. Goldstein, *Multilevel Statistical Models*, 4th ed. Chichester, U.K.: Wiley, 2010. <https://doi.org/10.1002/9780470973394>
- [40] H. Holling and B. Schmitz, Eds., *Handbuch Statistik, Methoden und Evaluation*. Göttingen, Germany: Hogrefe, 2010.
- [41] D. Holt and T. M. F. Smith, "Post stratification," *Journal of the Royal Statistical Society: Series A (General)*, vol. 142, no. 1, pp. 33–46, 1979. <https://doi.org/10.2307/2344652>
- [42] J.-C. Deville and C.-E. Särndal, "Calibration estimators in survey sampling," *Journal of the American Statistical Association*, vol. 87, no. 418, pp. 376–382, 1992. <https://doi.org/10.1080/01621459.1992.10475217>
- [43] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge, U.K.: Cambridge University Press, 2000.
- [44] T. J. VanderWeele, "Principles of confounder selection," *European Journal of Epidemiology*, vol. 34, no. 3, pp. 211–219, 2019. <https://doi.org/10.1007/s10654-019-00494-6>
- [45] S. Greenland, "Basic methods for sensitivity analysis of biases," *International Journal of Epidemiology*, vol. 25, no. 6, pp. 1107–1116, 1996. <https://doi.org/10.1093/ije/25.6.1107-a>
- [46] H. R. Banack *et al.*, "Modern sources of controls in case-control studies," *American Journal of Epidemiology*, vol. 194, no. 9, pp. 2631–2640, 2025. <https://doi.org/10.1093/aje/kwae437>
- [47] M. G. Findley, K. Kikuta, and M. Denly, "External validity," *Annual Review of Political Science*, vol. 24, pp. 365–393, 2021. <https://doi.org/10.1146/annurev-polisci-041719-102556>

- [48] T. Dunning, *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge, UK: Cambridge University Press, 2016.
- [49] S. D. Levitt and J. A. List, "On the generalizability of lab behaviour," *Canadian Journal of Economics*, vol. 40, no. 2, pp. 347–370, 2007. <https://doi.org/10.1111/j.1365-2966.2007.00412.x>
- [50] K. Bottema-Beutel, B. P. Lloyd, E. W. Carter, and J. M. Asmus, "Generalizability and decision studies to inform observational and experimental research in classroom settings," *American Journal on Intellectual and Developmental Disabilities*, vol. 119, no. 6, pp. 589–605, 2014. <https://doi.org/10.1352/1944-7558-119.6.589>
- [51] N. G. Holmes, J. Olsen, J. L. Thomas, and C. E. Wieman, "Value added or misattributed?" *Physical Review Physics Education Research*, vol. 13, no. 1, art. no. 010129, 2017. <https://doi.org/10.1103/PhysRevPhysEducRes.13.010129>
- [52] L. Kohnke and D. Fount, "Re-examining factors in AR/VR teacher training," *Educational Technology & Society*, vol. 28, no. 1, pp. 359–370, 2025.
- [53] J. Han, G. Liu, and Q. Zheng, "Prior knowledge as a moderator between signaling and learning performance," *Frontiers in Psychology*, vol. 14, p. 1118174, 2023. <https://doi.org/10.3389/fpsyg.2023.1118174>
- [54] C. C. Lewis *et al.*, "A systematic review of empirical studies examining mechanisms of implementation in health," *Implementation Science*, vol. 15, art. no. 21, 2020. <https://doi.org/10.1186/s13012-020-00983-3>
- [55] A. A. Christ, V. Capon-Sieber, U. Grob, and A. K. Praetorius, "Learning processes and their mediating role between teaching quality and student achievement: A systematic review," *Studies in Educational Evaluation*, vol. 75, art. no. 101209, 2022. <https://doi.org/10.1016/j.stueduc.2022.101209>
- [56] D. B. Clark, E. E. Tanner-Smith, and S. S. Killingsworth, "Digital games, design, and learning: A systematic review and meta-analysis," *Review of Educational Research*, vol. 86, no. 1, pp. 79–122, 2016. <https://doi.org/10.3102/0034654315582065>
- [57] E. Klieme and C. Pauli, "Quality of instruction and long-term learning outcomes," in *The Impact of Teaching Quality on Student Outcomes*, E. Klieme and U. Trautwein, Eds., Springer, 2019.
- [58] G. Makransky and 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*, vol. 33, pp. 937–958, 2021. <https://doi.org/10.1007/s10648-020-09586-2>
- [59] J. Einloft *et al.*, "Immersive medical training: A comprehensive longitudinal study," *BMC Medical Education*, vol. 24, art. no. 978, 2024. <https://doi.org/10.1186/s12909-024-05957-3>
- [60] F. L. Schmidt and J. E. Hunter, *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, 3rd ed. Thousand Oaks, CA, USA: Sage, 2015. <https://doi.org/10.4135/9781483398105>
- [61] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40–79, 2010. <https://doi.org/10.1214/09-SS054>
- [62] T. Takada *et al.*, "Internal-external cross-validation helped to evaluate the generalizability of prediction models in large clustered datasets," *Journal of Clinical Epidemiology*, vol. 137, pp. 83–91, 2021. <https://doi.org/10.1016/j.jclinepi.2021.03.025>
- [63] J. Buchner and M. Kerres, "Media comparison studies dominate comparative research on augmented reality in education," *Computers & Education*, vol. 195, art. no. 104711, 2023. <https://doi.org/10.1016/j.compedu.2022.104711>
- [64] M. Kerres, *Mediendidaktik: Konzeption und Entwicklung digitaler Lernangebote*, 5th ed. Berlin, Germany: De Gruyter, 2018. <https://doi.org/10.1515/9783110456837>
- [65] A. P. Lawson *et al.*, "Confounded or controlled?" *Educational Psychology Review*, vol. 36, 2024. <https://doi.org/10.1007/s10648-024-09908-8>

- [66] J. Buchner and M. Mulders, "Still trapped in media comparison? A systematic review of comparative research on immersive virtual reality in education," *Interactive Learning Environments*, pp. 1–16, 2026. <https://doi.org/10.1080/10494820.2026.2617986>
- [67] R. E. Mayer, "Incorporating motivation into multimedia learning," *Learning and Instruction*, vol. 29, pp. 171–173, 2014. <https://doi.org/10.1016/j.learninstruc.2013.04.003>
- [68] R. Moreno and R. E. Mayer, "Interactive multimodal learning environments," *Educational Psychology Review*, vol. 19, no. 3, pp. 309–326, 2007. <https://doi.org/10.1007/s10648-007-9047-2>
- [69] S. Klingenberg, M. L. M. Jørgensen, G. Dandanell, K. Skriver, A. Mottelson, and G. Makransky, "Teaching as a generative strategy in VR," *British Journal of Educational Technology*, vol. 51, pp. 2115–2138, 2020. <https://doi.org/10.1111/bjet.13029>
- [70] L. J. Cronbach and R. E. Snow, *Aptitudes and Instructional Methods: A Handbook for Research on Interactions*. New York, NY: Irvington, 1977.
- [71] C. J. Chen, S. C. Toh, and W. M. F. W. Ismail, "Are learning styles relevant to virtual reality?" *Journal of Research on Technology in Education*, vol. 38, no. 2, pp. 123–141, 2005. <https://doi.org/10.1080/15391523.2005.10782453>
- [72] T. C. Reeves and E. G. Oh, "The goals and methods of educational technology research over a quarter century (1989–2014)," *Educational Technology Research and Development*, vol. 65, no. 2, pp. 325–339, 2017. <https://doi.org/10.1007/s11423-016-9474-1>
- [73] L. A. Schindler, G. J. Burkholder, O. A. Morad, and C. Marsh, "Computer-based technology and student engagement: A critical review of the literature," *International Journal of Educational Technology in Higher Education*, vol. 14, art. no. 25, 2017. <https://doi.org/10.1186/s41239-017-0063-0>
- [74] D. Hillmayr, L. Ziernwald, F. Reinhold, S. I. Hofer, and K. M. Reiss, "The potential of digital tools to enhance mathematics and science learning in secondary schools: A context-specific meta-analysis," *Computers & Education*, vol. 153, art. no. 103897, 2020. <https://doi.org/10.1016/j.compedu.2020.103897>
- [75] R. M. Bernard *et al.*, "How does distance education compare with classroom instruction? A meta-analysis of the empirical literature," *Review of Educational Research*, vol. 74, no. 3, pp. 379–439, 2004. <https://doi.org/10.3102/00346543074003379>
- [76] R. E. Slavin, "Evidence-based education policies," *Educational Researcher*, vol. 31, no. 7, pp. 15–21, 2002. <https://doi.org/10.3102/0013189X031007015>
- [77] G. Makransky, L. Lilleholt, and A. Aaby, "Development and validation of the multimodal presence scale," *Computers in Human Behavior*, vol. 72, pp. 276–285, 2017. <https://doi.org/10.1016/j.chb.2017.02.066>
- [78] M. D. Merrill, "First principles of instructional design," *Educational Technology Research and Development*, vol. 50, no. 3, pp. 43–59, 2002. <https://doi.org/10.1007/BF02505024>
- [79] Design-Based Research Collective, "Design-based research: An emerging paradigm for educational inquiry," *Educational Researcher*, vol. 32, no. 1, pp. 5–8, 2003. <https://doi.org/10.3102/0013189X032001005>
- [80] R. C. Clark and R. E. Mayer, *e-Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning*, 4th ed. Hoboken, NJ, USA: Wiley, 2016. <https://doi.org/10.1002/9781119239086>
- [81] M. Mulders, J. Buchner, and M. Kerres, "A framework for the use of immersive virtual reality in learning environments," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 15, no. 24, pp. 208–224, 2020. <https://doi.org/10.3991/ijet.v15i24.16615>
- [82] O. A. Meyer, M. K. Omdahl, and G. Makransky, "Investigating the effect of pre-training when learning through immersive virtual reality and video: A media and methods experiment," *Computers & Education*, vol. 140, art. no. 103603, 2019. <https://doi.org/10.1016/j.compedu.2019.103603>

- [83] S. Klingenberg, R. Fischer, I. Zettler, and G. Makransky, "Facilitating learning in immersive VR," *Journal of Computer Assisted Learning*, vol. 39, no. 1, pp. 218–230, 2023. <https://doi.org/10.1111/jcal.12741>
- [84] J. Parong and R. E. Mayer, "Learning science in immersive virtual reality," *Journal of Educational Psychology*, vol. 110, no. 6, pp. 785–797, 2018. <https://doi.org/10.1037/edu0000241>
- [85] G. Makransky, N. K. Andreassen, S. Baceviciute, and R. E. Mayer, "Immersive virtual reality increases liking but not learning with a science simulation and generative learning strategies promote learning in immersive virtual reality," *Journal of Educational Psychology*, vol. 113, no. 4, pp. 719–735, 2021. <https://doi.org/10.1037/edu0000473>
- [86] D. Gašević, S. Dawson, T. Rogers, and D. Gasevic, "Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success," *The Internet and Higher Education*, vol. 28, pp. 68–84, 2016. <https://doi.org/10.1016/j.iheduc.2015.10.002>
- [87] C. I. Johnson, S. K. T. Bailey, B. L. Schroeder, and M. D. Marraffino, "Procedural learning in VR," *Technology, Mind, and Behavior*, vol. 3, no. 4, pp. 551–565, 2022. <https://doi.org/10.1037/tmb0000087>
- [88] S. Jin *et al.*, "Guided inquiry and prior knowledge in immersive VR environments," *Education and Information Technologies*, vol. 29, pp. 22447–22473, 2024. <https://doi.org/10.1007/s10639-024-12719-7>
- [89] Z. Taçgın, "The perceived effectiveness of immersive VR depending on prior knowledge," *Education and Information Technologies*, vol. 25, pp. 2791–2809, 2020. <https://doi.org/10.1007/s10639-019-10088-0>
- [90] B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill, and D. R. Krathwohl, *Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I: Cognitive Domain*. New York, NY, USA: David McKay, 1956.
- [91] J. Buchner and M. Mulders, "Ein Rahmenmodell für bedeutsame Forschung zu Bildungstechnologien," in *5th EdTech Research Forum and 1st Conference of the DGfE Working Group on Media Didactics*, University of Duisburg-Essen, Essen, Germany, 2024. [Online]. Available: <https://learninglab.uni-due.de/veranstaltung/15637>
- [92] E. R. Kahu and K. Nelson, "Student engagement in the educational interface," *Higher Education Research & Development*, vol. 37, no. 1, pp. 58–71, 2017. <https://doi.org/10.1080/07294360.2017.1344197>
- [93] M. T. H. Chi and R. Wylie, "The ICAP framework: Linking cognitive engagement to active learning outcomes," *Educational Psychologist*, vol. 49, no. 4, pp. 219–243, 2014. <https://doi.org/10.1080/00461520.2014.965823>
- [94] S. Chen, "The construct of cognitive load in interpreting and its measurement," *Perspectives*, vol. 25, no. 4, pp. 640–657, 2017. <https://doi.org/10.1080/0907676X.2016.1278026>
- [95] B. Spinath, J. Stiensmeier-Pelster, D. Schöne, and O. Dickhäuser, *SELLMO*. Hogrefe, 2002.
- [96] J. M. Keller, *Motivational Design for Learning and Performance: The ARCS Model Approach*. Springer, 2010. <https://doi.org/10.1007/978-1-4419-1250-3>
- [97] B. Breyer and M. Bluemke, "Deutsche Version der Positive and Negative Affect Schedule PANAS (GESIS Panel)," *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*, 2016. <https://doi.org/10.6102/zis242>
- [98] F. de Paula Forero Pataquiva and B. Klimova, "A systematic review of virtual reality in the acquisition of second language," *International Journal of Emerging Technologies in Learning (ijET)*, vol. 17, no. 15, pp. 43–53, 2022. <https://doi.org/10.3991/ijet.v17i15.31781>
- [99] A. Ashley-Welbeck and D. Vlachopoulos, "Teachers' perceptions on using augmented reality for language learning in Primary Years Programme (PYP) education," *International Journal of Emerging Technologies in Learning (ijET)*, vol. 15, no. 12, pp. 116–135, 2020. <https://doi.org/10.3991/ijet.v15i12.13499>

- [100] M. Sailer, R. Maier, S. Berger, T. Kastorff, and K. Stegmann, "Learning activities in technology-enhanced learning," *Learning and Individual Differences*, vol. 112, art. no. 102446, 2024. <https://doi.org/10.1016/j.lindif.2024.102446>
- [101] J. L. Plass and S. Pawar, "Toward a taxonomy of adaptivity for learning," *Journal of Research on Technology in Education*, vol. 52, no. 3, pp. 275–300, 2020. <https://doi.org/10.1080/15391523.2020.1719943>

7 AUTHORS

Miriam Mulders is a postdoctoral researcher at the Learning Lab of the University of Duisburg-Essen, Germany. Her research focuses on the theory, investigation, and design of learning environments supported by virtual and augmented reality (E-mail: miriam.mulders@uni-due.de).

Katja Buntins is a postdoctoral researcher at the Learning Lab of the University of Duisburg-Essen, Germany. Her research focuses on systematic reviews, research methodology, and synthesizing evidence in educational research (E-mail: katja.buntins@uni-due.de).