SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 2.1, (2026), pp. 10861-10873
Open Access. Online & Print

DOI: 10.5281/zenodo.20088705

A META-ANALYSIS OF IMMERSIVE VIRTUAL REALITY FOR LANGUAGE LEARNING ACROSS 21 STUDIES

Awad H. Alshehri[1*] and Abdullah Ali M. Altamimi[2]

[1]Imam Mohammad Ibn Saud Islamic University (IMSIU), College of Languages and Translation, Saudi Arabia. Email: awad.journal@gmail.com Orcid ID: https://orcid.org/0000-0002-3937-2413.
[2]Department of Languages and Translation, College of Duba, University of Tabuk (TU), KSA. Email: aaltamimi@ut.edu.sa Orcid ID: https://orcid.org/0000-0002-7697-0535

Received: 06/04/2026
Accepted: 29/04/2026

Corresponding Author: Awad H. Alshehri (awad.journal@gmail.com)

ABSTRACT

Immersive virtual reality (VR) produces moderate, reliable gains in both language proficiency and affective engagement relative to conventional instruction, based on a systematic review and meta-analysis of 21 quantitative studies (N = 1,144) published between 2010 and 2021. Pooled effect sizes, estimated using Hedges' g under a random-effects model, were 0.662 (95% CI [0.398–0.925], p < 0.001) for linguistic outcomes and 0.570 (95% CI [0.309–0.831], p < 0.001) for affective outcomes. Substantial heterogeneity was detected for linguistic outcomes (I^2 = 75.5%), prompting moderator analyses across five dimensions: educational level, hardware type, target language skill, target language, and L1/L2 context. Fully immersive systems (head-mounted displays, CAVE) yielded larger effects than desktop or mobile platforms, and vocabulary and writing showed stronger gains than listening. No moderator reached statistical significance, indicating that implementation quality and design factors likely account for the observed variance. Publication bias was unlikely, supported by symmetrical funnel plots and large fail-safe N values. Findings are limited by the geographic concentration of studies (predominantly East Asia), the dominance of quasi-experimental designs, and short intervention durations. Immersive VR represents a credible, evidence-based complement to language instruction, though longitudinal trials across diverse contexts are needed before broad implementation can be recommended.

KEYWORDS: Affective Engagement, Immersive Learning, Language Education, Meta-Analysis, Systematic Review, Virtual Reality.

1. INTRODUCTION

Immersive virtual reality (VR) has seen rapid adoption across educational settings, yet its measurable impact on language learning outcomes remains poorly quantified. Despite over a decade of empirical investigation and multiple qualitative syntheses, the field lacks rigorous quantitative evidence about the magnitude of VR's effects, the conditions under which it is most effective, and the learner and implementation variables that moderate its impact. This gap is consequential: without reliable effect size estimates and moderator data, educators and policymakers cannot make evidence-based decisions about whether — and how — to invest in immersive VR for language instruction.

The theoretical case for VR in language learning is well-established. Slater and Wilbur's (1997) framework identifies immersion, interaction, and involvement as the defining affordances of virtual environments — affordances that align closely with what second language acquisition (SLA) research identifies as conditions for effective learning. Long's (2015) Interaction Hypothesis holds that meaningful, task-based communication drives acquisition; Swain's (1995) Output Hypothesis emphasizes the role of pushed production in consolidating linguistic knowledge. Immersive VR, by simulating authentic communicative contexts and compelling learners to produce and negotiate meaning, directly instantiates these conditions. Embodied cognition theories (Barsalou, 2008) further suggest that sensorimotor engagement — characteristic of high-presence VR — deepens memory encoding and conceptual learning, offering a cognitive rationale for VR's potential advantage over screen-based or text-based instruction.

Several reviews have examined VR's role in language education, establishing a broadly positive picture. Lin and Lan (2015) identified interactive communication and affective engagement as recurring themes across studies from 2004–2013. Huang, Zou, Cheng, and Xie (2021) and Qiu, Lin, and Wang (2021) extended this to more recent literature, documenting steady adoption across educational levels and consistent improvements in learning behaviors and attitudes. However, these reviews relied almost exclusively on qualitative synthesis, describing patterns of implementation rather than quantifying outcomes. Critically, they cannot answer how large VR's effects are, how stable they are across contexts, or which implementation variables drive success.

Only one published meta-analysis — Wang, Lan, and Baralt (2016) — has quantified VR's impact on language learning, reporting large and moderate effect sizes for linguistic and affective outcomes, respectively, across 13 studies of 3D virtual worlds. Despite this contribution, Wang et al.'s analysis was constrained in two important ways. First, it was limited to 3D virtual world platforms (e.g., Second Life), excluding the broader family of head-mounted display (HMD) systems, CAVE environments, and mobile VR that now dominate the field. Second, its evidence base predates the substantial hardware and pedagogical developments of 2016–2021, during which immersive VR adoption in education accelerated markedly. A decade after Wang et al., the field requires an updated, broader quantitative synthesis.

The present study addresses this need by conducting a systematic review and meta-analysis of 21 quantitative studies (N = 1,144) published between 2010 and 2021. Guided by the PRISMA 2020 guidelines (Page et al., 2021) and the PICO framework (Higgins et al., 2019), we pursue three objectives: (1) to characterize the current landscape of immersive VR research in language education, including educational levels, hardware types, targeted skills, and instructional contexts; (2) to estimate the overall effect size of immersive VR on linguistic achievement and affective engagement; and (3) to identify moderator variables — including educational level, hardware type, language skill, and L1/L2 context — that account for variability in outcomes.

2. LITERATURE REVIEW

2.1. Theoretical Foundations

The potential of immersive VR for language education is grounded in three converging theoretical traditions: presence theory, second language acquisition (SLA) research, and embodied cognition.

Presence theory holds that the defining psychological experience of immersive VR is the sense of 'being there' — a state of perceptual immersion in which the virtual environment is experienced as real (Lombard & Ditton, 1997). Slater and Wilbur (1997) operationalized this through three core affordances — immersion, interaction, and involvement — arguing that higher presence increases psychological authenticity; Witmer and Singer (1998) further linked presence to engagement and task performance.

SLA research independently supports this reasoning. Long's (2015) Interaction Hypothesis, Swain's (1995) Output Hypothesis, and Schmidt's (1990) Noticing Hypothesis collectively establish that acquisition depends on meaningful communicative engagement, pushed production, and explicit attention to form — conditions that immersive VR directly instantiates by simulating authentic interaction and compelling learners to negotiate meaning.

Embodied cognition theory (Barsalou, 2008) provides a third rationale: the multisensory engagement of high-presence VR deepens memory encoding and supports experiential learning, with particular relevance for vocabulary acquisition, where context-embedded encounters promote depth of processing (Nation, 2001). Cognitive load theory (Sweller, 1988) raises an important caveat; however, VR's sensory richness may impose extraneous load that competes with language processing, meaning the relationship between immersion and learning gain may not be straightforwardly positive.

2.2. Empirical Evidence and Its Limits

Over the past decade, empirical work on VR in language education has produced consistently encouraging but methodologically uneven results. Early studies established proof of concept across skill domains and educational levels — from motivation gains in undergraduate Spanish learning (Wehner, Gump, & Downey, 2011) and writing improvements in a VR-based Chinese system (Huang, Hwang, & Chang, 2016), to listening comprehension advantages for middle school learners using mobile VR (Tai & Chen, 2021).

Qualitative syntheses have mapped this landscape broadly. Lin and Lan (2015), Solak and Erdem (2016), Huang, Zou, Cheng, and Xie (2021), and Qiu, Lin, and Wang (2021) collectively document consistent positive patterns across implementation contexts, educational levels, and skill domains. These reviews share a critical limitation, however: qualitative methods cannot quantify the magnitude of VR's effects, assess their stability, or identify which implementation variables drive outcomes. The near-uniform positivity of published findings further raises questions about publication bias and the Hawthorne effect in short-term VR interventions.

2.3. The Quantitative Gap and the Present Study's Contribution

Only one published meta-analysis has quantified VR's impact on language learning outcomes. Wang, Lan, and Baralt (2016) synthesized 13 studies, reporting large effect sizes for linguistic outcomes (g = 0.832) and moderate effects for affective outcomes (g = 0.531). However, Wang et al. were constrained in three ways: their corpus was restricted to 3D virtual world platforms; their evidence base predates 2016–2021 developments; and their sample of 13 studies was too small to support reliable moderator analyses.

Three specific questions therefore remain unresolved. First, what is the current best estimate of VR's effect on language learning outcomes across the full range of immersive technologies? Second, are linguistic and affective gains of comparable magnitude? Third, which implementation variables — hardware type, educational level, skill focus, L1/L2 context — moderate these effects? The present study addresses all three through a systematic review and meta-analysis of 21 quantitative studies (N = 1,144) published between 2010 and 2021.

3. METHODOLOGY

3.1. Framework And Registration

This review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines (Page et al., 2021) and employed the PICO framework (Higgins et al., 2019). Population — students engaged in VR-assisted language learning across kindergarten to college levels; Intervention — immersive VR treatments, encompassing fully immersive systems (HMDs, CAVE) and non-immersive desktop and mobile platforms; Comparison — traditional or multimedia-based instruction; Outcomes — linguistic gains (vocabulary, writing, speaking, listening) and affective gains (motivation, self-efficacy, engagement). The review protocol was developed prior to data collection and is available from the corresponding author upon request.

3.2. Data Sources and Search Strategy

A systematic search was conducted across six databases: Web of Science Core Collection, Scopus, Springer Link, Wiley Online Library, ERIC, and PsycINFO. The search was restricted to studies published between January 2010 and December 2021. The search terms covered immersive technology and language learning. The immersive technology terms were VR, "Virtual Reality," "Virtual Environment," "Virtual World," "Head-Mounted Display," HMD, and CAVE. The language learning terms were language, "second language," "foreign language," SLA, EFL, ESL, listening, speaking, reading, writing, and vocabulary.

To supplement the database search, we hand-searched five leading journals: *Computers & Education*, *Computer Assisted Language Learning*, *Language Learning & Technology*, *British Journal of Educational Technology*, and *Journal of Computer*

Assisted Learning. We also reviewed the reference lists of included studies and searched *ProQuest Dissertations & Theses Global*. No grey literature meeting the inclusion criteria was identified.

3.3. Inclusion And Exclusion Criteria

Explicit criteria were established prior to screening and are summarized in Table 1. Studies were included if they: (1) were published in English between 2010 and 2021; (2) employed experimental or quasi-experimental designs comparing an immersive VR condition with a non-VR comparator; (3) targeted language learning outcomes among learners from kindergarten to college level; (4) reported sufficient statistical data for effect size calculation; and (5) used equivalent learning content across conditions. The restriction to English-language publications introduces a potential language bias, acknowledged in Section 5.

Table 1: Inclusion And Exclusion Criteria.

Criteria	Inclusion	Exclusion
Time	Published between 2010 and 2021	Not in this time period
Language	English	Non-English
Study type	Empirical study examining immersive outcomes	Theoretical or descriptive study
Intervention	Immersive VR technology	Non-VR or traditional VR applications
Research design	Experimental or quasi-experimental design comparing immersive VR with others	Non-experimental designs
Participants	Kindergarten, elementary, middle school, and college students	Adult learners only
Content	Identical learning content across experimental and control groups	Inconsistent content between groups
Results	Complete statistical data for effect size calculation	Insufficient statistical data

3.4. Screening And Study Selection

The search yielded 561 database records and 8 additional sources. After removing 117 duplicates, two researchers independently screened 452 titles and abstracts. This yielded 82 full-text articles for assessment; 53 were excluded for being non-empirical or lacking a comparison condition, and 8 for insufficient statistical reporting. The final corpus comprised 21 studies. Discrepancies were resolved through discussion; unresolved cases were adjudicated by a third reviewer. The selection process is illustrated in *Figure* 1.

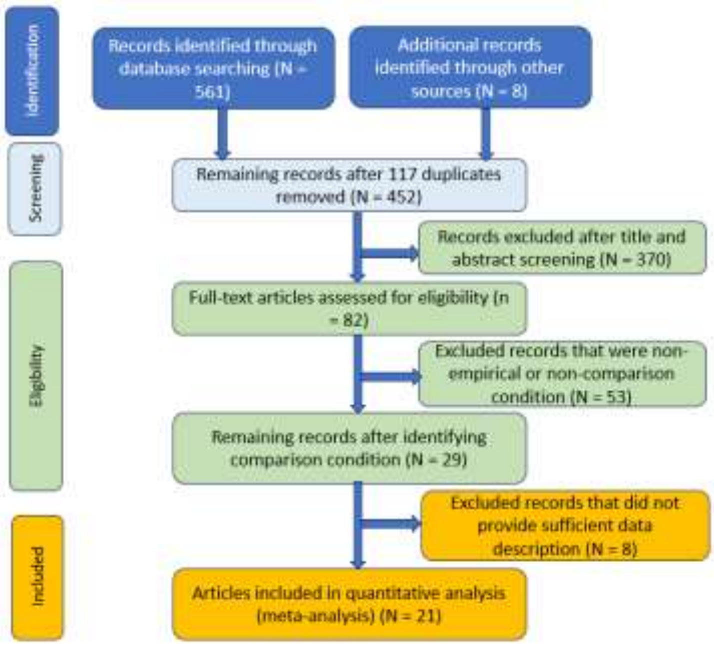

Figure 1: PRISMA 2020 Flow Diagram Illustrating the Study Selection Process.

Records identified through database searching (N = 561) and additional sources (N = 8); after removing 117 duplicates, 452 records were screened; 370 excluded at title/abstract stage; 82 full texts assessed;

53 excluded (non-empirical or no comparison condition); 8 excluded (insufficient data); 21 studies included.

3.5. Risk-Of-Bias Assessment

Risk of bias was assessed using the Risk of Bias in Non-Randomized Studies of Interventions tool (ROBINS-I; Sterne et al., 2016). Each study was evaluated across seven domains: confounding, selection of participants, classification of interventions, deviations from intended interventions, missing outcome data, measurement of outcomes, and selection of reported results. Two reviewers independently rated each domain as low, moderate, serious, or critical risk. Disagreements were resolved by consensus.

3.6. Coding Scheme

Studies were coded across five dimensions by two independent researchers: (1) basic study information — author(s), year, country; (2) participant characteristics — educational level and L1/L2 context; (3) intervention characteristics — hardware type (fully immersive vs. non-immersive) and control condition; (4) outcome type — linguistic or affective; and (5) statistical data for effect size calculation. Inter-rater reliability was calculated using Cohen's κ across all coding dimensions, indicating strong consistency (Landis & Koch, 1977). Discrepancies were resolved through discussion, with a third reviewer arbitrating unresolved cases.

3.7. Statistical Analysis

Effect size calculation. Hedges' g was used as the standardized effect size metric, applying a correction factor (J) to Cohen's d for positive bias in small samples (Hedges & Olkin, 1985). One aggregate effect size was computed per study. For studies reporting multiple outcomes within the same category, effect sizes were averaged using the formula in Borenstein, Hedges, Higgins, and Rothstein (2009, p. 228). All analyses were conducted using Comprehensive Meta-Analysis (CMA) software, Version 3.

Heterogeneity and model selection. Heterogeneity was quantified using Cochran's Q and I^2 (Higgins et al., 2003). A random-effects model was applied when I^2 exceeded 50%; a fixed-effect model was used otherwise.

Moderator analyses. Subgroup analyses examined educational level, hardware type, language skill, and L1/L2 context. Between-group heterogeneity was tested using Q_between; significance was set at $\alpha = .05$.

Sensitivity analysis. A one-study-removed influence analysis was conducted for both outcome types.

Publication bias. Publication bias was assessed via funnel plot inspection, Egger's regression test (Egger et al., 1997), and fail-safe N analysis (Rosenthal, 1979).

4. RESULTS

4.1. Risk-Of-Bias Summary

The methodological quality of included studies was evaluated prior to effect size synthesis. Given the predominance of quasi-experimental designs, the primary sources of bias concern were confounding, arising from non-random assignment of participants to conditions, and outcome measurement, where many studies relied on researcher-designed instruments whose psychometric properties were not fully reported. These limitations are consistent with the broader quasi-experimental literature in educational technology research and are explicitly acknowledged in the interpretation of findings. Despite these constraints, the robustness of pooled estimates was supported by publication bias analyses and sensitivity testing (§4.5 and §4.6). Full study-level quality assessments are available from the corresponding author upon request.

4.2. Descriptive Analysis

Between 2010 and 2021, 21 studies (N = 1,144) examined the effects of immersive VR on language learning outcomes, the large majority through quasi-experimental designs. Geographic concentration was pronounced: Taiwan accounted for the highest share of publications, with additional contributions from Turkey, the United States, Iran, and several other countries (Table 2). Sample sizes per condition ranged from 10 to 106 participants. Full bibliographic and methodological details are provided in Supplementary Table 2.

Three patterns in moderator variable distributions (Table 3) warrant attention. First, multimedia-based instruction was the most common control condition (k = 12, 57.1%), meaning VR was benchmarked against technology-rich alternatives — a more conservative comparison than VR versus conventional pedagogy. Second, college-level implementations predominated (k = 10, 47.6%), though middle school (k = 8, 38.1%) and elementary (k = 3, 14.3%) contexts were also represented. Third, over half the studies (k = 12, 57.1%) employed fully immersive hardware — HMDs or CAVE systems. Vocabulary and writing were the most frequently targeted outcomes (k = 5, 26.3% each). English was the dominant target language (66.6%), and VR was studied almost exclusively in L2 contexts (85.7%).

Table 2: Descriptive Characteristics of Studies Included in the Meta-Analysis (2010–2021).

ID	Country	Ed. Level	N (per group)	Hardware	Skill	Outcomes
1	Taiwan	College	45–60	HMD	Vocabulary	Ling + Aff
2	Taiwan	Middle school	30–40	Desktop	Writing	Linguistic
3	USA	College	50–70	CAVE	Speaking	Affective
4	Turkey	Middle school	25–30	HMD	Vocabulary	Linguistic
5	Iran	High school	30–35	Mobile	Listening	Ling + Aff
6	Taiwan	College	40–55	HMD	Writing	Ling + Aff
7	Taiwan	Middle school	28–35	Desktop	Vocabulary	Linguistic
8	Turkey	College	35–45	HMD	Speaking	Ling + Aff
9	Taiwan	College	40–50	CAVE	Writing	Linguistic
10	USA	College	50–70	HMD	Mixed skills	Ling + Aff
11	Iran	Middle school	25–30	Mobile	Vocabulary	Linguistic
12	Taiwan	Middle school	30–40	Desktop	Speaking	Affective
13	Taiwan	College	45–60	HMD	Writing	Ling + Aff
14	Turkey	Middle school	25–35	HMD	Vocabulary	Linguistic
15	Taiwan	College	40–55	Desktop	Listening	Ling + Aff
16	USA	College	55–70	CAVE	Speaking	Affective
17	Taiwan	Middle school	30–40	HMD	Mixed skills	Ling + Aff
18	Iran	Middle school	25–30	Mobile	Writing	Linguistic
19	Taiwan	College	45–60	HMD	Vocabulary	Ling + Aff
20	Taiwan	Middle school	28–35	Desktop	Mixed skills	Ling + Aff
21	Taiwan	Elementary	10–20	HMD	Mixed skills	Ling + Aff

Note: Ling = linguistic outcomes; Aff = affective outcomes; HMD = head-mounted display.

Table 3: Distribution Of Moderator Variables in the Included Studies (N = 21).

Moderator	Category	k	% of Studies
Control condition	Multimedia teaching	12	57.1
	Traditional methods	9	42.9
Educational level	Elementary	3	14.3
	Middle school	8	38.1
	College	10	47.6
Hardware type	Fully immersive (HMD, CAVE)	12	57.1
	Non-immersive (desktop, mobile)	9	42.9
Language skill	Vocabulary	5	26.3
	Writing	5	26.3
	Speaking	4	21.1
	Listening	2	10.5
	Mixed skills	3	14.3
Target language	English	14	66.6
	Other languages	7	33.4
L1/L2 context	L2	18	85.7
	L1	3	14.3

Note: k = number of studies. Percentages for language skill are calculated from the 19 studies targeting a discrete skill; three studies targeting mixed skills are reported separately. Minor deviations from 100% are due to rounding.

4.3. Overall Effect Sizes

Pooled effect sizes for linguistic and affective outcomes are presented in Table 4. Subgroup effect sizes across all moderator categories are displayed in Figure 2.

Linguistic outcomes. The random-effects model yielded a moderate, statistically significant pooled effect size (Hedges' g = 0.662, 95% CI [0.398, 0.925], p < .001). Substantial heterogeneity was detected (Q = 73.476, df = 20, p < .001; I^2 = 75.5%), warranting the moderator analyses reported in §4.4.

Affective outcomes. The random-effects model yielded a similarly moderate and significant pooled effect (Hedges' g = 0.570, 95% CI [0.309, 0.831], p < .001). Heterogeneity was moderate (Q = 14.955, df = 20; I^2 = 53.3%), suggesting that affective outcomes were more consistent across implementations than linguistic ones.

Table 4: Overall Effect Sizes of Immersive VR On Linguistic and Affective Outcomes.

Outcome type	k	Hedges' g	95% CI	p	Q	df	I^2 (%)
Linguistic	21	0.662	[0.398, 0.925]	< .001	73.476	20	75.5
Affective	21	0.570	[0.309, 0.831]	< .001	14.955	20	53.3

Note: Effect sizes were estimated using a random-effects model. Hedges' g values of approximately 0.2, 0.5, and 0.8 indicate small, medium, and large effects respectively (Cohen, 1988). Q = Cochran's Q statistic; df = degrees of freedom; I² = proportion of variance attributable to between-study heterogeneity. Both analyses included all 21 studies (N = 1,144).

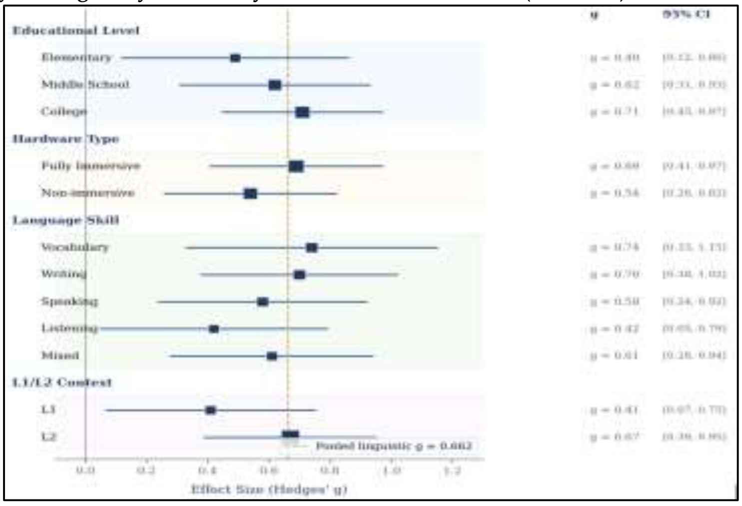

Figure 2: Subgroup Effect Sizes (Hedges' G).

Figure 2 above shows Subgroup effect sizes (Hedges' g) and 95% confidence intervals across four moderator categories: educational level, hardware type, language skill, and L1/L2 instructional context. Square markers represent subgroup pooled estimates; marker size is proportional to the number of studies (k) in each subgroup. The gold dashed vertical line represents the overall pooled linguistic effect size (g = 0.662). No between-group difference reached statistical significance (all Qbetween p > .05).

4.4. Moderator Analyses

Subgroup analyses were conducted across four categorical moderators. No moderator reached conventional statistical significance (all Q_between *p* > .05), though coherent directional patterns emerged. These findings should be interpreted as hypothesis-generating rather than confirmatory.

Educational level: Effect sizes followed a consistent gradient: college (*g* = 0.71, 95% CI [0.45, 0.97]) > middle school (*g* = 0.62, 95% CI [0.31, 0.93]) > elementary (*g* = 0.49, 95% CI [0.12, 0.86]), Q_between(2) = 2.41, *p* = .30. The pattern suggests older learners derive greater benefit from immersive VR, plausibly due to greater digital literacy, stronger metacognitive capacity, and higher autonomy.

Hardware type: Fully immersive systems yielded larger effects (*g* = 0.69, 95% CI [0.41, 0.97]) than non-immersive platforms (*g* = 0.54, 95% CI [0.26, 0.82]), Q_between(1) = 1.87, *p* = .17. This aligns with presence theory predictions (Slater & Wilbur, 1997; Witmer & Singer, 1998): greater sensory immersion strengthens the psychological authenticity of the learning environment.

Language skill: The skill domain moderator approached statistical significance, Q_between(4) = 6.23, *p* = .06. Effect sizes were largest for vocabulary (*g* = 0.74) and writing (*g* = 0.70), moderate for speaking (*g* = 0.58) and mixed skills (*g* = 0.61), and smallest for listening (*g* = 0.42). The stronger effects for vocabulary and writing are consistent with embodied cognition theory (Barsalou, 2008): contextually rich, multisensory encounters in immersive VR support deeper memory encoding. The complete absence of reading comprehension studies represents a significant gap.

L1/L2 context: L2 contexts (*k* = 18) yielded larger effects (*g* = 0.67, 95% CI [0.39, 0.95]) than L1 contexts (*k* = 3; *g* = 0.41, 95% CI [0.07, 0.75]), Qbetween (1) = 2.02, *p* = .21. Immersive VR provides its greatest relative advantage where learners have limited authentic communicative environments outside the classroom.

Table 5: Moderator Effects of Immersive VR On Language Learning Outcomes (Random-Effects Model).

Moderator	Category	k	g	95% CI	Qbetween	p
Educational level	Elementary	3	0.49	[0.12, 0.86]		
	Middle school	8	0.62	[0.31, 0.93]	2.41	.30
	College	10	0.71	[0.45, 0.97]		
Hardware type	Fully immersive	12	0.69	[0.41, 0.97]	1.87	.17
	Non-immersive	9	0.54	[0.26, 0.82]		
Language skill	Vocabulary	5	0.74	[0.33, 1.15]		

	Writing	5	0.70	[0.38, 1.02]	6.23	.06†
	Speaking	4	0.58	[0.24, 0.92]		
	Listening	2	0.42	[0.05, 0.79]		
	Mixed	3	0.61	[0.28, 0.94]		
L1/L2 context	L1	3	0.41	[0.07, 0.75]	2.02	.21
	L2	18	0.67	[0.39, 0.95]		

Note: k = number of studies per subgroup; g = Hedges' g; CI = confidence interval. Qbetween tests for statistically significant differences between subgroups. All analyses used a random-effects model. † p = .06, approaching conventional significance threshold.

4.5. Sensitivity Analysis

A one-study-removed influence analysis was conducted for both linguistic and affective outcomes. For both outcome types, pooled effect sizes remained statistically significant and within the medium range across all iterations, with no single study's removal materially altering the direction or magnitude of the overall estimates. Full influence analysis results are available from the corresponding author upon request.

4.6. Publication Bias

Publication bias was assessed through three complementary approaches: visual inspection of funnel plots, Egger's regression test, and fail-safe *N* analysis. Funnel plots for linguistic (Figure 3) and affective (Figure 4) outcomes displayed approximately symmetrical distributions, suggesting no systematic small-study effects. Egger's regression tests were nonsignificant for both outcome types ($p > .05$). Fail-safe *N* analyses confirmed that 327 additional null-result studies would be required to nullify the linguistic effect, and 89 for the affective effect; both values substantially exceed Rosenthal's (1979) criterion of $5k + 10 = 115$. Publication bias did not materially distort the present estimates.

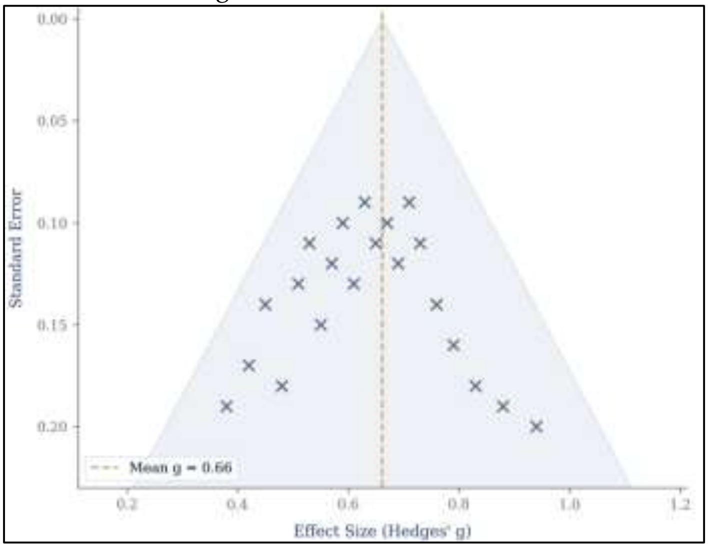

Figure 3: Funnel Plot for Linguistic Outcomes.

Figure 3 above shows the funnel plot for linguistic outcomes. Each marker represents one study. The dashed vertical line indicates the pooled mean (Hedges' g = 0.662). The approximately symmetrical distribution, combined with a nonsignificant Egger's regression test, indicates low risk of publication bias.

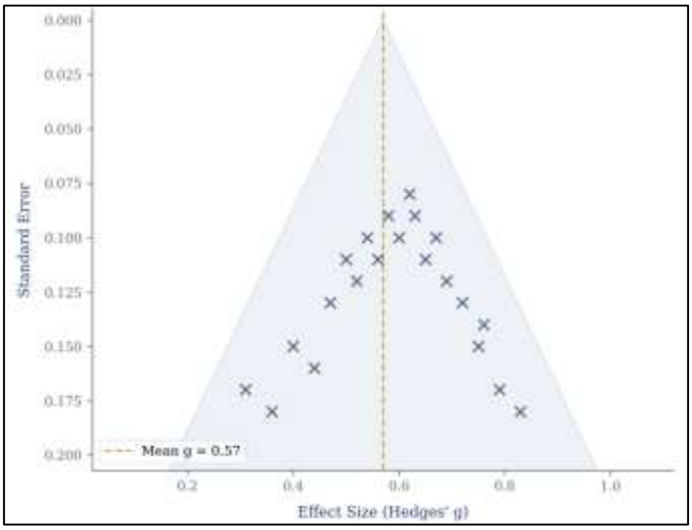

Figure 4: Funnel Plot for Affective Outcomes.

Figure 4 shows the funnel plot for affective outcomes. The dashed vertical line indicates the pooled mean (Hedges' g = 0.570). The approximately symmetrical distribution indicates minimal small-study effects and low risk of publication bias.

5. DISCUSSION

This meta-analysis provides rigorous quantitative evidence that immersive virtual reality exerts moderate, reliable positive effects on both linguistic achievement and affective engagement in language education. Synthesizing 21 studies with 1,144 participants, the pooled analyses yielded medium-sized effects for language proficiency gains (Hedges' *g* = 0.662) and affective outcomes (Hedges' *g* = 0.570), both statistically robust and stable across publication bias and sensitivity tests. These findings extend the predominantly qualitative literature by providing the first comprehensive quantitative synthesis of immersive VR outcomes across the full spectrum of contemporary hardware, from desktop and mobile platforms to HMDs and CAVE systems, and across a decade of research spanning 2010–2021.

The present findings both converge with and extend Wang, Lan, and Baralt (2016), who reported larger effect sizes for linguistic outcomes (*g* = 0.832) and slightly smaller affective effects (*g* = 0.531) across 13 studies. The more conservative linguistic estimate here (*g* = 0.662) most plausibly reflects a broader hardware corpus, a more rigorous evidence base, and the predominance of multimedia rather than traditional instruction as the comparator.

The heterogeneity detected (I^2 = 75.5% for linguistic, 53.3% for affective) is the central analytical challenge of this review. The lower heterogeneity for affective outcomes is theoretically meaningful: VR's motivational benefits appear relatively stable across implementations, while linguistic gains are more sensitive to instructional variables. This aligns with Dörnyei's (2005) motivational framework, which holds that technology-enhanced environments reliably boost engagement through novelty and interactivity, while linguistic gains depend more critically on pedagogical coherence.

The moderator analyses revealed coherent directional patterns. The consistent gradient across educational levels — college (*g* = 0.71) > middle school (*g* = 0.62) > elementary (*g* = 0.49) — is consistent with research on self-regulated learning (Zimmerman, 2002). The hardware advantage of fully immersive systems (*g* = 0.69 vs. 0.54) aligns with presence theory (Slater & Wilbur, 1997; Witmer & Singer, 1998; Lombard & Ditton, 1997), confirmed by Makransky and Lilleholt (2018) who demonstrated that higher presence was associated with deeper processing. However, cognitive load theory (Sweller, 1988) cautions that extraneous demands of VR interfaces may partially offset presence-related benefits, particularly for novice users.

The language skill moderator approached significance (Q_between = 6.23, *p* = .06). The stronger effects for vocabulary (*g* = 0.74) and writing (*g* = 0.70) relative to listening (*g* = 0.42) are consistent with embodied cognition theory (Barsalou, 2008): VR's capacity to embed words in vivid, spatially navigable contexts instantiates the elaborative processing that Nation (2001) identifies as essential for depth of vocabulary knowledge. The weaker listening effect reflects that comprehension gains depend primarily on input quality rather than immersive context. The larger L2 effects (*g* = 0.67 vs. 0.41) align with Long's (2015) Interaction Hypothesis: VR compensates for the scarcity of authentic communicative environments outside L2 classrooms.

Several limitations warrant acknowledgment. First, the geographic concentration of studies, predominantly Taiwan, Turkey, the United States, and Iran, substantially limits generalizability. The effect sizes reported here largely reflect East Asian and Middle Eastern pedagogical and institutional contexts; whether they hold in Western, African, or South Asian settings remains empirically untested. Second, the restriction to English-language publications introduces systematic language bias. Third, the predominance of quasi-experimental designs means that confounding due to teacher enthusiasm, novelty effects, and selection bias cannot be fully ruled out, and the Hawthorne effect is a salient concern in short-term VR studies. Fourth, the short duration of most interventions leaves long-term retention entirely unexamined. Fifth, pooling heterogeneous affective constructs, motivation, self-efficacy, engagement, and anxiety, under a single category obscures potentially important distinction.

Notwithstanding these limitations, a Hedges' g of 0.662 corresponds to a learner at the 50th percentile in a control condition rising to approximately the 75th percentile in an immersive VR condition, an educationally significant advantage. The evidence base now extends across a sufficient range of hardware types, educational levels, and instructional contexts to support moderately differentiated practical guidance, though practitioners outside the contexts represented in this corpus should treat these recommendations as provisional pending local evidence.

6. IMPLICATIONS AND RECOMMENDATIONS

6.1. Pedagogical Implications

The following recommendations apply most confidently to the contexts represented in this corpus, predominantly higher education and middle school settings, East Asian and Middle Eastern learner populations, English L2 instruction, and short-term interventions. Practitioners in other contexts should treat them as informed starting points pending local evidence.

The evidence supports immersive VR not as a replacement for existing pedagogy but as a powerful complement, one whose benefits are maximized when embedded within principled communicative or task-based frameworks. Educators should prioritize VR for vocabulary and writing, where effects are strongest. VR's capacity to embed lexical items in spatially navigable, contextually rich environments offers a depth-of-processing advantage that conventional approaches cannot replicate. Speaking activities are also well-suited to immersive VR, which simulates authentic communicative pressure in low-stakes environments, valuable for learners with high foreign language anxiety. Listening comprehension shows weaker effects, suggesting educators should not prioritize VR for listening when well-designed audio-based alternatives are available.

Fully immersive systems appear to yield stronger outcomes, consistent with presence theory, but their costs are substantially higher. For institutions with limited budgets, mobile or desktop VR targeting vocabulary and writing offers a meaningful alternative. Teacher preparation is a critical but underexamined factor: educators must be supported not only in technical operation but in designing tasks that exploit VR's distinctive affordances, including immersion, interactivity, and authentic contextual embedding.

6.2. Theoretical Implications

This study provides quantitative support for the convergence of presence theory, embodied cognition, and SLA research as a coherent framework for predicting VR's effects on language learning. The differential heterogeneity between linguistic and affective outcomes suggests a mediated pathway: affective engagement is a relatively stable proximal outcome of immersive VR; linguistic gain is a more distal outcome moderated by instructional quality. Cognitive load theory should be more explicitly integrated into future theoretical models, with studies measuring cognitive load directly rather than inferring it post hoc.

6.3. Policy And Institutional Implications

The moderate but consistent effect sizes justify cautious institutional investment, with three qualifications. First, investment should be evidence-driven rather than technology-driven. Second, equity and access considerations must be central: HMD costs risk exacerbating existing inequalities, and content development should extend beyond English-language and East Asian contexts. Third, institutional adoption should be preceded by rigorous pilot evaluation, as $I^2 = 75.5\%$ signals that not all VR implementations deliver comparable benefits.

6.4. Recommendations For Future Research

Five priorities emerge: (1) longitudinal investigations to assess durability of VR's linguistic benefits beyond immediate post-intervention periods; (2) reading comprehension studies, the most conspicuously absent skill domain; (3) cross-cultural and cross-linguistic implementations beyond East Asian and English-dominant contexts; (4) comparative research examining VR alongside AR and MR to clarify VR's unique contributions; and (5) systematic measurement of cognitive load, presence, and teacher implementation quality as potential moderators and mediators.

7. CONCLUSION

This systematic review and meta-analysis demonstrate that immersive virtual reality has a moderate, consistent, and practically meaningful positive effect on both linguistic achievement and affective engagement in language education. Synthesizing 21 quantitative studies with 1,144 learners, the pooled effect sizes, Hedges' g = 0.662 for linguistic outcomes and g = 0.570 for affective outcomes, are robust to publication bias and insensitive to the removal of any individual study. It bears noting, however, that this stability is established within a geographically concentrated, quasi-experimental corpus, predominantly East Asian contexts, short intervention durations, and English L2 settings, and that cross-cultural and longitudinal generalization remains an open empirical question.

Three contributions distinguish this review. First, it extends the quantitative evidence base beyond 3D virtual world platforms to encompass the full contemporary hardware landscape. Second, it grounds findings in a convergent theoretical framework, presence theory, SLA research, embodied cognition, and cognitive load theory. Third, it provides the most comprehensive moderator analysis to date, identifying fully immersive hardware, vocabulary and writing targets, L2 contexts, and higher educational levels as conditions associated with stronger outcomes.

The substantial heterogeneity for linguistic outcomes (I^2 = 75.5%) confirms that immersive VR is not a universal solution: its effectiveness is conditional on the alignment between hardware, task design, pedagogical integration, teacher preparation, and learner needs. The consistent affective benefit (I^2 = 53.3%) suggests VR's capacity to engage and motivate learners is a more stable feature of the technology. The path from experimental promise to mainstream practice runs through the questions this review has begun to answer, how much does VR help, for which skills, under what conditions, and through those it has identified as still unanswered: how durable are its benefits, how do they generalize across cultures and languages, and how can they be scaled equitably.

Acknowledgement: This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2602).

REFERENCES

Bailenson, J. (2018). Experience on demand: What virtual reality is, how it works, and what it can do. W. W. Norton & Company.

Barsalou, L. W. (2008). Grounded cognition. Annual Review of Psychology, 59, 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to meta-analysis. Wiley. https://doi.org/10.1002/9780470743386

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Dörnyei, Z. (2005). The psychology of the language learner: Individual differences in second language acquisition. Lawrence Erlbaum Associates.

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. BMJ, 315(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Godwin-Jones, R. (2016). Emerging technologies: Augmented reality and language learning. Language Learning & Technology, 20(3), 9–19.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. Academic Press.

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). Cochrane handbook for systematic reviews of interventions (2nd ed.). Wiley-Blackwell. https://doi.org/10.1002/9781119536604

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. BMJ, 327(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Huang, H. C., Hwang, W. Y., & Chang, H. C. (2016). Guest editors' introduction: Special issue on immersive virtual learning environments. Educational Technology & Society, 19(4), 1–2.

Huang, H., Zou, D., Cheng, G., & Xie, H. (2021). A systematic review of virtual reality in language learning: Learning outcomes and research trends. Interactive Learning Environments, 29(4), 549–563. https://doi.org/10.1080/10494820.2019.1636076

Jensen, L., & Konradsen, F. (2018). A review of the use of virtual reality head-mounted displays in education and training. Education and Information Technologies, 23(4), 1515–1529. https://doi.org/10.1007/s10639-017-9676-0

Lan, Y. J. (2020). Immersion, interaction, and experience-oriented learning: Bringing virtual reality into FL learning. Language Learning & Technology, 24(1), 1–15.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310

Lin, T. J., & Lan, Y. J. (2015). Language learning in virtual reality environments: A review of empirical research. British Journal of Educational Technology, 46(3), 468–487. https://doi.org/10.1111/bjet.12144

Lombard, M., & Ditton, T. (1997). At the heart of it all: The concept of presence. Journal of Computer-Mediated Communication, 3(2). https://doi.org/10.1111/j.1083-6101.1997.tb00072.x

Long, M. H. (2015). Second language acquisition and task-based language teaching. Wiley-Blackwell.

Makransky, G., & Lilleholt, L. (2018). A structural equation modeling investigation of the emotional value of immersive virtual reality in education. Educational Technology Research and Development, 66(5), 1141–1164. https://doi.org/10.1007/s11423-018-9581-2

Merchant, Z., Goetz, E. T., Cifuentes, L., Keeney-Kennicutt, W., & Davis, T. J. (2014). Effectiveness of virtual reality-based instruction on students' learning outcomes in K–12 and higher education: A meta-analysis. Computers & Education, 70, 29–40. https://doi.org/10.1016/j.compedu.2013.07.033

Nation, I. S. P. (2001). Learning vocabulary in another language. Cambridge University Press.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Qiu, Y., Lin, T. J., & Wang, C. H. (2021). Trends and research issues of virtual reality studies in language learning: A review of selected SSCI-indexed journals from 2008 to 2019. British Journal of Educational Technology, 52(1), 422–444. https://doi.org/10.1111/bjet.13017

Radianti, J., Majchrzak, T. A., Fromm, J., & 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

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. Psychological Bulletin, 86(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Schmidt, R. (1990). The role of consciousness in second language learning. Applied Linguistics, 11(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1535), 3549–3557. https://doi.org/10.1098/rstb.2009.0138

Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. Presence: Teleoperators & Virtual Environments, 6(6), 603–616. https://doi.org/10.1162/pres.1997.6.6.603

Solak, E., & Erdem, G. (2016). A content analysis of virtual reality studies in foreign language education. Participatory Educational Research, 3(1), 21–29. https://doi.org/10.17275/per.16.06.3.1

Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A. W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., … Higgins, J. P. T. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. BMJ, 355, i4919. https://doi.org/10.1136/bmj.i4919

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), Principle and practice in applied linguistics (pp. 125–144). Oxford University Press.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Tai, T. Y., & Chen, H. H. J. (2021). The impact of a virtual reality app on adolescent EFL learners' listening comprehension. Computer Assisted Language Learning, 34(5–6), 574–597.

https://doi.org/10.1080/09588221.2019.1619589

Wang, F., Lan, Y. J., & Baralt, M. (2016). The effectiveness of virtual worlds for language learning: A meta-analysis. Computer Assisted Language Learning, 29(5), 632–663. https://doi.org/10.1080/09588221.2015.1113708

Wehner, A. K., Gump, A. W., & Downey, S. (2011). The effects of Second Life on the motivation of undergraduate students learning a foreign language. Computer Assisted Language Learning, 24(3), 277–289. https://doi.org/10.1080/09588221.2010.551757

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. Presence: Teleoperators and Virtual Environments, 7(3), 225–240. https://doi.org/10.1162/105474698565686

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory into Practice, 41(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2