Beyond Compliance: Washback Dynamics of Embedded Assessment Practices in Flipped Learning Environments Among EFL Undergraduates

Tinur Rahmawati Harahap, Yulia Rizki Ramadhani, Rosni Harahap

Universitas Graha Nusantara, Kolonel Hamzah Lubis St. No.30, Padang Sidempuan, North Sumatera, 22711, Indonesia

*Corresponding author, email: tinurrahmawati@gmail.com

https://doi.org/10.17977/um065.v6.i10.2026.12

Article history
Submitted: 2 June 2026
Revised: 1 July 2026
Accepted: 8 July 2026
Published: 9 July 2026

Keywords
Assessment scheduling
EFL
Flipped learning
Mixed methods
Washback

Abstract
Flipped learning has transformed higher education through frequent embedded assessment; however, limited research has examined its washback effects on students' learning behavior. This study investigated positive and negative washback perceptions among undergraduate students in a flipped English as a Foreign Language (EFL) program and introduced the concept of a high-density washback environment (HDWE). A sequential explanatory mixed-methods design was employed. Quantitative data were collected from 60 undergraduate students at a public university in North Sumatra, Indonesia, using a validated 14-item Washback Perception Questionnaire and analyzed through descriptive statistics, paired-samples t-tests, and PLS-SEM confirmatory composite analysis. Qualitative data were obtained through semi-structured interviews and one focus group discussion involving six participants and analyzed using Reflexive Thematic Analysis. Positive washback was perceived at a high level and significantly exceeded moderate negative washback. Three qualitative themes emerged: self-regulated learning through pre-class assessment, cumulative assessment burden, and strategic compliance as ambivalent washback. The findings indicate that flipped learning assessment generally promotes positive learning behavior, although some responses reflect peer-mediated surface preparation rather than deep independent learning. The study proposes ambivalent washback as a distinct theoretical category and highlights assessment scheduling and instructional transparency as key factors influencing washback quality.

1. Introduction

Flipped learning is one of the most rapidly growing pedagogical innovations in higher education today. This model inverts the conventional sequence of instruction. Core content is delivered to students outside the classroom through videos, podcasts, or digital modules, while class time is used for discussion, group work, and the application of material to real-world problems (Bishop & Verleger, 2013; Hasan, Abidin, Hashim, Yaacob, & Fuzi, 2024; Zainuddin, Chu, & Perera, 2024). This shift is not merely technical; it reflects a deeper change in perspective, positioning students as active learners rather than passive recipients of information (Khayat, 2024; Ruengtam, 2026).

Lo and Hew (2017) noted a sharp increase in publications on the flipped classroom since 2012. The model has been applied across various fields, including medicine, engineering, mathematics, and language education. In Indonesia, flipped learning has also become increasingly widespread, particularly following the *Merdeka Belajar Kampus Merdeka* (*MBKM*) policy, which promotes student autonomy and experiential learning (Indonesian Ministry of Education Research, and Technology, 2020).

It is important to emphasize that flipped learning depends heavily on assessment. Because content delivery is shifted outside the classroom, instructors must ensure that students genuinely engage with the material before class. This is achieved through various forms of assessment, such as pre-class quizzes, in-class participation grading, reflection assignments, and group tasks. These assessments recur multiple times within a single week, unlike a final examination that occurs only once.

This condition is referred to in the present study as a high-density washback environment (HDWE), defined as a learning environment in which frequent and recurring assessments continuously shape how students' study. This differs from large-scale examinations, which typically occur only once and are distanced from day-to-day learning activities. The key question is whether such frequent assessment makes students more autonomous and motivated (positive washback), or whether it instead generates pressure and encourages

ISSN: 2797-3174

students to merely chase grades (negative washback). This question underlies the rationale for the present study.

Previous studies have shown that flipped learning has a positive impact on learning outcomes. Strelan, Osborn, and Palmer (2020), in a meta-analysis encompassing numerous studies, found that flipped learning improves students' academic achievement, particularly when accompanied by collaborative in-class activities. Li and Li (2022) similarly found that the flipped classroom enhanced student engagement in language learning.

However, other studies point to certain challenges. Abeysekera and Dawson (2015) explained that while flipped learning can increase learning motivation, it may also burden students who are not yet accustomed to regulating their own learning.

Washback theory, meanwhile, explains how assessment influences teaching and learning processes (Alderson & Wall, 1993). Cheng, Watanabe, and Curtis (2004) and Bailey (1996) distinguish between two types of washback. Positive washback occurs when assessment encourages deeper learning and intrinsic motivation. Negative washback occurs when assessment instead generates anxiety and leads students to study merely to pass. Green (2007) further argued that washback is not determined by the test format alone, but also by the context and by how instructors and students perceive the assessment.

The discussion above shows that research on flipped learning and research on washback have largely developed in isolation from one another. Studies on flipped learning, whether systematic reviews (O'Flaherty & Phillips, 2015), meta-analyses (Strelan et al., 2020), or narrative reviews (Lo & Hew, 2017), have generally focused on outcome-level questions, such as test scores or student satisfaction. In these studies, assessment is treated as a measure of learning outcomes rather than as a factor that shapes how students learn.

Conversely, washback research has predominantly examined large-scale assessments, such as university entrance examinations or international language proficiency tests (Cheng et al., 2004; Green, 2007). Assessment within flipped learning, however, is fundamentally different in nature. It is continuous, embedded within instruction, and relatively low-stakes, taking forms such as pre-class quizzes, reflective journals, and peer assessment. To date, no study has examined this type of assessment through the lens of washback theory.

A search of the Scopus and Web of Science databases using the keywords washback, backwash, and flipped classroom or flipped learning returned no studies that integrate these two areas of research. This indicates a clear gap in the literature, both in flipped learning scholarship and in washback scholarship. This gap is significant and warrants investigation. If assessment within flipped learning generates predominantly negative washback, the model's intended goal of fostering autonomous and intrinsically motivated learners may fail to be achieved. This issue is particularly relevant in Indonesia, where flipped learning is increasingly implemented under the MBKM policy, yet the washback effects of its embedded assessment practices remain unexamined.

Building on this gap, the present study aims to examine the washback effects of assessment practices within flipped learning. The study was conducted among undergraduate students in an English Language Education program at a public university in North Sumatra, Indonesia. It addresses three research questions: (1) To what extent do students experience positive washback, in the sense of freedom to learn, from assessment practices in flipped learning? (2) To what extent do students experience negative washback, in the sense of additional burden, from these same practices? (3) How do students' qualitative experiences help explain the quantitative washback patterns identified?

The findings of this study are expected to extend the application of washback theory from high-stakes testing contexts to formative assessment within innovative pedagogical models such as flipped learning. In addition, the findings are expected to offer practical guidance for instructors and institutions in designing assessment practices that more effectively support student learning.

2. Method

2.1. Research Design

This study employed a sequential explanatory mixed-methods design (Creswell & Creswell, 2023), in which quantitative data were collected and analyzed first, followed by qualitative data collection designed to explain and contextualize the initial quantitative findings. This design was selected because the research questions operate at two distinct but complementary levels of inquiry. Research Questions 1 and 2 concern the measurable extent of positive and negative washback perceptions across the student population and are therefore most appropriately addressed through quantitative survey data and structural measurement. Research Question 3 concerns the experiential logic underlying these perceptions, the "how" and "why" behind the quantitative patterns which requires the depth, contextual richness, and interpretive openness afforded by

ISSN: 2797-3174

qualitative inquiry. The sequential structure ensures that the qualitative strand is theoretically targeted: interview and focus group discussion (FGD) questions were developed after preliminary quantitative analysis, enabling the qualitative strand to probe specifically those constructs, patterns, and anomalies that emerged from the survey data, rather than operating as a parallel and independent investigation.

Data collection proceeded in three sequential phases. Phase 1 encompassed instrument development, expert validation, and pilot testing. Phase 2 involved quantitative data collection through the validated questionnaire administered to the full sample. Phase 3 involved qualitative data collection through semi-structured interviews and a focus group discussion with a purposively selected sub-sample. Mixed-methods integration was conducted in a fourth analytical stage, following the completion of both quantitative and qualitative analyses, using joint displays to systematically compare and synthesize findings across strands (Guetterman, Creswell, & Fetters, 2015).

2.2. Research Site and Participants

The study was conducted at a public university in North Sumatra, Indonesia, that had formally integrated flipped learning into its English Language Education undergraduate program curriculum since the 2022/2023 academic year. This site was selected for three reasons. First, flipped learning was institutionally embedded adopted as a formal curricular policy rather than implemented experimentally by individual instructors ensuring that participants had sustained, authentic experience with the model across multiple courses and semesters. Second, the program's assessment practices within the flipped learning context included all four types of assessment identified in the present study's conceptual framework: pre-class quizzes, in-class participation scoring, reflective journal submissions, and collaborative group task assessments, thereby constituting a genuine high-density washback environment (HDWE) as theorized in this study. Third, the Indonesian higher education context offers a particularly relevant empirical setting for this investigation, given the national MBKM policy framework's emphasis on outcome-based and student-centered learning, within which flipped learning is increasingly mandated (Indonesian Ministry of Education Research, and Technology, 2020).

The target population comprised all undergraduate students enrolled in the English Language Education program who had completed at least one full semester under the flipped learning curriculum. A purposive census of this population yielded a sample of N = 60 participants, comprising 42 female (70%) and 18 male (30%) students, with ages ranging from 19 to 23 years (M = 20.8, SD = 1.1). The gender distribution is consistent with the demographic profile of Indonesian English Language Education programs, in which female students represent the majority (Indonesian Ministry of Education Research, and Technology, 2020). The minimum sample size for the quantitative strand was determined a priori using GPower software (Faul, Erdfelder, Lang, & Buchner, 2007). Based on multiple linear regression, a medium effect size ($f^2 = 0.15$), a significance level of $\alpha = .05$, and a desired statistical power of .80, GPower yielded a minimum required sample of N = 55 participants (Hair, Henseler, Ringle, & Sarstedt, 2022). The achieved sample of N = 60 exceeded this threshold, providing adequate power for the planned analyses.

Six participants were selected from the quantitative sample using maximum variation sampling (Patton, 2015), a strategy designed to capture a diverse range of washback experiences rather than a representative cross-section. Selection criteria were operationalized across four dimensions of variation: (a) academic standing, including two high-performing, two mid-performing, and two lower-performing students based on GPA; (b) semester of study, including two students each from Semesters 3, 5, and 7, to capture variation in accumulated experience with the flipped learning model; (c) self-reported digital access, identified from questionnaire responses, to include participants with both high and limited access to the digital tools required for pre-class tasks; and (d) gender, including four female and two male participants, proportional to the quantitative sample. All six participants completed the quantitative survey and subsequently provided separate written informed consent for participation in the qualitative strand. Data adequacy was assessed through ongoing reflexive engagement during analysis: following completion of all six interviews and the FGD, the research team collectively determined that the data provided rich, diverse, and substantively sufficient accounts to address Research Question 3, with no new analytical dimensions emerging from later interviews relative to earlier ones (Braun & Clarke, 2021).

2.3. Instruments

Three instruments were used in this study: (1) a 14-item Likert-scale questionnaire for the quantitative strand, (2) a semi-structured individual interview guide, and (3) a focus group discussion guide for the qualitative strand.

2.3.1. Washback Perception Questionnaire

The primary quantitative instrument was a 14-item, five-point Likert-scale questionnaire developed specifically for this study to measure student perceptions of washback within flipped learning environments. Items were distributed across two latent constructs derived from washback theory (Alderson & Wall, 1993;

ISSN: 2797-3174

Cheng et al., 2004; Green, 2007). The questionnaire consisted of two latent constructs: Positive Washback and Negative Washback, each represented by seven items. Positive Washback (PW) measured the extent to which assessment practices in flipped learning promoted constructive learning behaviors and academic engagement. The indicators included independent learning, pre-class preparation, deeper content understanding, self-paced learning, classroom participation, confidence in English use, and critical thinking development. Negative Washback (NW) measured the extent to which assessment practices in flipped learning generated unfavorable or counterproductive learning effects. The indicators included excessive workload, time-management difficulties, anxiety and stress, technical constraints, unclear instructions, perceived assessment burden, and surface-level learning orientation. Higher scores on Positive Washback items indicated stronger perceptions of beneficial washback effects, whereas higher scores on Negative Washback items reflected stronger perceptions of adverse washback effects.

Item development followed a structured three-stage process. First, an initial item pool of 22 candidate items was generated through a synthesis of washback constructs derived from Alderson and Wall (1993), Cheng et al. (2004), and Green (2007), adapted linguistically and contextually for the flipped learning assessment environment. Second, items were evaluated through two rounds of expert consultation (described in Section 2.4), through which eight items were eliminated or merged based on feedback regarding redundancy, ambiguity, or insufficient construct specificity. Third, the remaining 14 items were subjected to pilot testing (described in Section 2.4), yielding the final instrument.

2.3.2. Interview Guide

The semi-structured individual interview guide comprised 12 open-ended questions organized into three thematic areas. The first area (4 questions) explored students' general experiences with assessment in flipped learning, including their perceptions of assessment purpose, frequency, and alignment with learning goals. The second area (5 questions) probed the behavioral and motivational consequences of assessment, asking participants to describe how assessment design in flipped learning influenced their preparation strategies, study habits, and engagement with course content. The third area (3 questions) examined constraining or negative assessment experiences, inviting participants to describe situations in which flipped learning assessment felt burdensome, anxiety-inducing, or counterproductive, and to identify factors that mediated these experiences.

2.3.3. Focus Group Discussion Guide

The FGD guide comprised eight structured prompts designed to elicit comparative group discussion regarding washback experiences across different courses, semesters, and assessment types. The FGD format was selected to complement individual interviews by generating interactional data, in which participants built upon, challenged, or confirmed each other's responses, thereby providing a social dimension that could not be captured through individual interviews alone. All prompts were designed to be open-ended and non-leading, with follow-up probing questions prepared to explore emerging themes in greater depth.

2.4. Validity and Reliability

Content validity of the Washback Perception Questionnaire was established through expert judgment using Aiken's V coefficient (Aiken, 1985). Six experts evaluated the initial 22 items in terms of relevance, clarity, and suitability. Items with Aiken's V values ≥ 0.80 were retained, resulting in 14 valid items (V = 0.82–0.94). The interview and FGD guides were also reviewed by two qualitative research specialists and refined through member checking procedures proposed Guba and Lincoln (2000). Pilot testing was conducted with 15 students outside the main sample. The initial 22-item instrument yielded Cronbach's Alpha of $\alpha = .71$, which increased to $\alpha = .83$ after item revision and removal. Four additional items with low corrected item-total correlations (< .30) were excluded, resulting in the final 14-item instrument (Positive Washback $\alpha = .79$; Negative Washback $\alpha = .81$).

The psychometric properties of the final instrument were evaluated using PLS-SEM in SmartPLS 4.0 through confirmatory composite analysis (CCA). Convergent validity was assessed using factor loadings, Average Variance Extracted (AVE), and Composite Reliability (CR), while discriminant validity was examined using the Fornell-Larcker criterion and Heterotrait-Monotrait ratio (HTMT). Internal consistency reliability was assessed using Cronbach's Alpha following the recommended thresholds proposed by Hair et al. (2022)

2.5. Data Collection

In Phase 1 (September 2024), the questionnaire underwent expert validation and pilot testing as described in Section 2.4. Revisions based on expert feedback and pilot results were completed before proceeding to Phase 2. In Phase 2 (October 2024), the validated 14-item questionnaire was administered to the N = 60 main sample participants during two scheduled class sessions, in the presence of the researchers and the course instructor, to ensure clarity of instructions and completeness of responses. Administration during class sessions was adopted to maximize response completeness; all 60 participants present in the sessions completed the questionnaire, yielding a 100% completion rate for the administered sessions. Questionnaires were distributed and collected in paper format to ensure equitable access regardless of individual device availability. In Phase 3

ISSN: 2797-3174

(November 2024), semi-structured face-to-face interviews were conducted individually with each of the six selected qualitative participants at the research site. Interviews were conducted in English or Indonesian based on participant preference; three participants elected to use English, two used Indonesian, and one used a combination of both languages. Each interview lasted between 40 and 55 minutes. Following the completion of all individual interviews, one FGD was conducted with all six participants simultaneously, lasting approximately 90 minutes. All sessions were audio-recorded with participants' written informed consent, transcribed verbatim, and returned to participants for member checking. Participants were invited to review, clarify, or amend their transcripts within five days of receiving them; three participants made minor clarifications to terminology, which were incorporated into the final transcripts.

2.6. Data Analysis

Quantitative data were analyzed using IBM SPSS Statistics 26.0 to generate descriptive statistics, including means, standard deviations, and frequency distributions. Paired-samples t-tests were conducted to compare Positive and Negative Washback scores at a significance level of $\alpha = .05$. The PLS-SEM measurement model was evaluated using SmartPLS 4.0 (Ringle, Wende, & Becker, 2022). Qualitative data from interviews and FGDs were analyzed using Reflexive Thematic Analysis proposed by Braun and Clarke (2021), following six analytical phases: familiarization, coding, theme development, theme review, theme definition, and report writing. Two researchers independently coded the data, producing strong inter-rater agreement (Cohen's Kappa $\kappa = .81$) based on the interpretation criteria of Landis and Koch (1977). Qualitative trustworthiness was ensured through credibility, transferability, dependability, and confirmability strategies following Lincoln and Guba, (1985). Mixed-methods integration was conducted through joint displays as proposed by Guetterman et al, (2015), enabling comparison and synthesis of quantitative and qualitative findings to generate integrated metainferences.

3. Results and Discussion

3.1. Results

3.1.1. Instrument Validity and Reliability

3.1.1.1. Content Validity: Aiken's V

Content validity of the 14-item Washback Perception Questionnaire was evaluated through expert judgment using Aiken's V coefficient (Aiken, 1985). Six content experts three specialists in English language assessment and washback research, two specialists in flipped learning curriculum design, and one specialist in quantitative educational measurement independently rated each item for relevance and clarity on a four-point scale. The minimum acceptable threshold was set at V = 0.80, following Aiken (1985) for a six-rater panel. Table 1 presents Aiken's V values for all 14 items.

Table 1. Content Validity Coefficients (Aiken's V) for the 14-Item Washback Perception Questionnaire (N raters = 6)

Item	Construct	Aiken's V	Significance
PW1	Positive Washback	0.92	p < .01
PW2	Positive Washback	0.94	p < .01
PW3	Positive Washback	0.89	p < .01
PW4	Positive Washback	0.83	p < .01
PW5	Positive Washback	0.86	p < .01
PW6	Positive Washback	0.82	p < .01
PW7	Positive Washback	0.88	p < .01
NW1	Negative Washback	0.91	p < .01
NW2	Negative Washback	0.87	p < .01
NW3	Negative Washback	0.90	p < .01
NW4	Negative Washback	0.82	p < .01
NW5	Negative Washback	0.83	p < .01
NW6	Negative Washback	0.85	p < .01
NW7	Negative Washback	0.84	p < .01
Range	—	0.82–0.94	—

Note. Minimum acceptable threshold: V ≥ 0.80 (Aiken, 1985). N raters = 6.

All 14 items exceeded the minimum threshold, with V values ranging from 0.82 to 0.94. Notably, items PW4 (learning autonomy) and PW6 (academic self-confidence) returned the lowest V values in the Positive Washback subscale (V = 0.83 and V = 0.82, respectively). Expert reviewers indicated in their written feedback that these two items required contextual specification, specifically, that the concept of "autonomy" in the Indonesian higher education context may be interpreted differently from its theoretical meaning in self-determination

ISSN: 2797-3174

theory (Ryan & Deci, 2000), as institutional expectations of student compliance are strong. Both items were retained following minor linguistic revision that clarified their intended meaning in context. Item NW4 (technical access barriers, V = 0.82) similarly required revision: two experts noted that the original phrasing conflated institutional internet access with personal device availability, and a revised item distinguished between these two dimensions. These minor revisions reflect the importance of contextual adaptation in instrument development for non-Western educational settings, a consideration noted but insufficiently addressed in many washback instrument validation studies (Cheng et al., 2004).

3.1.1.2. Measurement Model Assessment

The psychometric properties of the 14-item instrument were evaluated through confirmatory composite analysis (CCA) using PLS-SEM in SmartPLS 4.0 (Ringle et al., 2022). Table 2 presents the measurement model results, including standardized factor loadings, Average Variance Extracted (AVE), Composite Reliability (CR), Cronbach's Alpha (α), and the Heterotrait-Monotrait ratio (HTMT) for discriminant validity.

Table 2. Measurement Model Results: Factor Loadings, Reliability, Convergent Validity, and Discriminant Validity

Item	Factor Loading	AVE	CR	Cronbach's α	HTMT
Positive Washback		0.54	0.87	0.83	
PW1 – Learning independence	0.82				
PW2 – Pre-class motivation	0.85				—
PW3 – Content understanding	0.79				
PW4 – Learning autonomy	0.74				
PW5 – Discussion engagement	0.76				
PW6 – Academic self-confidence	0.71				
PW7 – Critical thinking development	0.81				
Negative Washback		0.51	0.84	0.79	
NW1 – Task overload	0.83				
NW2 – Time management difficulty	0.80				0.47
NW3 – Assessment-related anxiety	0.78				
NW4 – Technical access barriers	0.71				
NW5 – Instructional clarity problems	0.73				
NW6 – Perceived burden	0.77				
NW7 – Surface-level learning	0.75				

The measurement model demonstrates acceptable but not uniformly strong psychometric performance. Convergent validity was established for both constructs: all factor loadings met the minimum threshold of 0.70, AVE values exceeded 0.50, and CR values exceeded 0.70 (Hair et al., 2022). However, three observations warrant transparent discussion. First, the AVE for Negative Washback (AVE = 0.51) barely exceeds the minimum threshold, indicating that approximately 49% of the variance in NW indicators is attributable to measurement error rather than the latent construct. This is partly attributable to items NW4 (loading = 0.71) and NW5 (loading = 0.73), which exhibited weaker relationships with their construct than anticipated. For NW4, this pattern may reflect heterogeneous digital access conditions among participants, interview data subsequently revealed that several participants had inconsistent home internet access, making their experience with technical barriers less predictable than those with stable connectivity (a pattern discussed further in Section 3.3). Second, PW6 (academic self-confidence) demonstrated the lowest loading within the Positive Washback construct (0.71), suggesting that confidence-building may be a more contextually variable outcome of flipped learning assessment than independence or motivation. Third, Cronbach's Alpha for Negative Washback (α = 0.79) is acceptable but reflects modest internal consistency, likely due to the conceptual heterogeneity of negative washback experiences across the seven items (ranging from workload perceptions to surface-learning strategies).

Discriminant validity was confirmed through the HTMT ratio. The HTMT between Positive Washback and Negative Washback was 0.47, well below the 0.85 threshold (Hair et al., 2022), indicating that the two constructs are empirically distinct. The relatively low HTMT value (0.47) implies that the constructs share only approximately 22% of their variance, a finding that is theoretically meaningful. Washback theory posits that positive and negative washback can coexist within a single assessment context (Cheng et al., 2004; Green, 2007), and this moderate inter-construct independence is consistent with the theoretical expectation that students can simultaneously perceive assessment as motivating and burdensome, without these experiences fully cancelling each other out.

ISSN: 2797-3174

3.1.2. Quantitative Findings

3.1.2.1. Descriptive Statistics

Table 3 presents descriptive statistics for all 14 items and both construct-level scores. Category labels were assigned using a three-interval classification system applied to the five-point Likert scale: scores from 1.00–2.33 were classified as low, scores from 2.34–3.66 as moderate, and scores from 3.67–5.00 as high.

Table 3. Descriptive Statistics for Positive Washback and Negative Washback Items (N = 60)

Item	M	SD	Min	Max	Category
Positive Washback					
PW1 – Learning independence	4.08	0.71	2	5	High
PW2 – Pre-class motivation	3.95	0.74	2	5	High
PW3 – Content understanding	3.87	0.78	2	5	High
PW4 – Learning autonomy	3.72	0.82	1	5	High
PW5 – Discussion engagement	3.83	0.79	2	5	High
PW6 – Academic self-confidence	3.54	0.88	1	5	Moderate
PW7 – Critical thinking development	3.91	0.77	2	5	High
PW Total	3.84	0.64			High
Negative Washback					
NW1 – Task overload	3.42	0.91	1	5	Moderate
NW2 – Time management difficulty	3.38	0.87	1	5	Moderate
NW3 – Assessment-related anxiety	3.25	0.94	1	5	Moderate
NW4 – Technical access barriers	3.09	1.02	1	5	Moderate
NW5 – Instructional clarity problems	3.17	0.96	1	5	Moderate
NW6 – Perceived burden	3.21	0.93	1	5	Moderate
NW7 – Surface-level learning	3.31	0.89	1	5	Moderate
NW Total	3.26	0.72			Moderate

Overall, students rated Positive Washback at a high level (M = 3.84, SD = 0.64), while Negative Washback was rated at a moderate level (M = 3.26, SD = 0.72). A paired-samples t-test confirmed that the difference between construct-level means was statistically significant: $t(59) = 4.23$, $p < .001$, Cohen's $d = 0.55$ (medium effect). This indicates that students perceived significantly stronger positive washback than negative washback from flipped learning assessments, though the moderate-range negative washback scores signal that burden perceptions were not negligible and should not be dismissed as residual noise.

3.1.2.2. Positive Washback Perceptions

Among the seven Positive Washback items, PW1 (learning independence, M = 4.08, SD = 0.71) and PW2 (pre-class motivation, M = 3.95, SD = 0.74) received the highest mean ratings, indicating that students most consistently associated flipped learning assessments with autonomous study behavior and enhanced pre-class preparation motivation. PW7 (critical thinking development, M = 3.91, SD = 0.77) also scored highly, suggesting that students perceived cumulative assessments as promoting higher-order cognitive engagement, a finding consistent with O'Flaherty & Phillips (2015), who identified development of higher-order thinking as one of the most frequently reported benefits of flipped learning across disciplines.

Notably, PW6 (academic self-confidence, M = 3.54, SD = 0.88) was the only Positive Washback item that fell in the moderate rather than high category and also returned the highest standard deviation within the subscale (SD = 0.88). This variability indicates meaningful individual differences in how students experience assessment-related confidence in flipped environments. Inspection of the response distribution revealed that approximately 28% of respondents ($n \approx 17$) rated PW6 at 3 or below, while 72% rated it at 4 or 5. This bimodal tendency suggests that confidence-building effects may be contingent on factors not captured by the aggregate score such as prior academic performance, English proficiency level, or cumulative experience with the model and the qualitative data subsequently illuminate this heterogeneity (see Section 3.3).

PW4 (learning autonomy, M = 3.72, SD = 0.82) was the second-lowest Positive Washback item. The relatively higher standard deviation (0.82, compared to 0.71–0.77 for PW1, PW2, PW7) reflects variability in how students interpret and experience autonomy within the structured assessment demands of flipped learning, an ambiguity consistent with the theoretical tension identified by Abeysekera dan Dawson (2015), who noted that the self-directed demands of flipped learning can simultaneously foster and undermine perceived autonomy depending on the sufficiency of scaffolding.

3.1.2.3. Negative Washback Perceptions

All seven Negative Washback items fell in the moderate category (M = 3.09–3.42), indicating that negative washback was present and meaningful across the sample, though not dominant. NW1 (task overload, M = 3.42,

ISSN: 2797-3174

SD = 0.91) and NW2 (time management difficulty, M = 3.38, SD = 0.87) received the highest mean ratings within the subscale, positioning workload volume and temporal pressure as the primary sources of negative washback, a pattern replicated across multiple flipped learning studies (Abeysekera & Dawson, 2015; Strelan et al., 2020).

A methodologically notable finding concerns NW4 (technical access barriers, M = 3.09, SD = 1.02). This item returned both the lowest mean and the highest standard deviation among all 14 items. The higher dispersion (SD = 1.02, the only item exceeding 1.00) reflects genuine variability in students' digital access situations across the sample: while students in urban areas or with institutional internet subscriptions reported minimal technical barriers, students from semi-rural residential contexts or who relied on mobile data reported substantially greater access difficulties. This heterogeneity was not fully controlled for in the sampling procedure, a limitation acknowledged in Section 4 and it reduces the item's construct representativeness at the aggregate level, which is consistent with its lower factor loading in the measurement model (NW4 = 0.71, the lowest in the instrument). Future washback research in Indonesian HE settings should consider stratifying by digital access profile, as this variable appears to function as a meaningful moderator of negative washback experience in HDWE contexts.

NW7 (surface-level learning strategies, M = 3.31, SD = 0.89) was the second-highest NW item, suggesting that a meaningful proportion of students acknowledged adopting surface compliance behaviors completing pre-class tasks without deep engagement in response to the cumulative assessment demands of flipped learning. This finding is particularly consequential because surface-level learning directly contradicts the pedagogical rationale of the flipped model (Bishop & Verleger, 2013), which depends on students engaging substantively with pre-class content for in-class activities to function effectively. The qualitative strand explored this tension in greater depth, yielding the study's most theoretically significant contribution (see Section 3.3.3).

3.1.3. Qualitative Findings: Reflexive Thematic Analysis

Reflexive Thematic Analysis (Braun & Clarke, 2021) of interview transcripts and FGD data generated three themes that provided interpretive depth for the quantitative patterns described above. Table 4 presents the qualitative participant profiles, included here to contextualize the voices cited in subsequent analysis. A note on data completeness: Participant 5 (P5) was required to attend to a personal matter and was absent from approximately 25 minutes of the FGD session (between the second and third discussion prompts). Data from that portion of the session were therefore derived from five rather than six participants; this is noted where relevant in Table 4.

Table 4. Qualitative Participant Profiles

ID	Gender	Semester	GPA	Digital Access	PW Score	NW Score
P1	Female	7	3.72	High	4.29	2.86
P2	Female	5	3.14	Moderate	3.86	3.43
P3	Male	3	3.05	High	3.71	3.57
P4	Female	7	3.81	High	4.43	2.71
P5	Female	5	2.83	Limited	3.29	3.86
P6	Male	3	3.52	Moderate	3.57	3.71

3.1.3.1. Theme 1: Pre-class Assessment as a Self-Regulation Catalyst

The most consistently expressed theme across participants was the role of pre-class assessment events principally weekly quizzes and reading reflection tasks in structuring students' independent study behavior. Five of the six participants described a clear behavioral shift in how they prepared for class following enrollment in flipped learning courses, attributing this shift explicitly to the evaluative demands of pre-class tasks. P1, a high-performing Semester 7 student, reflected:

> "Before flipped class, I used to wait until before the exam. Now I actually read before every session because I know there will be a quiz. It becomes a habit after a while I do not even feel like I am doing it for the grade anymore".

P4 offered a complementary account that foregrounded the self-regulatory dimension:

> "The pre-class quiz forces me to plan my week. I know Wednesday is the flipped class, so I arrange my time on Monday and Tuesday for the reading. It is uncomfortable at first but now it is my routine".

These accounts are consistent with the autonomy-supportive mechanism theorized by Abeysekera and Dawson (2015) within self-determination theory Ryan and Deci (2000), when students internalize assessment demands as structuring rather than controlling, the extrinsic motivation of a quiz grade can transition toward more self-determined study behaviors. The quantitative scores for P1 (PW = 4.29) and P4 (PW = 4.43) are the

ISSN: 2797-3174

highest in the qualitative sub-sample, lending quantitative corroboration to the depth of positive washback they described.

P6, however, offered a partially divergent account that qualifies the generalizability of this theme. While acknowledging that pre-class tasks had changed his preparation habits, he expressed ambivalence about whether the change was substantive:

> "I prepare more, yes. But prepare how? I prepare to answer the quiz correctly, not always to understand deeply. I watch the video and note the things that I think will be asked".

P6's account introduces a distinction between behavioral compliance and cognitive engagement that was developed more fully in Theme 3, and which the aggregate quantitative data where P6 scored PW = 3.57, below the sub-sample mean of 3.86 only partially captures.

3.1.3.2. Theme 2: Cumulative Assessment Burden and Temporal Compression

All six participants reported experiencing periods of significant assessment-related stress, particularly during the mid-semester period (Weeks 7–9), when cumulative pre-class tasks intersected with concurrent submissions from other courses. Unlike Theme 1, which showed variation in intensity across participants, Theme 2 was notable for its universality: regardless of GPA, semester, or digital access profile, all participants described a subjectively overloaded assessment experience at some point in the semester.

P5, whose NW score was the highest in the qualitative sub-sample (NW = 3.86) and who reported limited digital access, articulated the cumulative dimension most directly:

> "Every week there is something quiz, journal, reflection. And then the other courses also have assignments. In Week 8, I had three quizzes, one presentation, and one reflection submission all in the same week. I could not sleep properly for four days. I felt like I was running without stopping."

P2 described a qualitatively different but structurally similar experience, noting that the temporal compression of assessment demands intersected with preparation quality:

> "When the deadline is the same day as the class, I do not have enough time to think about the material. I complete the task but I do not feel I actually learned. It is like going through the motion."

This account closely mirrors the surface-level learning strategy construct captured in NW7 (M = 3.31), and its emergence from a mid-GPA participant (P2, GPA = 3.14) who scored moderately on both subscales (PW = 3.86; NW = 3.43) suggests that the surface-deep learning distinction is not exclusively a function of academic ability. Assessment scheduling and temporal organization, variables within the control of course designers rather than students, may be more consequential moderators of washback quality than individual differences alone.

Participants in the FGD (with P5 present for the first portion) collectively identified mid-semester as the most vulnerable period, and several spontaneously suggested that staggered deadlines distributing reflective tasks across the week rather than concentrating them immediately before class sessions would substantially reduce the workload experience without diminishing its learning function. This observation carries practical implications for assessment design that are developed in the Implications section.

P3, the lowest-GPA participant in the qualitative sub-sample (GPA = 3.05) and a Semester 3 student new to the flipped format, added a dimension specific to novice learners:

> "I did not understand what was expected from me in the first month. The quiz came and I did not know what kind of answer they want more analysis or just summarize the reading? So I was anxious every time, not because I did not study, but because I did not know if I studied the right things."

This account points to instructional clarity as a critical but underappreciated dimension of washback in HDWE contexts. P3's experience aligns with the moderate mean of NW5 (instructional clarity problems, M = 3.17) but illuminates its mechanism: the anxiety induced by assessment in flipped environments may be driven less by the assessment's existence than by the transparency of its criteria — a factor related to assessment design quality rather than to flipped learning per se.

ISSN: 2797-3174

3.1.3.3. Theme 3: Strategic Compliance: Ambivalent Washback and the Study Group Effect

The most theoretically significant and unexpected finding of the qualitative strand was the emergence of a pattern that the research team termed strategic compliance a form of assessment engagement that superficially resembles positive washback but which operates through surface-level mechanisms inconsistent with the deep learning goals of the flipped model. Four of the six participants (P2, P3, P5, P6) disclosed, to varying degrees and with evident discomfort, that they regularly used peer-shared summaries, WhatsApp study group notes, or prior-year quiz compilations to complete pre-class tasks rather than engaging independently with the full assigned reading or video content.

P6 described this practice with notable analytical self-awareness:

> "In our cohort, there is always someone who makes a summary of the video. By the time I get to actually watch the video, I already have the notes from my friend. So I read the summary, sometimes I watch part of the video to make sure, and then I do the quiz. My score is usually good. But did I really learn the same as if I watched the whole thing? I am not sure."

P3 offered a normative account that contextualizes the practice within peer culture:

> "Everyone does it. It is not cheating we share notes, not answers. But the effect is maybe the same: we prepare, but we prepare together, not individually. The teacher cannot see this from the quiz score."

This finding reveals an important tension in the interpretation of aggregate quantitative data. The high mean of PW1 (learning independence, M = 4.08) and PW2 (pre-class motivation, M = 3.95) reflects that students do study before class and do feel motivated to prepare. However, the qualitative data reveal that the nature of this preparation may depart significantly from the autonomous, deep engagement that these items were designed to capture. Students report a sense of motivation and readiness that is functionally contingent on social scaffolding rather than independent engagement a phenomenon not captured by Likert-scale items that ask whether students "study more independently" without probing the mechanism of that study.

Theoretically, this finding extends washback theory in a direction that has received limited attention. Alderson and Wall (1993) and Green (2007) conceptualized washback primarily in terms of individual responses to assessment demands. The present study's finding suggests that in digitally connected peer learning environments, particularly in Indonesian higher education, where collectivist norm's structure academic cooperation (cf. the *gotong royong* cultural principle), washback may operate as a social rather than purely individual phenomenon. Assessment in flipped learning may generate collective preparation strategies that produce assessment compliance without commensurate individual learning, a dynamic that is invisible at the aggregate level.

3.1.4. Mixed-Methods Integration

Table 5 presents the joint display integrating quantitative and qualitative findings across the three research questions. Metainferences, integrated interpretations combining both strands are presented in the rightmost column.

Table 5. Joint Display: Integration of Quantitative and Qualitative Findings Across Research Questions

RQ	Quantitative Finding	Qualitative Theme	Integration Pattern	Metainference
RQ1: Positive washback	PW overall: M = 3.84 (High); PW1 highest (M = 4.08); PW6 lowest (M = 3.54, SD = 0.88)	Theme 1: Pre-class assessment as self-regulation catalyst; behavioral routinization; autonomy internalization (P1, P4); partial ambivalence (P6)	Convergence with qualification	Positive washback is broadly present and statistically significant; however, the qualitative data reveal that high PW scores at the aggregate level partially mask individual variation in the depth of autonomous engagement, particularly among lower-confidence students (PW6 heterogeneity)
RQ2: Negative washback	NW overall: M = 3.26 (Moderate); NW1 highest (M = 3.42);	Theme 2: Cumulative burden and temporal compression; universal across	Convergence	Negative washback is moderate but genuinely present across the full range of student profiles; the

ISSN: 2797-3174

RQ	Quantitative Finding	Qualitative Theme	Integration Pattern	Metainference
	NW4 lowest (M = 3.09, SD = 1.02)	participants regardless of GPA; mid-semester vulnerability identified; digital access as heterogeneous moderator		qualitative data explain NW4's high dispersion through heterogeneous digital access, and identify temporal design of assessment deadlines as a key modifiable source of burden
RQ3: Contextual factors	NW4 high SD (1.02) = digital access variability; PW6 high SD (0.88) = confidence heterogeneity; t-test: d = 0.55 (medium), positive > negative but gap not large	Theme 3: Strategic compliance; social preparation via peer networks; assessment completion decoupled from individual deep engagement; particularly among Semesters 3–5	Complementarity	The quantitative strand establishes that positive washback predominates at the population level; the qualitative strand reveals a critical mechanism social/strategic compliance that inflates PW scores while undermining the individual deep learning that the flipped model requires; this represents an ambivalent washback dynamic invisible to survey measurement alone

3.2. Discussion

3.2.1. Positive Washback: Autonomy, Motivation, and the Limits of Self-Report

The finding that students perceived stronger positive washback (M = 3.84) than negative washback (M = 3.26) lends support to the theoretical assumptions of Abeysekera and Dawson (2015), who argued that structured assessment embedded in flipped learning can satisfy the basic psychological needs identified in self-determination theory (Ryan & Deci, 2000), particularly autonomy and competence. High scores on learning independence and pre-class motivation suggest that students generally internalized assessment demands as part of their regular learning routines rather than experiencing them as externally imposed obligations.

This pattern resonates with, but also extends, recent washback research in EFL contexts. Washback is generally understood as the influence that tests, especially high-stakes ones, exert on test takers, teachers, students, parents, curriculum, and society, and this influence can be either harmful or beneficial. Montasseri and Ahmadi (2022), in a mixed-methods study of an interactional competence checklist used with 27 advanced EFL learners, found positive washback in areas including learners' familiarity with the assessment tool, their perception of its quality, their motivation, and their study habits. The present study's findings are broadly consistent with this pattern in that a low-stakes, formative assessment instrument generated positive perceptual shifts. However, an important point of divergence should be noted: Montasseri and Ahmadi's (2022) checklist was a single, discrete intervention compared against a control group, allowing the washback effect to be attributed to that specific tool. In the present HDWE context, by contrast, positive washback emerged from the *cumulative density* of multiple recurring assessment types operating simultaneously (pre-class quizzes, participation grading, reflection tasks, group submissions). This suggests that positive washback in flipped learning may not derive from any single well-designed instrument, as in Montasseri and Ahmadi's case, but from the structural rhythm of frequent, low-stakes accountability itself, a mechanism that has not been isolated in prior single-instrument washback studies.

A similar pattern of cautious optimism appears in Vietnamese EFL washback research. Ahmadi and Montasseri's findings, alongside Alqahtani's (2021) analysis of the Learning Outcome-Based English Language Assessment in Saudi Arabia, suggest that well-structured assessments can encourage more effective teaching methods and greater student engagement, indicating that assessment design plays a central role in shaping learning outcomes. The present study corroborates this design-sensitivity claim but adds a contextual layer largely absent from these single-country studies: in an Indonesian flipped learning setting, "design" cannot be separated from "frequency and embeddedness," since the same instrument administered once versus weekly may produce markedly different washback profiles. This is a dimension that single-snapshot washback studies, which typically examine one assessment tool at one point in time, are structurally unable to capture (Khan, Hassan, & Cheng, 2025).

ISSN: 2797-3174

It is equally important, however, to interrogate the limits of this positive finding rather than simply affirm it. Recent washback scholarship across Asian high-stakes testing contexts has increasingly recast washback not as a fixed property of a test but as a negotiated and evolving dialogue between assessment design and pedagogical practice, shaped by the interplay of policy, institutional culture, teacher professionalism, and student literacy. Read through this lens, the high mean positive washback score in the present study should not be interpreted as evidence that assessment design alone produced autonomous learning. Rather, it likely reflects a particular configuration of institutional culture (the collectivist, MBKM-influenced Indonesian context) interacting with assessment frequency (Cheng, 2026). These reframing matters because it tempers an overly instrument-centric reading of the quantitative result and motivates the more critical analysis of ambivalent washback developed in section 4.3.

3.2.2. Negative Washback: Workload, Time Pressure, and Assessment Scheduling

Despite the predominance of positive washback, moderate negative washback scores indicate that students still experienced workload pressure and time-management difficulties. This finding aligns with longstanding concerns in the flipped learning literature regarding workload accumulation, as documented by Lo and Hew (2017). It is also consistent with more recent flipped learning research: for instance, a 2025 study of flipped learning in a South Korean EFL grammar course found that the model produced mixed perceptual outcomes, with flipped learning failing to produce greater gains in grammar test scores compared to a traditional class group, despite its intended pedagogical advantages. While that study measured achievement rather than washback directly, it is suggestive of a broader pattern in recent flipped learning scholarship: positive design intentions do not automatically translate into uniformly positive student experience, and workload or pacing issues can offset cognitive gains (Harrison, 2025).

The qualitative findings in the present study sharpen this picture by locating the source of burden more precisely. Students' accounts suggest that perceived burden was associated primarily with assessment scheduling (the clustering and overlapping timing of multiple low-stakes tasks) rather than with the design or difficulty of the tasks themselves. This is a noteworthy refinement of the existing literature. Most prior negative-washback accounts, including Abeysekera and Dawson (2015) cognitive load argument, attribute burden to the inherent demands of self-regulation under flipped conditions in general terms, without disaggregating what specifically drives the cognitive load: task difficulty, task novelty, or scheduling density. The present study's finding that scheduling, not task design, is the proximate source of negative washback suggests a more actionable and more easily remediated source of strain than the literature has previously identified. This distinction also differentiates the present study from recent rubric-based washback research in Vietnam, where negative washback was traced to a narrowing effect in what was assessed. In that context, lecturers' selective focus on easily measurable components such as grammar accuracy, at the expense of content and critical thinking, was identified as a form of negative washback rooted in instrument design and teacher interpretation, not scheduling (Phuong, Le, & Pham, 2025). Taken together, these two studies indicate that negative washback in EFL contexts can originate from at least two structurally distinct sources, namely instrument-level narrowing (Phuong et al., 2025) and schedule-level density (the present study), a distinction that has not previously been made explicit in the washback literature and that carries different implications for remediation: the former calls for rubric or instrument redesign, while the latter calls for pacing and calendar-level intervention.

3.2.3. Ambivalent Washback: Strategic Compliance and Peer-Mediated Surface Learning

The most theoretically significant contribution of this study lies in the identification of strategic compliance as a distinct, ambivalent form of washback that does not map cleanly onto the positive-negative binary that has structured the field since Alderson and Wall (1993), Bailey (1996), and Cheng et al. (2004). These foundational frameworks conceptualize washback primarily as an individual cognitive or behavioral response to assessment, implicitly treating the learner as the unit of analysis. The present study's qualitative data complicate this assumption: students frequently relied on peer-shared summaries and collaborative preparation strategies, particularly within the collectivist learning culture characteristic of Indonesian higher education, to meet recurring assessment demands.

This finding adds a layer that recent washback scholarship has begun to gesture toward but has not yet fully theorized for formative, peer-mediated contexts. Recent rubric-washback research in Vietnam similarly found that lecturers' assessment practices could produce a narrowing effect, where teaching and evaluation converged on the most easily measurable, surface-level criteria, which is conceptually parallel to the present study's finding, except that in the Vietnamese case the narrowing originated from the teacher's interpretive practice, whereas in the present case it originated from peer networks operating around, rather than through, the instructor. This is an important structural distinction: ambivalent washback, as identified here, is socially distributed among learners rather than instructor-mediated, a dynamic not captured by prior washback models built around the teacher-student-test triad (Phuong et al., 2025).

ISSN: 2797-3174

The practical consequence is that high positive washback scores, as measured through self-report instruments, did not always correspond to deep, independent learning. In some cases, they reflected surface-level compliance that was nonetheless socially reinforced and experienced subjectively as positive. This raises a measurement concern. As recent washback scholarship has emphasized, the character of washback is forged through the interplay of policy design, institutional culture, teacher professionalism, and student literacy rather than predetermined by the test alone, and the present finding extends this argument by showing that peer culture itself can function as a mediating variable that inflates self-reported positive washback without necessarily producing the autonomous, self-regulated learning that flipped learning pedagogy aims to cultivate (Cheng, 2026).

This finding exposes a methodological limitation common to self-report washback measures more broadly: such instruments may capture behavioral completion (assignments submitted, quizzes attempted, participation recorded) without adequately capturing cognitive depth. Building on this distinction, the present study proposes *ambivalent washback* as a distinct theoretical category, describing assessment-induced behavior that satisfies the surface demands of assessment without necessarily promoting deep, autonomous learning. This category sits conceptually between the positive and negative poles of classical washback theory and is particularly salient in high-density, collectivist learning environments such as the Indonesian flipped learning context examined here, where prior single-instrument or individually-focused washback frameworks have limited explanatory power.

3.3. Implications

This study carries several theoretical and practical implications. Theoretically, the findings extend washback theory beyond the high-stakes testing contexts examined by into formative, assessment-intensive flipped learning environments, demonstrating that washback operates meaningfully even when individual assessment events are low-stakes but occur with high frequency. The introduction of ambivalent washback as a distinct theoretical category further refines the field's conventional positive-negative binary by accounting for surface-level compliance behaviors that satisfy assessment demands without producing deep autonomous learning. The findings also highlight a methodological limitation of self-report washback instruments, which tend to capture behavioral completion rather than cognitive depth, and therefore support the case for methodological pluralism in future washback research, including the integration of behavioral observation and learning analytics data. Furthermore, the social mediation of assessment responses through peer networks points to the relevance of sociocultural learning theory, particularly Vygotsky's (1978) concept of the zone of proximal development and Lave and Wenger (1991) notion of communities of practice, both of which foreground the role of collaborative social interaction in shaping how individuals respond to and learn from evaluative demands.

Practically, the findings indicate that flipped learning assessment generally promotes positive washback and is broadly aligned with the student autonomy and experiential learning goals of Indonesia's MBKM framework. However, negative washback in this study was associated primarily with assessment scheduling and workload concentration rather than with task design itself, suggesting that the problem is structural and temporal rather than inherent to the assessment instruments used. Instructors should therefore prioritize improvements in deadline distribution across the semester, provide clearer and more explicit assessment criteria to reduce uncertainty-driven compliance, and incorporate reflective activities that allow monitoring of the quality of student engagement rather than relying solely on quiz completion rates and scores as proxies for learning. At the policy level, the study supports the broader implementation of flipped learning within the MBKM framework while cautioning that positive washback outcomes are not automatic and depend on deliberate design decisions. Policymakers and institutional leaders should attend not only to pedagogical model adoption but also to assessment scheduling, instructional transparency, and digital equity. Unequal access to internet connectivity and technological resources risks producing washback inequities, where students with limited connectivity are structurally disadvantaged in meeting the recurring pre-class assessment demands that flipped learning requires, making digital infrastructure investment a critical institutional prerequisite for equitable and effective flipped learning implementation.

3.4. Limitations

This study was conducted within a single English Language Education program at one public university in North Sumatra, Indonesia; therefore, the findings should be interpreted within this limited context. The absence of institutional and disciplinary diversity restricts the generalizability of the results to other educational settings. In addition, the strategic compliance pattern identified in the qualitative findings may reflect collectivist learning cultures and may not be directly transferable to different cultural contexts.

Although the quantitative sample size (N = 60) met the minimum requirement based on G Power analysis, it remained relatively modest for PLS-SEM analysis. Several indicators, particularly NW4 and PW6, also showed relatively lower factor loadings, suggesting that items related to digital access and academic self-confidence may

ISSN: 2797-3174

require further refinement. The unequal gender distribution (70% female and 30% male) also limits the ability to examine gender-based washback differences, as discussed by Green (2007).

The qualitative strand involved only six participants, meaning that the ambivalent washback findings should be interpreted as analytical insights rather than statistically generalizable conclusions. Furthermore, the use of a cross-sectional design limits the study's ability to examine how washback perceptions develop over time. Longitudinal research across multiple semesters would provide a stronger understanding of the evolving nature of washback in flipped learning environments.

4. Conclusion

This study demonstrates that flipped learning creates a high-density washback environment in which recurring assessment activities shape students' study behavior, motivation, and learning strategies more intensively than conventional assessment contexts. The findings show that flipped learning generated stronger positive washback than negative washback, particularly in promoting learning independence, pre-class preparation, and critical engagement. However, qualitative evidence revealed that positive washback perceptions were sometimes driven by strategic compliance, where students fulfilled assessment demands collectively without fully engaging in deep autonomous learning. This pattern points to the need to expand washback theory beyond its traditional positive-negative binary to include ambivalent washback, referring to assessment-induced behaviors that are neither fully beneficial nor harmful, and that are particularly salient in collectivist learning cultures such as the Indonesian higher education context examined here. Practically, these findings suggest that assessment design in flipped learning should attend carefully to workload distribution, clarity of assessment criteria, and the role of peer networks in shaping learning behavior. Future research should examine ambivalent washback across disciplines and cultural contexts, investigate whether long-term exposure to flipped learning modifies its washback profile, and employ observational and learning analytics methods to capture cognitive depth beyond what self-report instruments can measure.

Author Contributions

All authors contributed substantially to the conception and design of the study, data collection, analysis and interpretation of data, manuscript drafting, and final approval of the submitted version.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Declaration on AI Use

Artificial intelligence (AI) tools were used solely to assist language editing and manuscript refinement. All interpretations, analyses, and final content remain the responsibility of the authors.

References

Abeysekera, L., & Dawson, P. (2015). Motivation and cognitive load in the flipped classroom: Definition, rationale and a call for research. *Higher Education Research & Development, 34*(1), 1–14. https://doi.org/10.1080/07294360.2014.934336

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement, 45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115–129. https://doi.org/10.1093/applin/14.2.115

Alqahtani, F. (2021). The impact of language testing washback in promoting teaching and learning processes: A theoretical review. *English Language Teaching, 14*(7), 21–31. https://doi.org/10.5539/elt.v14n7p21

Ardila, A. (2016). L. S. Vygotsky in the 21st century. *Psychology in Russia: State of the Art, 9*(4), 4–15. https://doi.org/10.11621/pir.2016.0401

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*(3), 257–279. https://doi.org/10.1177/026553229601300303

Bishop, J. L., & Verleger, M. A. (2013). *The flipped classroom: A survey of the research*. *ASEE Annual Conference & Exposition Proceedings*. https://doi.org/10.18260/1-2--22585

ISSN: 2797-3174

Braun, V., & Clarke, V. (2021). *Thematic analysis: A practical guide*. SAGE Publications.

Cheng, L. (2026). Reimagining the washback ecosystem: New directions from Asia's high-stakes testing landscapes. *Language Testing in Asia, 16*(1), 10. https://doi.org/10.1186/s40468-026-00424-w

Cheng, L., Watanabe, Y., & Curtis, A. (2004). *Washback in language testing: Research contexts and methods*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410609731

Creswell, J. W., & Creswell, J. D. (2023). *Research design: Qualitative, quantitative, and mixed methods approaches* (6th ed.). SAGE Publications.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education* (Vol. 25). Cambridge University Press & Cambridge ESOL.

Guba, E. G., & Lincoln, Y. S. (2000). Paradigmatic controversies, contradictions, and emerging confluences. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 163–188). SAGE Publications.

Guetterman, T. C., Creswell, J. W., & Fetters, M. D. (2015). Integrating quantitative and qualitative results in health science mixed methods research through joint displays. *Annals of Family Medicine, 13*(6), 554–561. https://doi.org/10.1370/afm.1865

Hair, J. F., Henseler, J., Ringle, C. M., & Sarstedt, M. (2022). *A primer on partial least squares structural equation modeling (PLS-SEM)* (3rd ed.). SAGE Publications.

Harrison, M. F. (2025). The effect of the flipped learning model on grammar test scores and student perceptions in a South Korean EFL setting. *Asian-Pacific Journal of Second and Foreign Language Education, 10*(1), 29. https://doi.org/10.1186/s40862-025-00332-1

Hasan, M. Z., Abidin, R. A. Z., Hashim, H. I. C., Yaacob, T. Z., & Fuzi, N. M. (2024). Revitalizing education in the 21st century: The evolution of the flipped classroom. In *Flipped classrooms and learning* (pp. 63–86). Springer.

Indonesian Ministry of Education, Research, and Technology. (2020). *Buku panduan Merdeka Belajar Kampus Merdeka*. Direktorat Jenderal Pendidikan Tinggi.

Khan, A., Hassan, N., & Cheng, L. (2025). Investigating the contextual factors mediating washback effects of a learning-oriented English language assessment in Malaysia. *Language Testing in Asia, 15*(1), 20. https://doi.org/10.1186/s40468-025-00359-8

Khayat, M. (2024). The effectiveness of the flipped learning approach on classroom perception and academic attitude in university students. *Research & Development in Medical Education, 13*(1). https://doi.org/10.34172/rdme.33224

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

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815355

Li, Z., & Li, J. (2022). Using the flipped classroom to promote learner engagement for the sustainable development of language skills: A mixed-methods study. *Sustainability, 14*(10), 5983. https://doi.org/10.3390/su14105983

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. SAGE Publications.

Lo, C. K., & Hew, K. F. (2017). A critical review of flipped classroom challenges in K–12 education: Possible solutions and recommendations for future research. *Research and Practice in Technology Enhanced Learning, 12*, 4. https://doi.org/10.1186/s41039-016-0044-2

Montasseri, Z., & Ahmadi, A. (2022). Washback effects of an interactional competence checklist. *Teaching English Language, 16*(2), 285–314.

O'Flaherty, J., & Phillips, C. (2015). The use of flipped classrooms in higher education: A scoping review. *The Internet and Higher Education, 25*, 85–95. https://doi.org/10.1016/j.iheduc.2015.02.002

Patton, M. Q. (2015). *Qualitative research and evaluation methods* (4th ed.). SAGE Publications.

Phuong, H. Y., Le, T. T., & Pham, T. T. (2025). Assessment for learning or assessment for scoring? Washback of analytic rubrics in Vietnamese EFL classrooms. *Language Testing in Asia, 15*(1), 32. https://doi.org/10.1186/s40468-025-00371-y

Ringle, C. M., Wende, S., & Becker, J.-M. (2022). *SmartPLS 4* [Computer software]. Retrieved from https://www.smartpls.com

Ruengtam, P. (2026). Reframing flipped studio learning for transformative pedagogy in architecture and built environment education. *Frontiers in Education, 11*, 1726937. https://doi.org/10.3389/feduc.2026.1726937

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Strelan, P., Osborn, A., & Palmer, E. (2020). The flipped classroom: A meta-analysis of effects on student performance across disciplines and education levels. *Educational Research Review, 30*, 100314. https://doi.org/10.1016/j.edurev.2020.100314

ISSN: 2797-3174

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Zainuddin, Z., Chu, S. K. W., & Perera, C. J. (2024). Flipped classroom. In *Gamification in a flipped classroom: Pedagogical methods and best practices* (pp. 11–65). Springer. https://doi.org/10.1007/978-981-97-2219-8

ISSN: 2797-3174