Journal of Information and Organizational Sciences
Volume 49, Number 2 (2025)
Journal homepage: jios.foi.hr

JIOS

DOI: 10.31341/jios.49.2.9
Open Access

UDC 37.091.33-047.37
Original Scientific Paper

Evaluating The Flipped Classroom Approach in Computer Science Curricula

Ruben Picek[1] and Samuel Picek[2*]

[1]Faculty of Organization and Informatics, University of Zagreb, Varaždin, Croatia
[2]InputLayer GmbH, Zürich, CH
[*]Correspondence: sam@inputlayer.ai

PAPER INFO

Paper history:
Received 14 May 2025
Accepted 18 November 2025

Citation:
Picek, R. and Picek, S. (2025). Evaluating The Flipped Classroom Approach in Computer Science Curricula. In Journal of Information and Organizational Sciences, vol. 49, no. 2, pp. 313-327

ABSTRACT

Active, technology-supported learning accelerated during and after COVID-19, yet evidence from non-programming computer science courses remains limited. This paper contributes (i) a focused review of flipped classroom (FC) studies in CS program (2020-2024) and (ii) a three-year case study of how the flipped classroom enhances the teaching of IT Service Management (ITSM) as a discipline in the computer science program in an online university environment, during and after the COVID-19 pandemic. The FC design combined pre-lecture micro-videos and auto graded quizzes with in-classroom clarification and post classroom activities (project).

Using LMS telemetry, course outcomes, and an end of semester survey across three academic years (2021/22-2023/24), we examined engagement-achievement links with non-parametric, rank based correlations (Spearman ρ), regularized logistic regression, and comparisons across empirically defined engagement tertiles. Results show consistent, practically meaningful associations between quality weighted engagement (quiz participation and performance) and both passing and final grades, with survey perceptions aligning to the behavioral signals. While strictly non-causal, the pattern is robust across methods and suggests actionable uses: early identification of at-risk students and design guidance that emphasizes short, well scaffolded videos and steady formative assessment.

Keywords: flipped classroom, flipped learning case study, computer science education, IT Service Management (ITSM), learning analytics, engagement, logistic regression, experience and achievements using flipped classroom

1. Introduction

Digital technologies and the COVID-19 period accelerated the adoption of online and blended teaching models in higher education. Approaches such as flipped learning, problem based learning, and project based learning are now widely used, supported by Learning Management Systems (LMS) that provide continuous access to resources and traceable engagement data.

The flipped classroom is an innovative teaching and learning strategy where students first engage with materials (e.g., short videos and quizzes) before class and then use class time for guided practice and discussion. Prior studies in computing report that this model is associated with greater student activity and opportunities for interaction, while also noting variability across courses, teachers, and implementations (e.g., O'Flaherty & Phillips, 2015; Long et al., 2017; Sosa Díaz et al., 2021; Gong et al., 2023). Implementations range from fully flipped courses to selectively flipped units, often relying on 5-10 minute videos, low-stakes pre-class assessments, and collaborative in-class tasks.

Despite growing use in programming and other CS subjects, evidence from non-programming CS courses remains limited. This paper contributes a focused review of flipped-learning studies in computer science (2020-2024) and a three year case study (2021-2023) of an IT Service Management (ITSM) course delivered in online/blended modes. Our analyses rely on observational data; we examine associations rather than causal effects.

We address two questions:

- RQ1: To what extent is engagement in flipped activities associated with achievement in an ITSM course?
- RQ2: What are students' perceptions and experiences of learning CS topics through flipped learning?

2. Methodology

The goal of this research is to provide both a theoretical and practical contribution. First, we conducted a comprehensive literature review to identify and synthesize key factors related to implementing the flipped classroom (FC) in higher education computer science (CS) programs. Second, we carried out a case study applying the FC approach in the IT Service Management (ITSM) course taken in the final year of the vocational study program Information Technology and Business Digitalization at the Faculty of Organization and Informatics, University of Zagreb. The case study spans three academic years: 2021/22, 2022/23, and 2023/24. A flipped learning design was developed for the course, covering both conceptual and assessment components.

2.1. Literature review

We followed the three phase approach of Divjak et al. (2022): P1 - Extraction, P2 - Abstract Analysis, and P3 - Detailed Examination. We searched Scopus, Web of Science Core Collection (WoS), and the IEEE Xplore Digital Library on 31 May 2024.

Time window and fields. To capture developments during and after the COVID-19 period, we limited the search from January 2020 to May 2024. Searches targeted Title, Abstract, and Keywords (using database-specific field tags).

Query terms. The core concept combined the flipped classroom with the CS domain. Representative queries were:

- *Scopus:* `TITLE-ABS-KEY ( flipped AND classroom AND computer AND science ) AND PUBYEAR > 2019 AND LIMIT-TO ( SUBJAREA, "COMP" )`
- *WoS:* `TI=(flipped AND classroom AND computer AND science) OR AB=(flipped AND classroom AND computer AND science) OR AK=(flipped AND classroom AND computer AND science); Timespan: 2020-01-01 to 2024-05-31; Category: Computer Science-Interdisciplinary Applications`
- *IEEE:* `(("Document Title":flipped AND "Document Title":classroom AND "Document Title":computer AND "Document Title":science) OR ("Abstract":flipped AND "Abstract":classroom AND "Abstract":computer AND "Abstract":science) OR ("Author Keywords":flipped AND "Author Keywords":classroom AND "Author Keywords":computer AND "Author Keywords":science)) Timespan:2020-2024`

Eligibility criteria. We included peer reviewed journal articles and conference papers in English that (i) focus on higher-education CS or closely related computing courses, and (ii) empirically examine FC implementations, experiences, outcomes, or design. We excluded editorials, posters, theses, and papers outside higher education or outside CS.

Screening and deduplication. Records from all sources were merged and deduplicated (by DOI/title). Titles/abstracts were screened in P2, and full texts examined in P3, with disagreements resolved by discussion.

Yield. The searches returned 112 papers (Scopus 80; WoS 12; IEEE Xplore 20). After deduplication and Phase-1 screening (abstract review), 36 studies remained (Scopus 33; WoS 1; IEEE Xplore 2). If a reader is interested in obtaining the full list of papers analyzed in this study, they may either replicate the search queries or request the author to provide the corresponding BibTeX files.

2.2. Case study

We implemented the flipped classroom (FC) in IT Service Management (ITSM), a final year course in the undergraduate vocational program *Information Technology and Business Digitalization* at the Faculty of Organization and Informatics, University of Zagreb. The case study includes 3 generations of students in academic years 2021/22-2023/24; N = 147; enrollments: 45, 49, 53 students. Delivery was fully online in 2021/22 (Moodle + BigBlueButton) and blended in 2022/23 and 2023/24.

To achieve learning outcomes, in addition to attending lectures, students in a team of 3 members have to solve weekly assignments (projects) related to the concepts of IT service management, which are based on ITIL best practices and the service lifecycle. After the introductory topics, the classes were organized using the FC approach. The FC activities can be categorized into three segments: pre-lecture, in-classroom, and post-classroom activities.

Pre-lecture activities: Over 10 weeks, students had to prepare in advance for each lecture by watching the prepared video content (public YouTube videos uploaded by IT consultants who conduct ITIL training or courses for ITIL certification preparation) and accessing assessments before the lecture, set on the Learning Management System (Moodle). The course follows a model of continuous monitoring/assessment (applicable only to the exam period for continuous assessment), according to which a student can earn up to 10% of the total points for participating in the FC activities. The use of FC was not mandatory, and with each assessment students could earn up to 1 point (10 points in total). Students who did not participate in FC at all did not receive any points from this category during the continuous assessment exam period. All other exam periods followed the traditional oral and written format. The video materials used for FC were also available alongside the exam materials, so students were able to watch them, but of course now without the possibility of completing the assessments.

Each assessment was based on 3 questions (the database for each assessment contained 6 questions), in which the student had to show that he or she understood the new concepts to be considered in the lecture that would follow and apply them in team based weekly tasks.

In-classroom activities: Immediately before the lecture, the teacher analyzed the success of solving each (of the 6) questions and paid special attention to those parts of the material that had the lowest percentage of accuracy.

Post-classroom activities: After each lecture, students solved weekly assignments as a team, which were evaluated at the end of the semester.

In this course, students' knowledge was assessed using multiple techniques and instruments, providing a comprehensive evaluation of both theoretical understanding and practical application. Assessment methods included two written online quizzes (colloquia), which resembled certification tests but were simplified. The final colloquium also included oral questions to further evaluate students' comprehension. Additionally, students, as a team, completed a project in a Virtual IT Company, which involved weekly tasks aligned with the lectures. During the semester, each team presented one completed task according to the schedule. Both the team tasks and presentations were graded, and the project concluded with a team-based oral assessment covering the entire project. Summative evaluation followed the continuous-assessment model: project 40%, two written colloquia 50%, and FC activity assessment points 10%.

For our case study analysis, we derived a set of engagement metrics from FC activity assessment and behavior:

- **Base metrics:** participation rate, number of tests attempted, average test score, total test score, composite tests engagement score (participation and performance), and three consistency measures (overall, stability, early-late).
- **Percentile analogues:** per-metric within group percentiles to enable scale free comparisons.

Achievement was captured by **final grade** (1-5 scale) and **passed** (binary) variables.

We conducted descriptive and inferential analyses, with test selection driven by distributional diagnostics:

1. **Distribution checks.** We assessed normality with the Shapiro-Wilk test. Given pervasive non-normality in outcomes, all bivariate associations were estimated with Spearman's rank correlation (ρ). We checked two-sided p-values and interpreted effect sizes by the magnitude of $|\rho|$ (≈ 0.10 small, ≈ 0.30 moderate, ≥ 0.50 large).
2. **Engagement-achievement correlations.** For each engagement metric (base and percentile variants) paired with each achievement metric, we computed Spearman's ρ with two-sided significance tests and summarized effect sizes via $|\rho|$.
3. **Grouped comparisons by engagement level.** To examine distributional differences in outcomes across engagement strata, we formed Low/Middle/High groups using the 33rd and 67th percentiles

of the composite engagement score within the pooled sample (i.e., ≤33rd, 33rd–67th, >67th). Because outcomes were non-normal, we used Kruskal-Wallis tests with epsilon-squared (ε^2) as effect size (≈0.01 small, 0.06 medium, 0.14 large). We report group n, means/medians/SDs, p-values, ε^2, and lay summaries (e.g., pass-rate differences).

4. **Pass/fail prediction (operational framing).** To assess whether engagement signals separate pass vs. fail cases, we fit an L2-regularized logistic regression with standardized percentile-scaled engagement predictors (to reduce collinearity and unit effects) and a passed variable as the target. We used an 80/20 stratified split (seed = 42) and 5-fold stratified CV. Metrics included accuracy, precision, recall, F1, and AUC-ROC; we also report the cross-validated AUC mean ± SD as a stability check. Coefficients and odds ratios are presented for interpretability (positive coefficients indicate higher odds of passing, holding other features constant).

Other notes on the case study analysis:

- **Multiple comparisons.** We report unadjusted p-values and focus interpretation on effect sizes and consistency across methods (correlations, group tests, classification).
- **Robustness.** Emphasis is placed on rank-based statistics and Spearman's ρ where normality is not supported.
- **Non-causal interpretation.** All analyses are observational and associational; language and conclusions avoid causal attribution.

2.3. Perception survey

At the end of the last lecture of the semester (in all three academic years), and before the evaluation of the project and the final exam, an online survey was conducted among all enrolled undergraduate students to gather their opinions about their experience using the flipped classroom. The instrument comprised 16 statements across four sections, adapted from prior studies to ensure content validity: 7 Likert-type items (1-10), 7 multiple-choice items, and 2 open-ended prompts. Items covered (i) familiarity with modern teaching concepts, (ii) perceptions of flipped classroom (FC) materials and experience, (iii) views on FC assessments, and (iv) self-reported impact and suggestions. Participation was voluntary and approved by the Faculty Ethics Committee.

3. Results

This section reports two complementary strands: (i) a focused review of flipped classroom (FC) studies in computer science programs (2020-2024) and (ii) a three year case study of an online/blended IT Service Management (ITSM) course that adopted an FC approach. We first synthesize the literature to contextualize our setting, and then present quantitative and qualitative evidence from the case study.

3.1. Review of flipped classroom studies in computer science programs

This review synthesizes recent work on flipped classrooms (FC) in university computer science programs from various perspectives. Across studies, a common pattern emerges: short pre-class videos paired with brief quizzes, in-class clarification and active work, and post-class practice or projects (e.g., Aldalur et al., 2022; Gong et al., 2023). Videos are typically concise (≈5-10 minutes), often hosted on YouTube, though length varies by topic. Evidence from a STEM meta-analysis indicates a modest positive effect of FC over traditional formats, especially when pre-class checks, mixed individual/group in-class activities, and post-class quizzes are used (Gong et al., 2023). From a program and instructor perspective, Aldalur et al. highlight those outcomes and workload depend on scope (single course vs. multi-course rollout). Students report higher engagement and continuity, while teachers value richer contact time but face substantial upfront effort, particularly for producing clear, concise videos. Design and technology choices matter: LMS-based quizzes and analytics are common, and studies often blend FC with gamification to sustain motivation (e.g., Algayres et al., 2021; Olivindo et al., 2021). Challenges recur, balancing depth with short videos, supporting large or heterogeneous groups, and maintaining interaction online (Steinmaurer & Gütl, 2023; Ossovski, 2022). Instructor adoption is shaped more by experience and enabling conditions (time, support, tools) than by age or gender (Bakheet & Gravell, 2021b,c).

Most empirical work centers on programming (CS1/CS2) but also spans databases, software engineering/testing, distributed systems, data management, and broader STEM (Aldalur et al., 2022; Araujo et al., 2020; Zamora-Hernandez et al., 2022). Design guidance converges on careful learning design: align FC elements with outcomes, scaffold pre-class work, structure active in-class tasks, and use analytics for timely feedback (Mithun & Luo, 2020). The literature supports FC as a viable approach in CS when thoughtfully designed and institutionally supported, while noting practical constraints, chiefly instructor time, video quality and length, and sustaining student engagement outside class. Table 1 presents the main findings of the literature review.

Category	Findings
Student Impact	FC improves the perception of learning, engagement, and continuity across academic years (Aldalur et al., 2022). Students prefer flexibility: they can pause, rewatch, and come to class prepared (Gong et al., 2023). Quizzes, group and individual activities during class enhance performance (Gallaugher, 2023). Gamification and rewards positively affect engagement (Algayres et al., 2021). Some skepticism remains regarding engagement without direct classroom instruction (Urquiza-Fuentes, 2023)
Teacher Perspective	FC improves the teaching experience but demands significant resources (Aldalur et al., 2022). Creating video content is the most demanding part (Bakheet & Gravell, 2021b). Teaching experience (more than age or gender) influences FC adoption (Bakheet & Gravell, 2021c). Eleven factors were identified that influence FC adoption, including technical self-efficacy and student motivation (Bakheet & Gravell, 2021c).
Pedagogical Design and Course Structure	Successful FC implementation depends on careful instructional design and material adaptation (Aldalur et al., 2022). All studies follow a similar structure: pre-class (video + quiz), in-class (discussion, practice) and post-class (assignments) (Gong et al., 2023). Iterative adaptation is needed based on subject, students, and available resources (Aldalur et al., 2022). Video materials typically last 5–10 minutes (Mithun & Luo, 2020).
Technological and Organizational Aspects	The use of YouTube and other tools is common, but challenges include technical issues and content balancing (Steinmaurer & Gütl, 2023). Asynchronous-synchronous combinations (video + live stream) are well accepted (Aldalur et al., 2022). Face-to-face communication remains irreplaceable (Steinmaurer & Gütl, 2023). In large groups with heterogeneous prior knowledge, FC helps with customization (Ossovski, 2022).
Limitations and Challenges	Difficult content is sometimes better taught in traditional settings than with FC (Elgrably & Oliveira, 2022). Preparing video content and learning materials is time-consuming for teachers (Bakheet & Gravell, 2021a). Some students show lower enjoyment and higher absence rates (Bakheet & Gravell, 2021b). Barriers include gender, language, and time availability (Bakheet & Gravell, 2021b).

Table 1. Main findings of the conducted review.

3.2. Case study - analytic roadmap

Our dataset consists of LMS telemetry data, derived engagement metrics (base and percentile-scaled), and course outcomes, complemented by an end-of-semester survey. We first describe the variables and their empirical distributions, including normality checks and handling of missing data. We then analyze *engagement ↔ achievement* associations using correlation analysis. Next, we examine the predictive strength of those correlations to verify implementation of early interventions. We then assess distributional separation and compare outcomes across engagement tertiles, defined by empirical 32nd and 68th percentiles, using non-parametric tests and effect sizes. Finally, we triangulate the results with survey responses to align student perceptions with observed engagement and achievement patterns. All analysis can be found in the following Github repository: *https://github.com/rubenpicek/fc*.

3.2.1. Variable descriptions and measurements

All assessment variables summarize student activity across ten weekly flipped classroom assessments administered during weeks 4-13. Based on distributional diagnostics and our analytic stance, histograms and Shapiro-Wilk tests indicate non-normality for most variables (notably, participation and attempts are left-skewed; grades are discrete; pass is Bernoulli) (Figure 1a, Figure 1b, Figure 2). Accordingly, we prioritized non-parametric procedures throughout (Spearman's ρ for associations; Kruskal-Wallis for group contrasts).

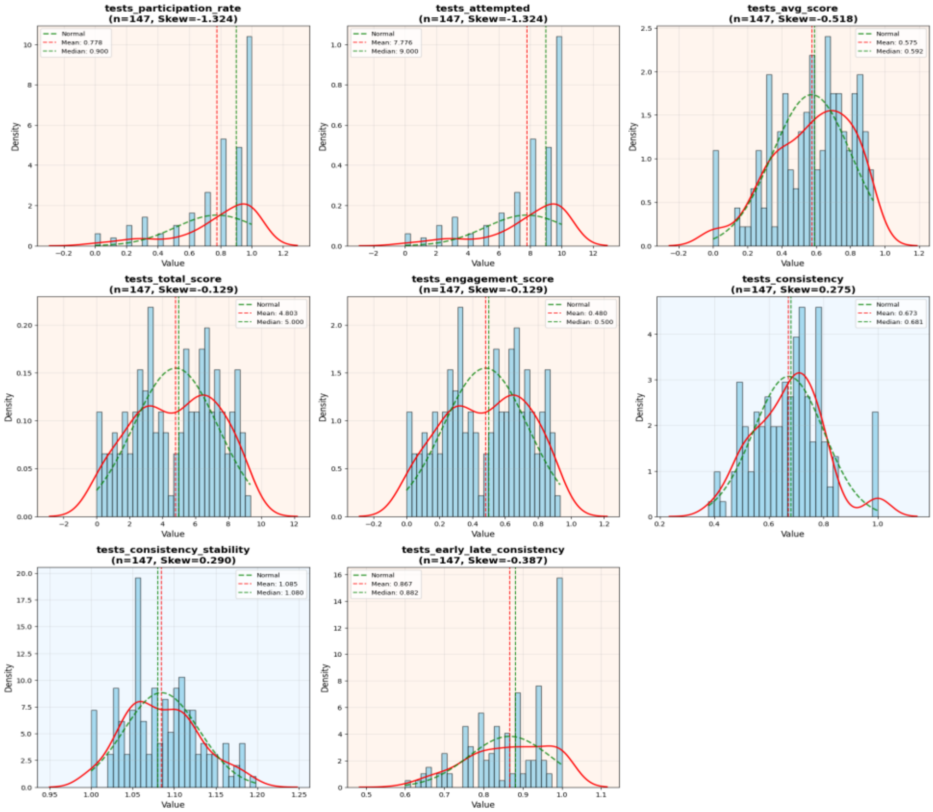


Figure 1a. Histograms of participation, performance and consistency metrics with normality assessment

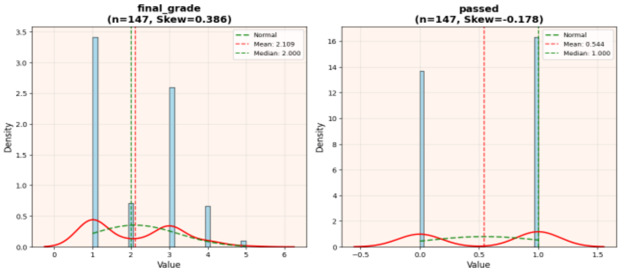


Figure 1b. Histograms of achievement metrics with normality assessment

Together, these variables capture complementary facets of flipped classroom engagement - whether students participated, how well they performed, and how steadily their performance evolved across the ten assessments - alongside two course outcomes (final grade and pass). The histograms and normality tests support our use of robust, rank-based statistics (Spearman's ρ; Kruskal-Wallis) across analyses. These distributional facts also anchor interpretation in the sections that follow (correlation mapping, predictive modeling, tertile comparisons, and survey triangulation), avoiding causal claims and focusing on association strength and practical separability.

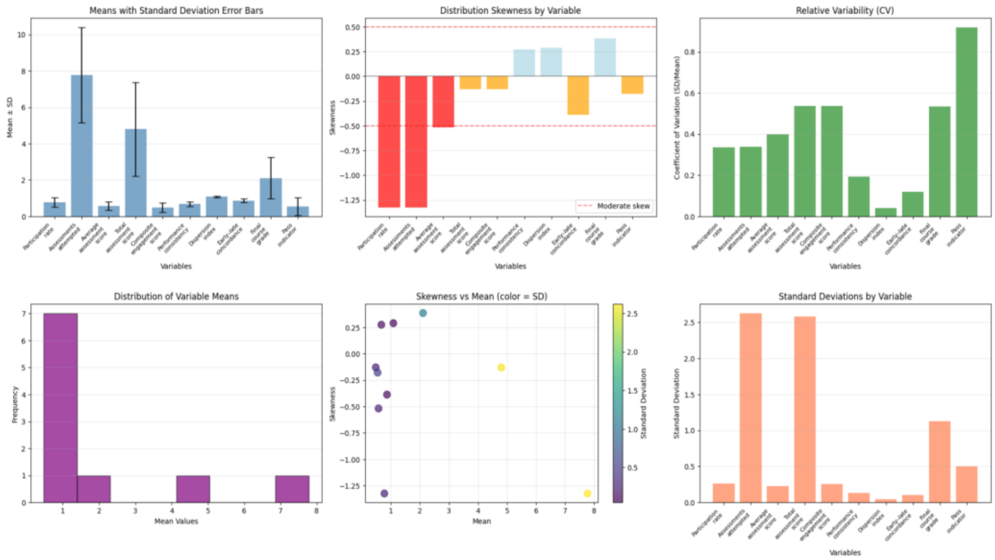


Figure 2. Dataset distribution summary statistics

Key Measurement Insights: Sample size: N = 147. All variables showed non-normal distributions (Shapiro-Wilk), with participation rate and assessments attempted exhibiting strong negative skew (< -1.0), indicating left-heavy engagement patterns. The pass rate was 54.4%, with students attempting an average of 77.8% of the ten assessments. The engagement score (participation × average score) combines frequency and quality of assessment activity, serving as the primary predictor for subsequent analyses.

3.2.2. Correlation analysis

We next examine how each engagement indicator associates with the achievement outcomes using Spearman's rank correlation (ρ). This choice follows the distributional diagnostics: variables are bounded and non-normal, making rank-based measures preferable to parametric alternatives. Spearman's ρ captures monotonic relationships without assuming linearity and is robust to outliers and scale differences. For interpretability, we show ρ in the correlation heatmaps (Figure 3a) and report ρ² as an effect-size analogue (Figure 3c) in the rank domain in the companion panels (that is, the proportion of variability in the ranks of the outcome that is monotonically associated with the ranks of the predictor).

Across engagement-outcome pairs, two-sided p-values indicated that core activity/performance metrics are highly significant ($p < .001$). Measures that combine frequency and quality - the total assessment score and the composite engagement index (participation rate × mean score) - show the strongest monotonic relationships with achievement (final grade: $\rho \approx 0.48–0.49$, $\rho^2 \approx 0.24-0.25$; pass/fail: $\rho \approx 0.36-0.37$, $\rho^2 \approx 0.13-0.14$; medium effects). Pure activity indicators (assessments attempted, participation rate) are also significant ($p \leq .001$) but smaller in magnitude (final grade: $\rho \approx 0.40$, $\rho^2 \approx 0.15$; pass/fail: $\rho \approx 0.29$, $\rho^2 \approx 0.08-0.09$). The average score sits between composite and activity-only measures (final grade: $\rho \approx 0.40$, $\rho^2 \approx 0.15-0.16$; pass/fail: $\rho \approx 0.31$, $\rho^2 \approx 0.08-0.09$). Raw temporal regularity features contribute little: raw consistency and consistency stability are almost negligible ($\rho = 0.07$; $p = 0.014$; $\rho^2 = 0.000$ / $\rho^2 = 0.006$), and early-late consistency has a weak relation (final grade: $\rho = 0.18-0.19$, $p = 0.014$, $\rho^2 = 0.047$; pass: $\rho = 0.14-0.15$, $p = 0.01$, $\rho^2 = 0.022$), indicating that once overall participation and mean score are taken into account, the timing pattern of performance (earlier vs. later assessments) adds minimal explanatory value.

Expressing engagement variables as percentiles (Figure 3b and Figure 3d) produces a modest uplift in association strength, with the improvement concentrated in activity indicators such as participation and attempts. The substantive ordering of predictors is unchanged - the composite engagement index and total assessment score remain the strongest correlates of both outcomes. Approximately two-thirds of the 32 engagement-outcome pairs are statistically significant at $p < .05$ under the percentile specification, mirroring the raw-scale results.

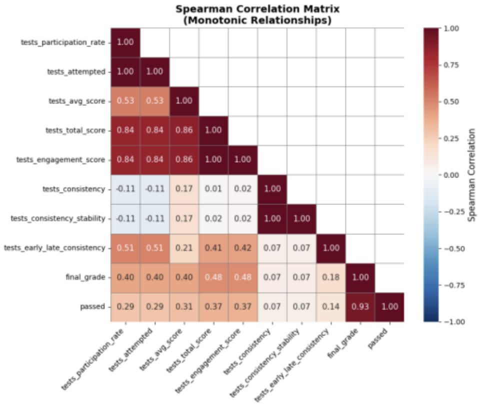


Figure 3a. Correlations of values

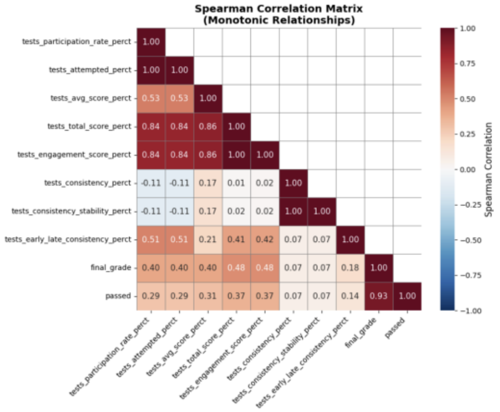


Figure 3b. Correlations of percentiles

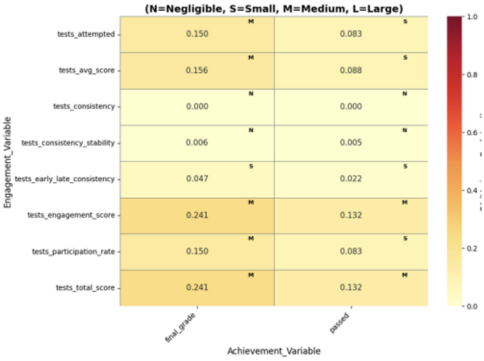


Figure 3c. Effect sizes of values

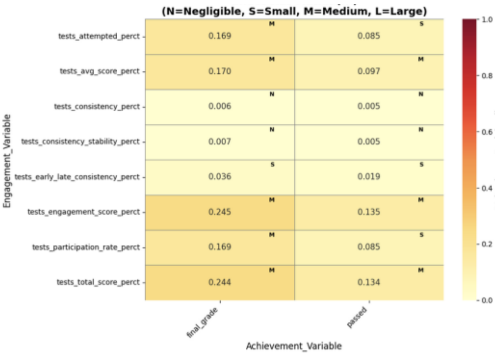


Figure 3d. Effect size of percentiles

3.2.3. Predictive modeling

Having established robust bivariate relationships between engagement signals and outcomes using Spearman's ρ (reported as both correlation coefficients and ρ^2 as rank-domain variance measures), the next step is to examine whether these signals retain unique, conditional value when considered together. Correlations are informative but inherently pairwise; they neither control for the strong interdependence among the engagement variables (e.g., participation, total score, composite engagement) nor indicate how their joint combination translates into actionable predictions. Because one of our outcomes is binary (pass/fail) and most predictors are non-normal, an L2-regularized logistic regression provides an appropriate multivariate framework: it models the log-odds of passing without assuming normality of predictors, quantifies direction and magnitude through interpretable odds ratios, and allows assessment of out-of-sample discrimination (ROC-AUC) for early-warning purposes. Regression enables us to move from "are these variables associated?" to "which signals add independent information, and how well do they work together to identify students at risk," while remaining strictly predictive rather than causal.

We estimated an L2-regularized logistic regression to predict passed boolean value, from the percentile-scaled engagement indicators (participation rate, attempts, average score, total score, composite engagement, and the three regularity measures). The modelling set comprised 147 complete cases (54.4% passed). Using a stratified 80/20 split preserved class balance (train 64/53 pass/fail; test 16/14).

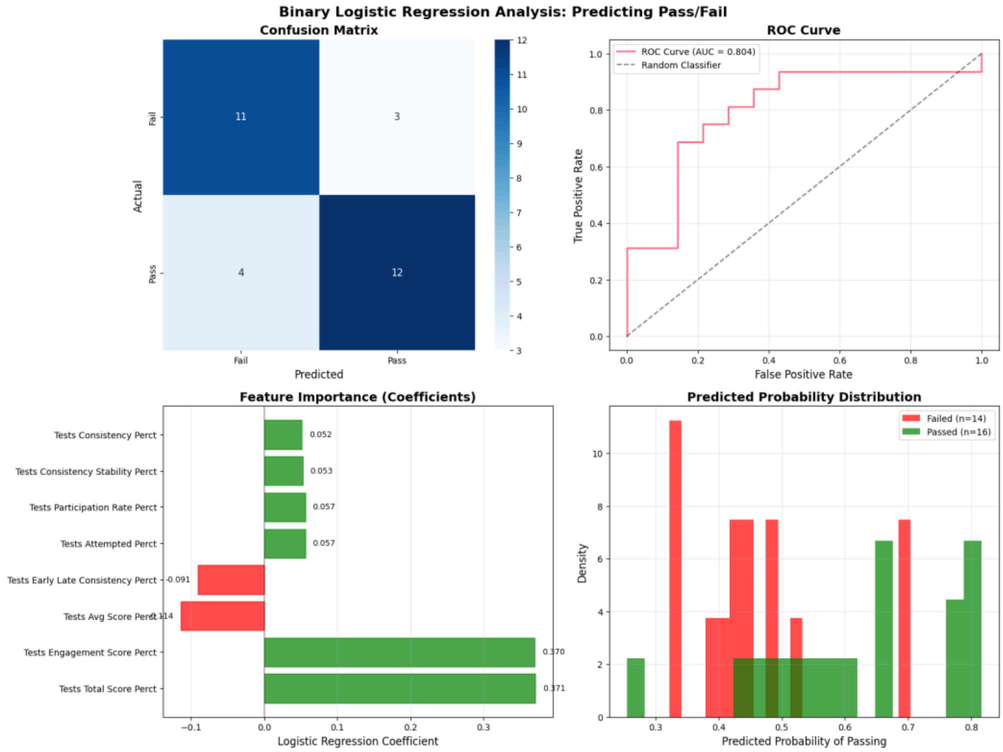


Figure 4. Binary classification model performance evaluation

The feature importance chart (Figure 4) reveals a key insight: quality matters more than quantity. The total assessment score and composite engagement are approximately three to four times more influential than simple participation counts. This suggests that merely completing assessments is insufficient - understanding and mastery, reflected in quiz scores, are crucial. On the held-out fold, the model achieved ROC-AUC ≈ 0.804 with accuracy ≈ 0.77-0.78 (23/30 correct; precision ≈ 0.77, recall ≈ 0.77, F1 ≈ 0.77) (Figure 4).

A stratified 5-fold cross-validation yielded a mean AUC ≈ 0.716, indicating consistent discrimination across folds (Figure 5). Model diagnostics support healthy generalization. Learning curves show the training and validation AUCs converging near 0.73 as sample size grows, with no widening gap - evidence against overfitting. The regularization path is smooth; performance plateaus for C in the 1-10 range, with no signs of dramatic under or over-penalization. Predicted probability distributions exhibit clear separation: most non-pass cases cluster around 0.30-0.50, while pass cases concentrate around 0.60-0.80. The model is slightly conservative (few extreme probabilities), which is desirable for calibrated early warning use.

Coefficient patterns align with the correlation results and provide an interpretable ranking of signals. Two predictors carry the largest positive weights (Figure 4): total assessment score (β ≈ 0.371; odds ratio ≈ 1.45 per percentile unit) and the composite engagement index (β ≈ 0.370; OR ≈ 1.45). Participation intensity remains beneficial once performance is controlled for (attempts and participation rate both have small positive coefficients, β ≈ 0.057). The regularity features contribute modestly: tests_consistency and consistency_stability are positive but small, while early-late consistency is slightly negative. Average score becomes weakly negative after conditioning on total score and composite engagement - an expected suppression effect given strong inter-correlations among activity/performance indicators. Taken together, these results indicate that joint quantity and quality engagement signals dominate predictive value; simple activity counts add only a small increment; timing regularity matters, albeit minimally. The ROC curve, stable learning curves, balanced confusion matrix, and coherent feature weights all point to a well-performing, well-calibrated model that is useful for risk identification (Figures 4 and 5).

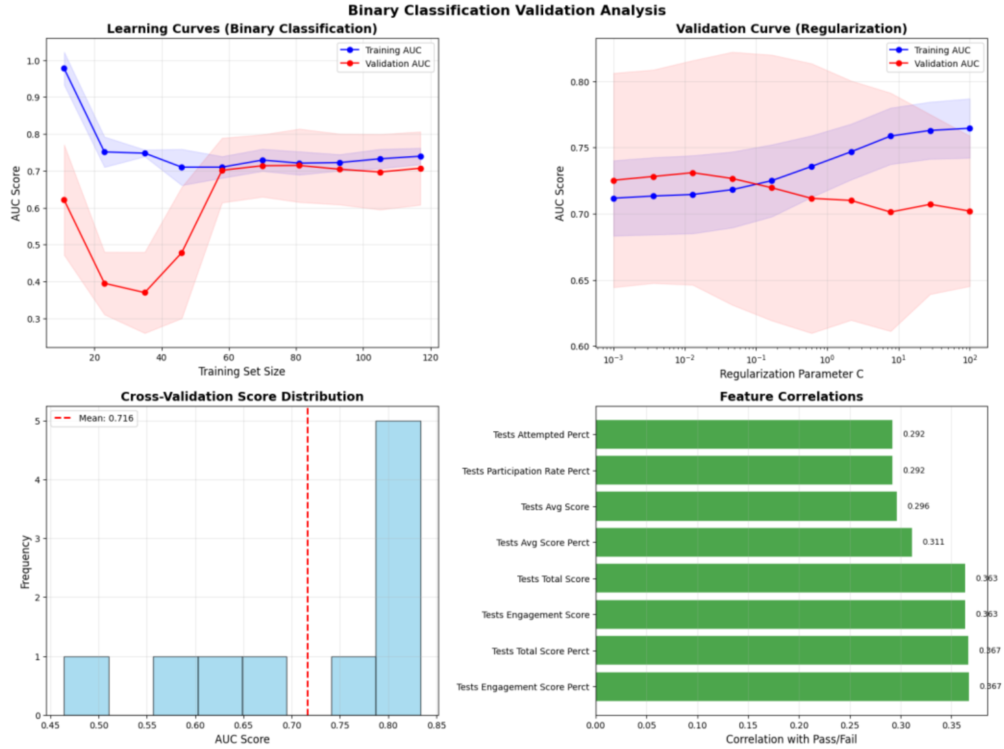


Figure 5. Model validation and generalization assessment

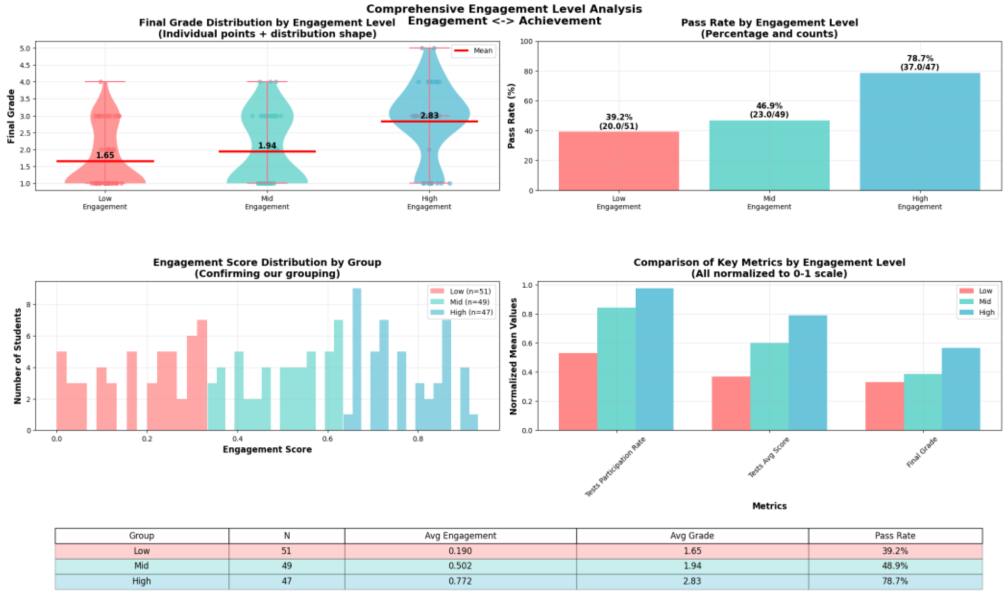
The model's accuracy (ROC-AUC ≈ 0.80) indicates that it correctly ranks students by risk approximately 80% of the time. If one randomly selects a passing and a failing student, the model would correctly identify which is which four out of five times. This performance is strong enough for practical use: flagging the bottom 20-30% of students for proactive outreach would capture most at-risk cases while keeping the intervention manageable.

3.2.4. Tertile comparison

To translate the earlier associations into a form that is directly interpretable for teaching practice, we segmented students into three engagement levels and compared outcomes across these groups. We used the composite engagement score and created approximate tertiles with empirical cut points at the 32nd and 68th percentiles (0.333 and 0.634) to avoid boundary ties. This yielded Low (n = 51), Mid (n = 49), and High (n = 47) groups with mean engagement scores of 0.190, 0.502, and 0.772, respectively.

Because all focal variables are non-normal, we tested group differences with Kruskal-Wallis and reported epsilon-squared (ε^2) as effect size (≈.01 small, ≈.06 medium, ≥ .14 large). The goal here is descriptive: to assess how strongly achievement varies across pragmatic engagement bands.
The tertile split produces clearly separated strata (manipulation check: $H = 129.78$, $p < .001$; $\varepsilon^2 = .887$). Achievement varies monotonically with engagement. Final grades increase from 1.65 (Low) to 1.94 (Mid) to 2.83 (High) (Figure 6); the overall difference is highly significant ($H = 27.43$, $p < .001$) with a large effect ($\varepsilon^2 = .179$). Pass rates show the same pattern - 39.2% (Low), 48.9% (Mid), 78.7% (High) - again significant ($H = 16.42$, $p < .001$) with a medium effect ($\varepsilon^2 = .102$). Process indicators confirm the segmentation: participation rate rises from 0.531 to 0.843 to 0.977 ($H = 86.71$, $p < .001$; $\varepsilon^2 = .588$), and average quiz score from 0.371 to 0.602 to 0.791 ($H = 90.55$, $p < .001$; $\varepsilon^2 = .628$).

In practical terms, students in the top engagement band earn, on average, 1.18 more grade points (on a 1-5 scale) and are about 39.5 percentage points more likely to pass than those in the bottom band. These differences are substantial and consistent with the correlation and regression findings, while remaining strictly associational.



Group	N	Avg Engagement	Avg Grade	Pass Rate
Low	51	0.190	1.65	39.2%
Mid	49	0.502	1.94	48.9%
High	47	0.772	2.83	78.7%

Figure 6. Engagement impact on academic performance

3.2.5. Survey triangulation

To triangulate findings across analytical methods, we administered a post-course perception survey (N=147) comprising 14 substantive items (excluding self-reported impact and suggestion). The instrument captured: familiarity with FC pedagogy, perceived learning benefits and materials quality, self-reported engagement behaviors, assessment experiences, and overall evaluation. Questions used mixed formats (Likert 1-10, categorical, count, percentage) administered in Croatian. Repository (https://github.com/rubenpicek/fc) provides complete definitions and survey questions. Two items (materials_quality, materials_relevance) formed a validated materials perception scale (α = 0.845, ρ = 0.706). Twelve individual indicators (num_learning_methods, knowledge_before, knowledge_after, learning_style_pref, quiz_difficulty, video_watching_freq, quiz_timing_pref, quiz_helpfulness, time_spent_learning, quiz_participation_rate, overall_satisfaction, perceived_fc_help) captured distinct FC dimensions.

Triangulation across methods:

Correlation ↔ Survey: Students rating materials favorably reported greater FC impact on grades (perceived_fc_help: $\rho = +0.35$, $p < .001$), mirroring correlation findings where composite engagement (participation × score) showed strongest association with final grade ($\rho \approx 0.48$–0.49). The parallel effect sizes - moderate correlations in both behavioral ($\rho^2 \approx 0.24$) and perceptual domains ($\rho^2 \approx 0.12$) - suggest students accurately perceive the achievement benefit of quality engagement.

Regression ↔ Survey: Quiz_participation_rate dominated satisfaction ratings ($\rho = +0.59$, $p < .001$) over materials perception ($\rho = +0.05$, ns), confirming regression results where total_score and composite_engagement carried largest coefficients (OR ≈ 1.45), while simple participation added minimal unique variance. Students' satisfaction thus tracks the conditional importance hierarchy revealed by multivariate modeling: what you achieve in assessments matters more than mere attendance.

Tertile ↔ Survey: Students self-reporting higher quiz_participation_rate also reported materials helped task completion ($\rho = 0.57$-$.59$) and exam preparation ($\rho = 0.61$-0.62). This aligns with tertile findings: High-engagement students (participation 0.98, score 0.79) passed at 78.7% versus Low-engagement (participation 0.53, score 0.37) at 39.2%. The magnitude of perceptual differences (0.57-0.62 correlations) mirrors the practical significance of behavioral tertile gaps (39.5 percentage-point pass-rate difference), validating that subjective utility aligns with objective stratification.

Cross-method convergence: All four methods independently identify quality-weighted participation (doing assessments well, not just often) as the dominant signal: correlation magnitude ordering, regression coefficient size, tertile pass-rate gaps, and survey satisfaction drivers converge on the same pattern. Survey perceptions validate behavioral telemetry: students who engaged more consistently both performed better objectively and perceived greater benefit subjectively. This multi-method alignment strengthens confidence in the engagement-achievement relationship while maintaining observational interpretation.

4. Discussion

This study extends flipped classroom (FC) evidence to a non-programming CS course (IT Service Management) and integrates four lenses-distributions, correlations, predictive modeling, and tertile contrasts-triangulated with a perception survey. The findings cohere around a simple message: quality-weighted engagement with the formative FC assessments tracks achievement in a strong and practically meaningful way, while timing regularity adds little. We interpret all links as associational, not causal.

Positioning in prior work. Our focused review (2020-2024) shows that FC research in CS is still dominated by programming courses and converges on a common pattern of pre-class micro-content with short quizzes, in-class clarification, and post-class practice. The present case adopts that pattern in ITSM and reaches conclusions consistent with the review: students value clear, concise pre-class materials; formative, auto-graded checks are workable at scale; and in-class time is reallocated to targeted explanation and application. The analyzed studies also indicate that FC approaches foster motivation and enhance students' acceptance and engagement (Aldalur et al., 2022; Algayres et al., 2021; Olivindo et al., 2021; Ossovski, 2022).

RQ1: Engagement and achievement. Non-parametric associations show that the strongest monotonic relationships with final grade and pass/fail are the total assessment score and a composite engagement index (participation rate × mean assessment score): for final grade, $\rho \approx 0.48$-0.49 ($\rho^2 \approx 0.24$-0.25); for pass/fail, $\rho \approx 0.37$-0.38 ($\rho^2 \approx 0.13$-0.14). Pure activity measures (attempted/participation) are smaller but clearly significant (e.g., final grade $\rho \approx 0.40$). Regularity features (consistency, stability, early-late balance) contribute only marginally once overall participation and average score are considered. Percentile scaling yields a small average uplift in ρ^2 without changing the ordering of predictors.

A regularized logistic model trained on percentile-scaled engagement indicators achieves test AUC ≈ 0.804 with ∼0.77 accuracy and cross-validated AUC ≈ 0.716, with stable learning curves and a smooth regularization path - evidence of a healthy, well-calibrated classifier. The largest positive coefficients are again total score and composite engagement (odds ratios ≈ 1.45 per standardized unit); participation intensity adds a modest positive increment; regularity features are small; average score turns slightly negative when total/composite are included, a suppression pattern consistent with their inter-correlations. Predicted probabilities separate pass/fail cases sensibly (fail ≈ 0.30-0.50; pass ≈ 0.60-0.80) and remain conservative rather than overconfident.

To translate these signals into practice, tertile comparisons using empirical 32nd/68th cut points show significant monotonic gaps: final grade rises from 1.65 (Low) to 1.94 (Mid) to 2.83 (High) (Kruskal-Wallis $p < .001$; $\varepsilon^2 \approx 0.18$), and pass rate from 39.2% to 48.9% to 78.7% ($p < .001$; $\varepsilon^2 \approx 0.10$). Process indicators confirm separation (participation $0.53 \rightarrow 0.84 \rightarrow 0.98$; average assessment score $0.37 \rightarrow 0.60 \rightarrow 0.79$). Practically, a student in the top engagement band earns about +1.18 grade points (on a 1-5 scale) and is approximately 40 percentage points more likely to pass than a student in the bottom band. Analyzed studies also confirm a general positive influence: FC improves students' final grades and overall performance (Bakheet & Gravell, 2021a), (Aldalur et al., 2022).

Practical use. These results support a simple, actionable workflow: compute a composite engagement index each week from LMS data; flag students in the bottom tertile or with predicted pass probability below ∼0.5 for a light-touch intervention (brief check-in, reminder to complete the next quiz, or targeted pointers to the specific video/quiz they missed). Because a one-unit increase (standardized - z-score) in total score or composite engagement raises the modeled odds of passing by ∼45% (OR = e^0.371 ≈ **1.45, Figure 4**), small, achievable increments-finishing one more assessment or raising an assessment average on the next attempt-can yield meaningful gains. At the course level, teachers can (i) allocate feedback and tutorial time first to low-engagement students, (ii) keep micro-videos tightly aligned with the quiz that follows, and (iii) consider modest incentives (time allowances, brief feedback, limited re-attempts) that encourage steady participation rather than last-minute bursts. Several suggestions for effective FC design are also supported by previous work (Bakheet & Gravell, 2021a), (Sibia & Liut, 2022).

RQ2 - Student perceptions. Survey results align closely with the telemetry and outcomes. Students report that FC increases interest in weekly topics and that the videos materially help them follow lectures and prepare for exams; team-based follow-ups are also viewed as useful. This aligns with research showing that the FC approach improved students' perception of their learning, increased engagement, and ensured valuable continuity across academic years (Aldalur et al., 2022; Gong et al., 2023; Gallaugher, 2023). Behaviors are imperfect-many students skim or do not finish videos-and fewer than 10% consistently watch all materials before attempting the quiz. Still, students who report fuller viewing and more frequent assessment access tend to pass more often and earn higher grades, mirroring the quantitative patterns. Satisfaction with assessment design is high; requests focus on more time, richer feedback, occasional native-language videos, and (limited) multiple attempts rather than on different content.

Implications for design and teaching. The converging evidence suggests several actionable priorities for FC in non-programming CS courses:

- **Weight quality and regularity of formative work.** Total assessment score and the composite engagement index carry the most information; simple counts help but are insufficient.
- **Use engagement signals for early support.** The logistic model's calibrated probabilities (AUC ≈ 0.80) and the tertile gaps provide workable thresholds for nudging and tutorial allocation-strictly for prediction, not attribution.
- **Tune pre-class materials and assessment flow.** Keep micro-videos concise (Mithun & Luo, 2020); align each quiz tightly with its video; consider allowing limited re-attempts and providing short, descriptive feedback. Requests for more time and occasional native-language content are reasonable enhancements.
- **Communicate expectations.** Some students still attempt assessments without finishing videos; clearer guidance and light-touch requirements (or small grade weight) can encourage better sequencing.

Limitations. The study is observational, single-site, and focused on a single course; unmeasured factors (e.g. prior knowledge, motivation) may influence both engagement and outcomes. Grades are discrete and pass/fail is coarse; engagement is operationalized through the ten FC assessments and may not capture all study behaviors. Survey data are self-reported and subject to recall and desirability biases. Results speak to associations in this context, rather than causal effects.

Future work. To move from association to credible causal claims about whether flipped learning outperforms traditional formats, the next phase should be a multi-institution, multi-course study-especially in non-programming CS subjects-with sufficient power for heterogeneity analysis. Designs could include cluster randomization or stepped-wedge rollouts at the section or instructor level, complemented by quasi-experimental strategies (propensity scoring, difference-in-differences, synthetic controls) where randomization is infeasible. A pre-registered causal analysis plan should integrate Bradford Hill considerations-temporality (engagement precedes outcomes), consistency (replication across groups and sites), dose-response (greater sustained engagement aligns with stronger effects), plausibility and coherence (linking mechanisms to learning theory), experiment (randomized or natural experiments), specificity (outcomes most affected are those targeted by FC), and analogy. Richer telemetry (fine-grained video interaction, forum discourse, attendance, project checkpoints) and background covariates (prior GPA, programming experience, workload) would help address confounding and support sensitivity analyses. Parallel work should assess calibration, fairness, and cost-effectiveness of early-warning models and examine which FC design elements (video length, feedback timing, limited retakes) drive the largest marginal gains.

If such evidence shows robust, generalizable benefits, institutions can act at scale: embed FC as a program-level design (rather than course-by-course), fund content production and analytics infrastructure, provide faculty development on learning design and data use, and formalize early-support protocols triggered by engagement signals. If effects are context-dependent, policy can target FC to courses and student segments where it is most effective, adjust assessment weightings (e.g., small stakes, frequent checks), and iterate on specific design levers (chunking, feedback, team tasks) that show causal lift. Either way, a larger, causally informed evidence base would enable a defensible curriculum strategy rather than ad-hoc adoption.

5. Conclusion

This study set out to clarify what flipped learning means for computer science education beyond programming courses and to share practical lessons from introducing it in an IT Service Management course. The literature points to clear benefits when pre-class preparation is paired with active class time; our case study shows that the same pattern is workable and well-received in ITSM. Students generally valued the videos, pre-class quizzes, and team tasks, and they described the format as helpful for following lectures and preparing for assessments.

The broad takeaway for teaching practice is simple: steady, quality-focused engagement on frequent assessments across the semester tracks with better achievement. Teachers can act on this by (1) setting clear weekly expectations, (2) using short, well-scaffolded videos, (3) keeping pre-class quizzes low-stakes but predictable, and (4) using LMS analytics to spot and support students who fall behind early. Students consistently asked for a bit more time on quizzes, clearer feedback, occasional native-language materials, and-in moderation-retakes; these are concrete, low-cost adjustments that align with the spirit of formative assessment. For course teams, the design workload is front-loaded: planning, producing, and curating materials matters as much as what happens in the classroom.

At the institutional level, scaling flipped learning responsibly means supporting content production, offering staff development on learning design and analytics, and setting light-touch policies for continuous assessment and early support. Data governance and simple dashboards that highlight weekly engagement can make interventions timely without being intrusive.

Our findings do not claim causality, but they do offer practical guidance: when flipped learning is thoughtfully designed and consistently supported, students engage more steadily-and that is the kind of behavior teachers and institutions can nurture.

References

Aldalur, I., Markiegi, U., Iturbe, M., Roman, I., & Illarramendi, M. (2024). An experience in the implementation of the flipped classroom instructional model in the computer science degree. Engineering Reports, 6(4). Scopus. https://doi.org/10.1002/eng2.12754

Aldalur, I., Markiegi, U., Valencia, X., Cuenca, J., & Illarramendi, M. (2022). E-Learning Experience with Flipped Classroom Quizzes Using Kahoot, Moodle and Google Forms: A Comparative Study. ACM Int. Conf. Proc. Ser., 82-89. Scopus. https://doi.org/10.1145/3572549.3572563

Algayres, M., Triantafyllou, E., Werthmann, L., Zotou, M., Efthimios, T., Malliarakis, C., Dermentzi, E., Lopez, R., Jatten, E., & Tarabanis, K. (2021). Collaborative Game Design for Learning: The Challenges of Adaptive Game-Based Learning for the Flipped Classroom. In Brooks A., Brooks E.I., & Jonathan D. (Eds.), Lect. Notes Inst. Comput. Sci. Soc. Informatics Telecommun. Eng.: Vol. 367 LNICST (pp. 228-242). Springer Science and Business Media Deutschland GmbH; Scopus. https://doi.org/10.1007/978-3-030-73426-8_13

Araujo, P., Costa, C., Viana, W., De Lima Veras, N., & Farias, E. J. P. (2020). Automatic Personalisation of Study Guides in Flipped Classroom: A Case Study in a Distributed Systems Course. Proc. Front. Educ. Conf. FIE, 2020-October. Scopus. https://doi.org/10.1109/FIE44824.2020.9274186

Bakheet, E. M., & Gravell, A. M. (2021a). Investigating computer science instructors behavioral intention to adopt the flipped classroom applying an extended utaut model: The role of age, gender, and experience. International Journal of Information and Education Technology, 11(12), 631-637. Scopus. https://doi.org/10.18178/IJIET.2021.11.12.1574

Bakheet, E. M., & Gravell, A. M. (2021b). Significant factors influencing computer science instructor's behavioral intentions to adopt the flipped classroom: A global quantitative study. Int. Conf. Educ. Inf. Technol., ICEIT, 150-155. Scopus. https://doi.org/10.1109/ICEIT51700.2021.9375590

Bakheet, E. M., & Gravell, A. M. (2021c). Would Flipped Classroom be My Approach in Teaching Computing Courses: Literature Review. Int. Conf. Inf. Educ. Technol., ICIET, 166-170. Scopus. https://doi.org/10.1109/ICIET51873.2021.9419631

Divjak, B., Rienties, B., Iniesto, F., Vondra, P., & Žižak, M. (2022). Flipped classrooms in higher education during the COVID-19 pandemic: Findings and future research recommendations. International Journal of Educational Technology in Higher Education, 19(1), 9. https://doi.org/10.1186/s41239-021-00316-4

Elgrably, I. S., & Ronaldo Bezerra Oliveira, S. (2022). Using flipped classroom to promote active learning and engagement in a Software Testing subject remotely during the COVID-19 pandemic. Proc. Front. Educ. Conf. FIE, 2022-October. Scopus. https://doi.org/10.1109/FIE56618.2022.9962379

Gallaugher, J. (2023). Physical Computing: Empowering Students with Hardware Programming. Annu. Am. Conf. Inf. Syst., AMCIS. 29th Annual Americas Conference on Information Systems, AMCIS 2023. Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85192898139&partnerID=40&md5=90f6d030fbd09d305eed6448bcaee293

Gong, J., Cai, S., & Cheng, M. (2023). Exploring the Effectiveness of Flipped Classroom on STEM Student Achievement: A Meta-analysis. Technology, Knowledge and Learning. Scopus. https://doi.org/10.1007/s10758-023-09700-7

Hannaoui, M., Janous, Y. E., El-Hassouny, E. H., El Hachhach, J., Askam, A., & Habibi, I. (2023). The Impact of the Online Flipped Classroom on the Learning Outcomes and Motivation of Nursing Students in Computer Science. In Lazaar M., En-Naimi E.M., Zouhair A., Al Achhab M., & Mahboub O. (Eds.), Lect. Notes Networks Syst.: Vol. 625 LNNS (pp. 138-147). Springer Science and Business Media Deutschland GmbH; Scopus. https://doi.org/10.1007/978-3-031-28387-1_13

J. Urquiza-Fuentes. (2023). A Study About the Impact of Rewards on Student's Engagement with the Flipped Classroom Methodology. 2023 IEEE Global Engineering Education Conference (EDUCON), 1-6. https://doi.org/10.1109/EDUCON54358.2023.10125267

Long, T., Cummins, J., & Waugh, M. (2017). Use of the flipped classroom instructional model in higher education: Teachers' perspectives. Journal of Computing in Higher Education, 29(2), 179-200. https://doi.org/10.1007/s12528-016-9119-8

Mithun, S., & Luo, X. (2020). Design and Evaluate the Factors for Flipped Classrooms for Data Management Courses. Proc. Front. Educ. Conf. FIE, 2020-October. Scopus. https://doi.org/10.1109/FIE44824.2020.9274201

Nouri, J. (2016). The flipped classroom: For active, effective and increased learning - especially for low achievers. International Journal of Educational Technology in Higher Education, 13(1), 33. https://doi.org/10.1186/s41239-016-0032-z

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

Olivindo, M., Veras, N., Viana, W., Cortés, M., & Rocha, L. (2021). Gamifying Flipped Classes: An Experience Report in Software Engineering Remote Teaching. ACM Int. Conf. Proc. Ser., 143-152. Scopus. https://doi.org/10.1145/3474624.3476971

Ossovski, E. (2022). Digitally Supported Introductory University Teaching in Computer Science Considering Heterogeneous Groups. ICER - Proc. ACM Conf. Int. Comput. Educ. Res., 2, 10-11. Scopus. https://doi.org/10.1145/3501709.3544296

Sibia, N., & Liut, M. (2022). The Positive Effects of using Reflective Prompts in a Database Course. In Aivaloglou E., Fletcher G., & Miedema D. (Eds.), Proc. ACM SIGMOD Int. Workshop Data Syst. Educ.: Bridging Educ. Pract. Educ. Res., DataEd (pp. 32-37). Association for Computing Machinery, Inc; Scopus. https://doi.org/10.1145/3531072.3535323

Sosa Díaz, M. J., Guerra Antequera, J., & Cerezo Pizarro, M. (2021). Flipped Classroom in the Context of Higher Education: Learning, Satisfaction and Interaction. Education Sciences, 11(8), 416. https://doi.org/10.3390/educsci11080416

Steinmaurer, A., & Gütl, C. (2023). Implementation and Experiences of a Flipped Lecture Hall-A Fully Online Introductory Programming Course. In Auer M.E., Pachatz W., & Rüütmann T. (Eds.), Lect. Notes Networks Syst.: Vol. 633 LNNS (pp. 832-843). Springer Science and Business Media Deutschland GmbH; Scopus. https://doi.org/10.1007/978-3-031-26876-2_79

Yu, Y., Fu, Y., Chen, Y., & Li, M. (2023). MOOCs Based Blending Teaching Reform for Integrated College Computer Course. In Hong W. & Weng Y. (Eds.), Commun. Comput. Info. Sci.: Vol. 1811 CCIS (pp. 293-304). Springer Science and Business Media Deutschland GmbH; Scopus. https://doi.org/10.1007/978-981-99-2443-1_26

Zamora-Hernandez, I., Rodriguez-Paz, M. X., Gonzalez-Mendivil, J. A., & Palomares-Moctezuma, J. A. (2022). A Teaching Model for a Five-week Electric Circuits Course for Engineering Students. ACM Int. Conf. Proc. Ser., 76-81. Scopus. https://doi.org/10.1145/3572549.3572562