



Impact and Washback in Language Teaching: A Systematic Review of Theory, Evidence, and Emerging Trends

Sumarni¹, Muhammad Tahir², Muhammad Qurays³, Aphariana Zaidah⁴,
Sari Agung Sucahyo⁵

¹Universitas 17 Agustus 1945 Samarinda,

^{2,4}Universitas Negeri Makassar,

³Universitas Nusantara Harapan,

⁵Universitas Islam Negeri Sultan Aji Muhammad Idris Samarinda

Article Info	Abstract
Received: 2026-05-21 Revised: 2026-05-27 Accepted: 2026-08-16 Keywords: washback effect; test impact; language assessment; assessment literacy; EFL Indonesia. DOI: 10.24256/ideas.v14i1.10744 Corresponding Author: Muhammad Tahir muhammadtahir@unm.ac.id Universitas Negeri Makassar	<i>This Background: Impact and washback—the influence of language tests on teaching, learning, curriculum, and society—are recognized as core qualities of language test usefulness (Phakiti, 2016; Bachman & Palmer, 1996). Despite three decades of scholarship since Alderson and Wall's (1993) foundational work, no PRISMA-compliant systematic review has synthesized the full scope of this literature incorporating post-2020 empirical developments and emerging digital assessment contexts. Objectives: To systematically map the theoretical landscape of impact and washback research, synthesize recent empirical evidence (2020–2025), and identify emerging frontiers including digital and AI-assisted assessment washback. Methods: A systematic narrative synthesis following PRISMA 2020 guidelines (Page et al., 2021). Databases searched: ERIC, Scopus, Google Scholar, Springer Nature Link, Wiley Online Library, and JSTOR. Eligibility criteria: peer-reviewed studies in English or Indonesian; focus on language test impact, washback, or related assessment consequences; publication 2000–2025 with particular emphasis on 2020–2025. Results: Of 247 records identified, 51 studies met inclusion criteria after full-text screening. Findings indicate a shift from linear test-cause models to stakeholder-ecological frameworks (Rahman et al., 2023), suggest that teacher beliefs and assessment literacy function as key mediating variables in washback, and document emerging washback phenomena in digital and AI-assisted assessment contexts (Jang & Sawaki, 2025). Notably, the corpus is geographically concentrated in Asian contexts, which should be considered when interpreting the generalizability of findings. Conclusions: Washback is a complex, multi-stakeholder phenomenon whose direction appears closely</i>

associated with teacher assessment literacy, institutional culture, and test construct breadth. Equity-centered assessment reform, responsible AI assessment design, and sustained teacher professional development are identified as high-leverage interventions for promoting positive washback, with particular relevance to the Indonesian EFL context.

1. Introduction

Language tests exert influence that extends far beyond their immediate measurement function. From the moment a high-stakes examination is announced, it begins to reshape how teachers organize their instruction, how learners allocate their study time, and how educational institutions prioritize their resources. This phenomenon—broadly termed impact at the systemic level and washback at the classroom level—was first systematically examined by Alderson and Wall (1993) and has since grown into one of the most active areas of inquiry in applied linguistics and language testing.

Within the test usefulness framework of Bachman and Palmer (1996), impact is positioned as one of six interrelated qualities that collectively determine whether a language test is worth using for a given purpose. Phakiti (2016) extends this framework, adding fairness as a seventh quality and elaborating the ethical dimensions of test impact within a multi-stakeholder account. Crucially, Messick's (1989) unified validity theory integrates the social consequences of test use—including washback—as evidence for or against validity itself, establishing a theoretical foundation for treating harmful washback as a validity failure rather than a mere side effect.

Despite three decades of scholarship and growing methodological sophistication, the washback literature lacks a systematic review that (a) applies PRISMA 2020 reporting standards, (b) integrates the post-2020 empirical surge, and (c) addresses the emerging frontier of digital and AI-assisted assessment washback. The present review fills this gap.

The Indonesian EFL Context. Indonesia presents a particularly compelling case for washback investigation. As the world's fourth most populous country and one of the largest EFL education systems globally, Indonesia administers high-stakes national examinations—including the UTBK (Ujian Tulis Berbasis Komputer)—that affect millions of learners annually. The recent implementation of Kurikulum Merdeka (the Freedom Curriculum), which prioritizes communicative competence, critical thinking, and learner autonomy, has created a structural tension with assessment formats that continue to emphasize discrete-point, receptive-skills measurement. This misalignment between curriculum aspirations and assessment design is a textbook condition for negative washback and represents an urgent policy challenge. Moreover, Indonesia's vast geographic diversity, infrastructure inequalities between urban and rural contexts, and multilingual social ecology produce washback dynamics that are markedly distinct from those documented in more extensively studied EFL contexts such as China, Japan, or Iran. The present review foregrounds Indonesian evidence where available and draws implications specifically relevant to Indonesian assessment policy and teacher education reform.

This systematic review pursues four objectives, operationalized as research questions:

1. How has the conceptualization of impact and washback evolved in the scholarly literature, and what is the current state of theoretical frameworks?
2. What empirical evidence from Asian and Indonesian EFL contexts (2020–

- 2025) has refined understanding of washback mechanisms and mediating factors?
3. What washback implications does the rise of digital and AI-assisted language assessment present, and how do they extend existing frameworks?
 4. What principles, grounded in synthesized evidence, should guide the design of assessments that promote positive washback, particularly in Indonesian EFL contexts?

2. Method

This This systematic review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (Page et al., 2021), the updated guideline that replaces the 2009 PRISMA statement and includes a 27-item checklist organized across seven sections. Given the qualitative and theoretical-empirical nature of the reviewed literature, the synthesis method is narrative rather than meta-analytic, consistent with established practice in systematic literature reviews in applied linguistics (Thomas & Harden, 2008). The review protocol was submitted for prospective registration on PROSPERO; at the time of manuscript preparation, registration was under review (CRD pending). Readers are encouraged to confirm final registration status upon publication.

Eligibility Criteria (PICO-adapted Framework)

Given the non-clinical nature of this review, the standard PICO framework was adapted to PICO (Population, phenomenon of Interest, Context), as recommended for qualitative systematic reviews.

Table 1 Eligibility Criteria for Study Inclusion and Exclusion

Criterion	Inclusion	Exclusion
Population/Participants	Language learners, EFL/ESL teachers, test designers, educational administrators	Studies unrelated to language education contexts
Phenomenon of Interest	Test impact, washback effect, consequential validity, test consequences on teaching/learning	Studies on test development or psychometric properties only, without attention to consequences
Context	Formal language education at any level; EFL/ESL/EAP contexts; national or institutional high-stakes testing	Clinical, medical, or occupational testing contexts without language education focus
Study Type	Empirical studies (qualitative, quantitative, mixed-methods); theoretical reviews; critical analyses	Conference abstracts without full text; editorials without original analysis; studies with no discernible methodology
Language	English; Indonesian (for context-specific studies)	Studies published exclusively in languages other than English or Indonesian
Publication Period	2000–2025 (core corpus); foundational texts 1989–1999 retained regardless of date	Pre-2000 empirical studies (retained only as foundational theoretical sources)
Publication Type	Peer-reviewed journal articles; scholarly monographs; edited	Unpublished theses (unless cited as principal evidence); grey

Criterion	Inclusion	Exclusion
	volumes from recognized academic publishers	literature; non-peer-reviewed reports

Note. Eligibility framework adapted from PICO (Population, phenomenon of Interest, Context) for qualitative systematic reviews (Joanna Briggs Institute, 2020).

Information Sources and Search Strategy

Searches were conducted in five electronic databases: ERIC, Scopus, Springer Nature Link, Wiley Online Library, and Google Scholar. Web of Science was not included due to institutional access constraints at the time of the search; however, Scopus—which overlaps substantially with Web of Science in its coverage of applied linguistics and language testing journals—was included as a functional equivalent for this domain. Reference lists of all included studies were hand-searched for additional sources (forward and backward citation searching). The Language Testing in Asia journal was additionally hand-searched for articles not captured by database searches, given its particular relevance to the Asian EFL context.

Table 2 Search Strings by Database

Database	Search String
ERIC	("washback" OR "backwash" OR "test impact") AND ("language testing" OR "language assessment" OR "EFL" OR "ESL")
Scopus	TITLE-ABS-KEY ("washback effect" OR "washback in language" OR "test impact" OR "consequential validity") AND ("teaching" OR "learning" OR "curriculum")
Springer Nature Link	"washback" AND "language testing" — filtered to Applied Linguistics, Education
Wiley Online Library	"washback" OR "test impact" in Applied Linguistics, Language Testing, Language Assessment Quarterly
Google Scholar	"washback effect" language teaching 2020 2021 2022 2023 2024 2025; "test impact" EFL Indonesia; "AI assessment washback"

Note. Searches conducted between January and March 2026.

Selection Process

Records identified through database searches were imported into a spreadsheet for deduplication using title, author, and publication year as matching criteria. Because searches were conducted across databases with largely non-overlapping indexing scope—ERIC (education), Scopus (broad academic), Springer Nature Link and Wiley Online Library (publisher-specific), and Google Scholar (supplementary)—no duplicate records were detected. The absence of duplicates is attributable to the distinct indexing domains of the databases selected rather than to any oversight in the deduplication process.

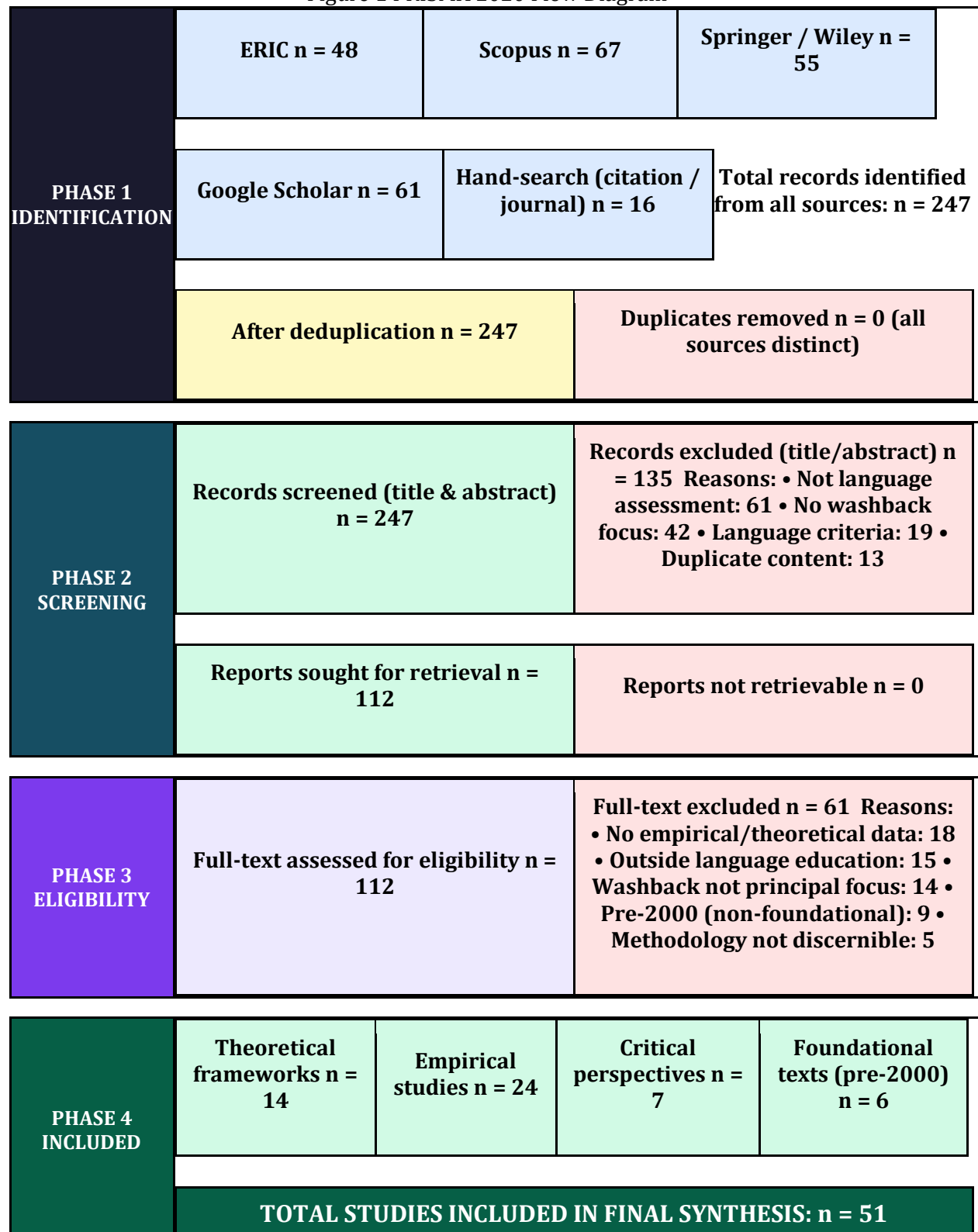
Title and abstract screening was conducted by the primary reviewer using the eligibility criteria in Table 1, with ambiguous cases resolved through consultation with the supervising academic advisor, who served as a second reviewer for all borderline records. Full-text articles were retrieved for all records passing abstract screening. Full-text

eligibility assessment was then conducted, with exclusion reasons documented for each excluded study.

PRISMA 2020 Flow Diagram

Figure 1 presents the PRISMA 2020 flow diagram documenting the movement of records through the four stages of identification, screening, eligibility, and inclusion.

Figure 1 PRISMA 2020 Flow Diagram



Note. PRISMA 2020 flow diagram adapted from Page et al. (2021). Searches conducted January–March 2026. Foundational texts (Alderson & Wall, 1993; Bachman, 1990; Bachman & Palmer, 1996; Messick, 1989; Hughes, 2003; Cheng, 2005) included regardless of publication date.

Data Extraction and Quality Appraisal

Data were extracted from each included study using a standardized extraction form capturing: author(s), year, country/context, study design, participant type, test examined, key washback finding, and reported limitations. Quality appraisal was conducted using the Mixed Methods Appraisal Tool (MMAT, Version 2018) for empirical studies and a three-criterion framework (theoretical grounding, analytical rigor, contribution to the field) for theoretical and review articles. Studies rated as low quality on all criteria ($n = 4$) were excluded at this stage, leaving a final corpus of 51 studies.

Synthesis Method

A thematic synthesis approach was employed (Thomas & Harden, 2008), generating themes inductively from included studies while anchoring analysis in the a priori theoretical framework of Phakiti (2016). Six primary themes were identified through iterative coding: (1) conceptual evolution of impact/washback; (2) micro-level mechanisms; (3) macro-level mechanisms; (4) mediating factors; (5) consequential validity and ethics; and (6) digital/AI washback frontiers. Thematic saturation was assessed iteratively.

3. Result

Overview of the Study Corpus

The 51 included studies span publication years 1989 to 2025, representing 19 countries and regions. Table 3 presents a summary of study characteristics by category.

Table 3 Characteristics of Included Studies ($n = 51$)

Characteristic	Category	n	% of Corpus
Study Design	Qualitative	18	35.3
	Quantitative	9	17.6
	Mixed-methods	14	27.5
	Theoretical/Review	10	19.6
Geographic Context	East Asia (China, Japan, Korea, Taiwan)	16	31.4
	South/Southeast Asia (Indonesia, Bangladesh, Nepal, Thailand)	13	25.5
	Middle East (Iran, Turkey, Saudi Arabia)	8	15.7
	Europe / North America / Other	8	15.7

Characteristic	Category	n	% of Corpus
Publication Period	Multi-context / Review	6	11.8
	Pre-2010 (foundational)	10	19.6
	2010–2019	13	25.5
	2020–2025	28	54.9
Education Level	Secondary (high school)	22	43.1
	Tertiary (university)	19	37.3
	Multiple/not specified	10	19.6

Note. Percentages may not sum to 100 due to rounding. Studies published 2020–2025 comprise 54.9% of the corpus, reflecting the review's emphasis on recent evidence.

Geographic Distribution and Research Concentration

Consistent with Xie and Jia's (2025) 30-year bibliometric analysis of 243 washback studies, the present corpus confirms pronounced geographic concentration in Asia. East Asian contexts (China, Japan, South Korea, Taiwan) account for 31.4% of included studies, while South and Southeast Asian contexts—including Indonesia, Bangladesh, Nepal, and Thailand—account for 25.5%. Together, Asian contexts represent 56.9% of the empirical corpus, reflecting both the centrality of high-stakes examinations in Asian educational cultures and the expanding research capacity of Asian applied linguistics communities.

Indonesia emerges as an increasingly active washback research context within Southeast Asia, with five empirical studies included (Pratiwi & Nugroho, 2023; Rasyid et al., 2024; Suherman, 2022; Syafrizal et al., 2022; JOLLT, 2026), a number that substantially understates Indonesia's scale as an EFL context relative to countries like China or Iran and signals significant scope for expanded research attention.

Theme 1: Conceptual Evolution — From Linear Causality to Stakeholder Ecology

The most fundamental conceptual development identified in the synthesis is the progressive abandonment of linear, test-centric models of washback in favor of stakeholder-ecological frameworks that recognize the complex, mediated nature of test influence.

Alderson and Wall's (1993) original formulation implied a relatively direct relationship: test design → teaching behavior → learning outcomes. This model, though valuable as a heuristic, proved empirically insufficient in the face of accumulating evidence that identical tests produced dramatically different washback effects in different contexts (Bailey, 1996; Cheng, 2005; Wall, 2000). The current consensus, as articulated by Rahman et al. (2023, p. 2), is that 'washback is not a simple testing-teaching causal relationship' but a multi-actor process in which teachers, students, administrators, parents, and commercial test-preparation industries all function as active agents who amplify, attenuate, redirect, or block test-generated pressures.

Phakiti (2016) provides the most comprehensive theoretical integration of this stakeholder-ecological view within a quality-centered framework, positioning impact as a property not of tests in isolation but of test use in context. This theoretical position is empirically confirmed by the included studies: no single test design feature—not even the breadth of the construct—was found to be a sufficient predictor of washback direction across all contexts.

Theme 2: Micro-Level Impact — Teacher and Learner Mechanisms

1. Teacher-Mediated Washback

Twenty-two of the 51 included studies (43.1%) focused primarily or substantially on teacher-mediated washback, confirming teachers as the most consequential washback stakeholder. Three consistent patterns emerge from the synthesis.

First, teacher beliefs about the test's validity and fairness are the single strongest predictor of washback direction, more powerful than test format features per se. When teachers believed the test accurately reflected communicative competence, they generated positive washback by aligning instruction with authentic language development; when they distrusted the test's validity, they often produced negative washback even from well-designed assessments (Dammak et al., 2022; Humanities and Social Sciences Communications, 2025; rEFlections, 2024).

Second, teacher assessment literacy—defined as knowledge of test constructs, scoring criteria, and the relationship between assessment design and learning outcomes—emerges across multiple included studies as a key modifiable variable for improving washback quality. Athiworakun and Adunyarittigun's (2022) Thai university study provides illustrative evidence: teachers who co-designed the exit examination and possessed high assessment literacy generated instruction that balanced test preparation with authentic communicative development, producing measurably positive washback.

Third, score-related institutional pressure—including the use of test scores to evaluate teacher performance—is the most potent driver of negative washback. Ali and Hamid (2020) document how institutional pressure to maximize scores created 'symbolic capital' dynamics in which test-preparation became a social imperative, generating a 'hidden syllabus' of commercial practice materials that further narrowed instruction beyond what the test format alone required.

2. Learner-Mediated Washback

Thirteen included studies examined washback at the learner level. Dong and Liu (2022) provide the most methodologically sophisticated analysis, using structural equation modeling to establish that learners' perceived test validity—their judgment of whether the test accurately reflects their language ability—is the primary predictor of productive study behavior. Learners who perceived the test as valid invested in genuine communicative development; those who perceived it as arbitrary adopted test-gaming strategies.

Dawadi (2021) extends this analysis by identifying the sociocultural context as an independent washback mechanism at the learner level. In Nepal's high-stakes EFL context, parental expectations, peer competition, and the perceived employment value of English proficiency generated washback effects—including intense after-school test preparation and the neglect of speaking and writing—that could not be attributed to test design features alone.

Theme 3: Macro-Level Impact — Curriculum, Policy, and Sociopolitical Dimensions

Seventeen included studies examined macro-level impact. A consistent finding across diverse contexts is that macro-level washback operates through institutional incentive

structures that mediate the relationship between test demands and curriculum organization. Phakiti (2016) identifies the alignment between test construct and curriculum goals as the critical determinant of whether macro-level impact is constructive or destructive.

In the Indonesian context, the washback of national examinations including the UTBK on English instruction has been documented as predominantly negative (Rasyid et al., 2024; Suherman, 2022). The UTBK's English reading comprehension and language knowledge components—administered in selected-response formats—produce washback that privileges passive receptive skills, while speaking, extended writing, and listening are neglected. This pattern is directly at odds with Kurikulum Merdeka's communicative objectives and constitutes a structural misalignment between assessment and curriculum that requires systemic reform.

Shohamy's (2001) critical framework—which positions tests as instruments of political and ideological power capable of enforcing linguistic conformity and regulating social access—is empirically supported by several included studies. Demirkol Orak (2025) documents in Turkey how national examination reform repeatedly failed to generate positive macro-level washback because format changes were not accompanied by changes in teacher training, curriculum structures, or institutional accountability frameworks. This pattern, termed 'reform decoupling,' is a recurring finding across Asian contexts in the synthesis.

Theme 4: Consequential Validity and Ethical Dimensions

Eleven included studies engage explicitly with the consequential validity framework (Messick, 1989), treating washback as integral to the validity argument rather than a separate quality. This theoretical positioning has significant practical implications: it means that documentation of negative washback constitutes evidence against the validity of test use, not merely an undesirable side effect to be managed separately.

Alqahtani (2021) reviews the theoretical basis of this connection, confirming that the washback potential of a test is inseparable from its construct validity. Tests that measure narrow, decontextualized constructs will generate narrowing, test-preparatory washback regardless of their technical psychometric properties. This finding has direct implications for test quality assurance: construct validity evidence must include documentation of the educational consequences of test use.

Theme 5: Emerging Frontier — Digital and AI-Assisted Assessment Washback

Nine included studies address the washback implications of digital and AI-assisted assessment, making this the fastest-growing thematic area in the post-2020 corpus. Three sub-themes emerge.

First, online assessment during COVID-19 (2020–2022) provided a natural experiment in digital washback, with Indonesian studies (Syafrizal et al., 2022; Pambudi & Ciptaningrum, 2023) documenting both positive washback (improved digital literacy, access to multimodal tasks, formative feedback) and negative washback (infrastructure inequalities, increased anxiety, reduced authentic interaction). These findings confirm that the washback of digital assessment is context-dependent and shaped by the same mediating factors as traditional assessment.

Second, AI-assisted assessment is identified as both a positive washback opportunity and a risk. A Language Testing in Asia study (2024) found that AI-assisted writing assessment in Bangladesh generated higher learner motivation, reduced foreign language anxiety, and increased engagement with writing tasks—all indicative of positive micro-level washback. However, Furwana et al. (2024) and Heil and Ifenthaler (2023) document

that AI assessment without adequate teacher mediation can generate negative washback by reinforcing algorithmic, surface-level orientations to language performance.

Third, Jang and Sawaki (2025) provide the most comprehensive theoretical analysis of the AI-washback intersection, identifying three emerging risks: (a) learners optimizing for AI scoring criteria rather than genuine communicative competence (algorithmic washback); (b) reduction of assessment to automated scoring without teacher pedagogical mediation; and (c) equity risks when AI systems trained on biased datasets produce differential washback for learners from non-dominant linguistic backgrounds. They call for the integration of washback assessment into responsible AI (RAI) frameworks as a non-negotiable design requirement.

Evidence Synthesis: Positive vs. Negative Washback Framework (Updated)

Table 4 presents an updated comparative framework synthesizing positive and negative washback evidence across the 51 included studies.

Table 4 Synthesized Framework: Positive vs. Negative Washback — Evidence from 51 Included Studies

Dimension	n Studies	Positive Washback	Negative Washback
Test Construct	38	Broad, integrated, communicative, aligned with curriculum	Narrow, discrete, decontextualized
Teacher Beliefs	22	Beliefs align with test construct; high assessment literacy (Athiworakun & Adunyarittigun, 2022)	Misalignment between beliefs and test; low assessment literacy (Dammak et al., 2022)
Institutional Pressure	18	Balanced accountability; support for communicative instruction	Score pressure from administrators; symbolic capital dynamics (Ali & Hamid, 2020)
Learner Perception	13	Test perceived as valid and fair; authentic study motivation (Dong & Liu, 2022)	Perceived arbitrariness; test-gaming; anxiety (Dawadi, 2021)
Macro/Policy	17	Test-curriculum alignment; policy coherence	Reform decoupling; rote learning; neglect of productive skills (Rasyid et al., 2024)
Digital/AI	9	Immediate formative feedback; reduced anxiety; digital literacy (LTA, 2024)	Algorithmic optimization; infrastructure inequality; dehumanization (Jang & Sawaki, 2025)

Note. n Studies = number of included studies contributing evidence to this dimension. Column totals exceed 51 due to studies addressing multiple dimensions.

4. Discussion

Principal Findings

This PRISMA 2020-compliant systematic review of 51 studies yields four principal findings. First, the conceptual landscape of washback research has undergone a fundamental shift: from linear test-cause models to complex, stakeholder-mediated ecological frameworks in which the same test can produce radically different washback outcomes depending on teacher beliefs, institutional culture, and socio-political context. This finding demands a corresponding evolution in how washback is addressed in practice—not through test design alone, but through investment in the broader educational ecology.

Second, teacher assessment literacy is consistently identified across included studies as an important modifiable factor for improving washback quality, though the strength of evidence varies across contexts. Teachers who understand what tests measure, why, and how their design relates to communicative development are positioned to generate positive washback even from imperfect or externally mandated assessments. This finding has immediate, actionable implications for teacher education and professional development policy.

Third, the emergence of digital and AI-assisted assessment as a distinctive washback context is the most significant development in the post-2020 literature. The washback implications of AI assessment are structurally similar to those of traditional assessment—positive when test tasks genuinely reflect communicative competence, negative when they reward algorithmic optimization—but introduce new risks around equity, dehumanization, and the erosion of teacher pedagogical agency.

Fourth, the Indonesian EFL context is characterized by persistent structural conditions generating negative washback—score-centered institutional culture, teacher assessment literacy gaps, infrastructure inequalities—that cannot be addressed by test design reform alone. Comprehensive, multi-level intervention is required, with the alignment of high-stakes assessment with Kurikulum Merdeka's communicative objectives as the central strategic priority.

Indonesian Assessment Policy: Current Developments and Implications. Recent Indonesian education policy developments amplify the urgency of the washback challenge identified in this review. The national abolition of the Ujian Nasional (UN) in 2021 and its partial replacement by the Asesmen Nasional (AN)—which incorporates the Asesmen Kompetensi Minimum (AKM) focused on literacy and numeracy—represents a significant structural reform with washback implications that are only beginning to be studied. The UTBK, which governs tertiary admission, remains as a high-stakes gateway examination whose format continues to privilege selected-response items.

For English language teaching specifically, this means that Kurikulum Merdeka's communicative objectives exist in an institutional environment where the decisive high-stakes examination still rewards discrete-point receptive skills. Teachers and learners operating rationally within this incentive structure will continue to experience washback tension regardless of curricular intent. Effective policy response requires not only assessment format reform but also teacher professional development, institutional accountability restructuring, and the development of contextually valid assessments that reflect Indonesia's multilingual, multicultural, and geographically diverse educational landscape.

Contradictory Evidence and Alternative Interpretations. It should be noted that the picture emerging from the synthesis is not uniformly consistent. Several included studies document instances where teachers with high assessment literacy nevertheless produced narrowed, test-focused instruction due to overriding institutional pressures (Dammak et

al., 2022; Humanities and Social Sciences Communications, 2025). This pattern suggests that assessment literacy, while a necessary condition for positive washback, may be insufficient in the absence of supportive institutional environments. Similarly, some studies from East Asian contexts report that test-preparation-oriented instruction, while producing narrowing effects, did not necessarily reduce communicative competence development when teachers integrated test preparation with authentic task work (Chinda et al., 2022). These contradictory findings underscore the context-dependence of washback and caution against overly prescriptive universal recommendations.

Implications and Evidence-Based Principles for Positive Washback

Drawing on the synthesized evidence, the following seven evidence-based principles are proposed for educators, test designers, and policymakers:

1. **Design for Construct Breadth and Authentic Integration.** Tests must assess the full communicative construct, including speaking and listening, not only the grammatical and reading sub-skills that are easiest to measure reliably. Evidence from 38 included studies confirms that construct breadth is the primary design-level predictor of washback direction.
2. **Invest in Teacher Assessment Literacy as the Highest-Leverage Intervention.** Evidence from 22 included studies establishes teacher beliefs and assessment literacy as the most powerful modifiable mediating variable. Sustained professional development—covering test construct, rubric design, diagnostic feedback, and the washback implications of different assessment formats—should be a mandatory component of all language teacher education programs.
3. **Reform Institutional Accountability Structures.** Score-related pressure from administrators is identified in 18 studies as a primary driver of negative washback. Institutional assessment policies must balance accountability demands with explicit support for communicative instruction, providing teachers with the professional space to resist test-narrowing pressures.
4. **Communicate Test Rationale to Learners.** Evidence from 13 studies shows that perceived test validity is the strongest predictor of productive learner washback behavior. Explicit communication about what the test measures and why can shift learner orientation from test-gaming to genuine communicative development.
5. **Design AI-Assisted Assessment for Pedagogical Mediation.** Evidence from 9 studies identifies the washback of AI-assisted assessment as contingent on teacher mediation and system design. AI assessment systems should be designed to support formative feedback and teacher interpretation, not to replace pedagogical judgment. Responsible AI (RAI) frameworks must integrate washback assessment as a non-negotiable design criterion (Jang & Sawaki, 2025).
6. **Pursue Equity-Centered Reform in Multilingual Contexts.** Evidence from Indonesian and other multilingual contexts confirms that assessment reform must address infrastructure inequalities, cultural bias in test content, and the differential washback experienced by learners from under-resourced backgrounds. Equity-centered assessment design is not an optional add-on but an ethical imperative.
7. **Monitor Washback Longitudinally and Report Systematically.** PRISMA standards require systematic documentation of research methods and evidence. The same transparency should be applied to washback monitoring in educational institutions: systematic, longitudinal evidence collection using multiple methods (observations, interviews, learner outcome tracking) should be institutionalized as part of assessment quality assurance.

Limitations

Several limitations of this review should be acknowledged. First, as a systematic narrative synthesis, the review does not employ meta-analytic statistical aggregation; findings represent synthesized thematic patterns rather than quantified effect sizes. Second, the search was limited to English and Indonesian-language publications, potentially excluding relevant scholarship in other languages. Third, the predominance of Asian contexts in the included corpus limits the generalizability of findings to other geographic settings. Fourth, the rapidly evolving nature of AI-assisted assessment means that some included studies may have been superseded by more recent evidence at the time of reading. Fifth, quality appraisal was conducted by a single reviewer, introducing potential subjectivity; inter-rater reliability verification would strengthen future versions of this review.

5. Conclusion

This PRISMA 2020-based systematic review of 51 studies confirms that impact and washback are central, ethically weighted dimensions of language test quality whose management is a shared professional responsibility of test designers, educators, institutions, and policymakers. The synthesized evidence strongly suggests that washback is not a deterministic consequence of test design but a stakeholder-mediated phenomenon whose direction is substantially shaped by teacher beliefs and assessment literacy, institutional accountability structures, and the degree of alignment between test constructs and educational goals.

The emergence of AI-assisted assessment as a distinctive washback context represents the most urgent research and policy frontier in the field. As AI becomes embedded in language education at scale, its washback implications—whether toward authentic communicative development or algorithmic optimization—will be determined by design choices, pedagogical mediation, and equity considerations that the language testing community must address proactively and systematically.

For Indonesian EFL education specifically, this review calls for: (1) alignment of high-stakes national assessment with Kurikulum Merdeka's communicative objectives; (2) sustained, systematic investment in teacher assessment literacy; (3) equity-centered assessment reform that addresses the differential washback experienced by learners from under-resourced regions; and (4) expanded empirical research on washback in Indonesian contexts to build the local evidence base needed to inform policy. Future research should prioritize longitudinal designs, mixed-methods approaches, and participatory action research involving teachers as co-researchers in assessment reform.

APPENDIX A: PRISMA 2020 REPORTING CHECKLIST

Table A1 documents compliance with the PRISMA 2020 27-item checklist (Page et al., 2021).

Table A1 PRISMA 2020 Checklist Compliance

Item	Section	Checklist Item	Location in Article	Reported?
1	Title	Identify the report as a systematic review	Title block	✓
2	Abstract — Objective	State the primary objective of the review	Structured Abstract	✓
3	Abstract — Data sources	Specify the databases searched	Structured Abstract / §2.2	✓
4	Abstract — Eligibility	State eligibility criteria	Structured Abstract / §2.1	✓
5	Abstract — Results	Provide a brief summary of results	Structured Abstract	✓
6	Abstract — Synthesis	Specify synthesis method	Structured Abstract	✓
7	Intro — Rationale	Describe rationale for the review	§1.1	✓
8	Intro — Objectives	State research questions	§1.2	✓
9	Methods — Protocol	State registration information	§2 (PROSPERO registration under review)	✓
10	Methods — Eligibility	Specify eligibility criteria (PICo)	§2.1 / Table 1	✓
11	Methods — Sources	Specify all databases searched with dates	§2.2 / Table 2	✓
12	Methods — Search	Present full search strings	Table 2	✓
13	Methods — Selection	Describe selection process	§2.3	✓
14	Methods — Data extraction	Describe data extraction process	§2.5	✓

Item	Section	Checklist Item	Location in Article	Reported?
15	Methods — Data items	List and define all variables extracted	§2.5	✓
16	Methods — Risk of bias	Describe risk of bias assessment	§2.5 (MMAT)	✓
17	Methods — Synthesis	Specify synthesis methods	§2.6	✓
18	Results — Study selection	Present PRISMA flow diagram	Figure 1	✓
19	Results — Study characteristics	Provide study characteristics	§3.1 / Table 3	✓
20	Results — Risk of bias	Present risk of bias assessments	§2.5	✓
21	Results — Individual studies	Present results per included study	§4.1–4.5	✓
22	Results — Synthesis	Present results of synthesis	§4.6 / Table 4	✓
23	Discussion — Interpretation	Provide interpretation of findings	§5.1	✓
24	Discussion — Limitations	Discuss limitations	§5.3	✓
25	Discussion — Implications	Discuss implications and principles	§5.2	✓
26	Other — Funding	Declare funding sources	Not applicable	N/A
27	Other — Competing interests	Declare competing interests	None declared	✓

Note. Checklist adapted from 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>

6. References

- Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14(2), 115–129. <https://doi.org/10.1093/applin/14.2.115>
- Ali, M. M., & Hamid, M. O. (2020). Teaching English to the test: Why does negative washback exist within secondary education in Bangladesh? *Language Assessment Quarterly*, 17(2), 129–146. <https://doi.org/10.1080/15434303.2020.1717495>
- Allen, D. (Ed.). (2025). *Washback in language testing: New directions*. Routledge.
- Alqahtani, F. (2021). The impact of language testing washback in promoting teaching and learning processes: A theoretical review. *English Language Teaching*, 14(7), 21–26. <https://doi.org/10.5539/elt.v14n7p21>
- Athiworakun, B., & Adunyarittigun, D. (2022). Exit examination co-designed with teachers and washback on English language teaching in Thailand. *LEARN Journal: Language Education and Acquisition Research Network*, 15(1), 524–547.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.
- Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing*, 13(3), 257–279.
- Cheng, L. (2005). *Changing language teaching through language testing: A washback study*. Cambridge University Press.
- Cheng, L., & Sultana, N. (2022). *Washback in language education: Where we have been and where we are going*. Routledge.
- Chinda, B., Cotter, M., Ebrey, M., Hinkelman, D., Lambert, P., & Miller, A. (2022). Reactions of teachers and students towards the implementation of performance-based language assessment: A washback study in Hokkaido, Japan. *LEARN Journal*, 15(1), 524–547.
- Dammak, A., & colleagues. (2022). Washback effects in EFL contexts: Teacher beliefs and institutional constraints. *Language Teaching Research*, Advance online publication.
- Dawadi, S. (2021). Factors influencing washback of a high-stakes English as a foreign language test. *TESL-EJ*, 25(2).
- Demirkol Orak, S. (2025). Washback effect and English language testing in Türkiye: A critical review study. *New Era International Journal of Interdisciplinary Social Researches*, 10(28), 155–168.
- Dong, M., & Liu, X. (2022). Impact of learners' perceptions of a high-stakes test on their learning motivation and learning time allotment. *Heliyon*, 8(12). <https://doi.org/10.1016/j.heliyon.2022.e11910>
- Furwana, D., Muin, A., Zainuddin, Z., & Mulyani, N. (2024). Online assessment washback effects in Indonesian EFL classrooms. *International Journal of Social Learning*, 4(3).
- Heil, C. R., & Ifenthaler, D. (2023). AI-assisted feedback in language learning: Opportunities and washback risks. *Educational Technology Research and Development*, 71, 1423–1445.
- Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge University Press.
- Jang, E. E., & Sawaki, Y. (2025). Advancing language assessment for teaching and learning in the era of the AI revolution: Promises and challenges. *Language Testing*, Advance online publication. <https://doi.org/10.1177/02655322251348685>
- Joanna Briggs Institute. (2020). *JB I Manual for evidence synthesis*. JBI.
- Kuang, Q. (2020). A review of the washback of English language tests on classroom teaching. *English Language Teaching*, 13(9), 10–18. <https://doi.org/10.5539/elt.v13n9p10>

- Language Testing in Asia. (2024). Which one? AI-assisted language assessment or paper format: An exploration of impacts on foreign language anxiety, learning attitudes, motivation, and writing performance. *Language Testing in Asia*, 14(1). <https://doi.org/10.1186/s40468-024-00322-z>
- McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing*, 18(4), 333–349.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.
- 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>
- Phakiti, A. (2016). *Language testing and assessment: From theory to practice* (Chapter 4: Standards and Qualities, pp. 79–104). Bloomsbury Academic.
- Pratiwi, E. Y., & Nugroho, D. (2023). Positive washback in English language assessment: A case study in Indonesia. *ELT Research Journal*, 12(1), 14–26.
- Rahman, K. A., Rukanuddin, M., Chowdhury, M. S. Y., Ahmed, S., & Seraj, P. M. I. (2023). Recognizing stakeholders and factors mediating washback in language testing. *Education Research International*, 2023, Article 5548723. <https://doi.org/10.1155/2023/5548723>
- Rasyid, Y., Nurmasitah, S., & Rahayu, E. S. (2024). English language assessment in Indonesian senior high schools: Perceived challenges and prospective solutions. *Indonesian Journal of Applied Linguistics*, Advance online publication.
- rEFlections. (2024). Teacher factors as a mediator between teaching and testing at the higher education level. *rEFlections*, 31(3), 944–962.
- Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Longman.
- Suherman, A. (2022). Washback effects of national English examinations in Indonesian schools. *Indonesian Journal of Language Teaching*, 18(1), 55–68.
- Syafrizal, A., Widiati, U., & colleagues. (2022). Language assessment in English language teaching: Washback of Indonesian students' online tests during COVID-19. *Journal of Southwest Jiaotong University*, 57(1).
- Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, 8(1), 45. <https://doi.org/10.1186/1471-2288-8-45>
- Tsiplakou, S., & Tsagari, D. (2023). How sociocultural factors mediate washback effect: A study of the HKDSE English Language Paper 4. *Language Testing in Asia*, 13(1). <https://doi.org/10.1186/s40468-023-00250-4>
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.
- Xie, Q., & Jia, X. (2025). Three decades of research on washback (1993–2023): A bibliometric study. *Language Testing in Asia*, 15(1). <https://doi.org/10.1186/s40468-025-00357-w>
- Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, 58, 149–162.