Jupiter: Jurnal Pendidikan Terapan
Homepage : https://journal.diginus.id/jupiter/index
Email: jurnalpendidikanterapan@gmail.com

p-ISSN : 2985-3214; e-ISSN : 2964-3171
Jupiter, Vol. 4, No. 3, July 2026
Page 392-406

CLIL-Based Learning Tools Mediated by Asynchronous Video for Multicultural Education in Bilingual Elementary Teacher Preparation

Novi Setyasto[1*], Zhao Xinzhe[2]
[1]Universitas Negeri Semarang, Semarang, Indonesia
[2]Liuzhou City Vocational College, Liuzhou, China

ARTICLE INFO

Article history:
Submitted: July 03, 2026
Final Revised: July 21, 2026
Accepted: July 24, 2026
Published: July 31, 2026

Keywords:
CLIL; Asynchronous Video; Multicultural Education; Bilingual Teacher Education.

ABSTRACT

Purpose: This study developed, validated, and evaluated CLIL-based learning tools (a semester learning plan, eight thematic teaching modules, and assessment instruments) mediated by asynchronous video for a multicultural education course to address the gap between bilingual policy and instructional practice in Indonesian elementary teacher preparation. **Methods:** A mixed-methods design combined the Four-D development model with a 16-week quasi-experimental pretest–posttest control-group study in which two intact classes totalling 74 pre-service teachers (37 per group) were assigned to conditions without randomization. Data from expert validation questionnaires, a validated cognitive test, structured observations, response questionnaires, and interviews were analysed using ANCOVA, normalized gain, and thematic analysis. **Findings**: All components of the tools met the expert-validation criterion in the highest ("very valid") category, with high inter-validator agreement. The experimental group achieved significantly higher covariate-adjusted post-test scores than the control group (adjusted mean difference = 12.14, 95% CI [7.23, 17.04], partial η^2 = .26), together with a higher class-average normalized gain (g = 0.50 vs. 0.29) and, descriptively, higher mastery attainment (76% vs. 54%). Observed engagement in the experimental group rose across the semester and most students (86.5%) responded positively; lower baseline English proficiency was descriptively associated with non-mastery. **Research Implications**: The validated tools offer an adaptable template for bilingual teacher education, provided implementation includes diagnostic assessment, differentiated scaffolding for lower-proficiency learners, and reliable technical support. **Originality**: This study provides rare controlled quasi-experimental evidence from Indonesian teacher education on integrating CLIL, asynchronous video-mediated communication, and multicultural content as one coherent instructional package.

Doi: https://doi.org/10.61255/jupiter.v4i3.1537

INTRODUCTION

The globalization of education demands graduates who can integrate subject-matter expertise with proficiency in more than one language. In response, a growing number of Indonesian higher education institutions have instituted bilingual programmes designed to enhance the global competitiveness of their graduates (Cummins, 1992; Garcia & Wei, 2014; Wright & Baker, 2017). Universitas Negeri Semarang (UNNES), a conservation-oriented and internationally oriented university, implemented bilingual classes in its Primary School Teacher Education (PGSD) program to prepare future teachers for culturally and linguistically diverse classrooms.

The official policy, however, does not always match the reality of classroom instruction. A preliminary needs assessment conducted by the researchers in the PGSD bilingual program during the 2022/2023 academic year documented a substantial policy-practice gap. The assessment comprised analysis of course syllabi and lesson plans, structured classroom observations, and interviews with six of the 20 lecturers teaching the bilingual classes and eight students (see the Define phase in the Method section). Most course documents were still written in Indonesian, and only about one in four lecturers demonstrated consistent bilingual delivery. Classroom observations showed that

English was largely confined to opening and closing routines, while core content was delivered predominantly in Indonesian. The instructional materials for the multicultural education course, meanwhile, were available only in English, without scaffolded bilingual support. These conditions created a mismatch between the program's stated goals and its instructional practice, and they were accompanied by students' self-reported anxiety about speaking English, a factor known to depress language learning outcomes (Horwitz, 2001). Filling the gap demands a systematic and integrated approach to content and language, and it also requires helping learners overcome their fear and anxiety about using English.

Content and Language Integrated Learning (CLIL) offers a theoretically grounded response. CLIL is a dual-focused approach in which an additional language serves as the medium for teaching subject content and developing language proficiency simultaneously (Coyle et al., 2010; Hidalgo & Ortega-Sánchez, 2023). Its conceptual core is the 4Cs Framework (Content, Communication, Cognition, and Culture). It provides an integrated structure for curriculum and lesson design (Coyle, 2008; Coyle et al., 2010). A growing evidence base indicates that CLIL can enhance receptive and productive language skills, build learner confidence, and foster intercultural understanding (Dalton-Puffer, 2011; Goris et al., 2019). A recent multilevel meta-analysis of 38 studies found a medium-to-large advantage of CLIL instruction for English development relative to mainstream instruction ($d = 0.73$ in the short term) (Ho et al., 2023). At the same time, scholars caution that benefits are not automatic. Substantial bilingual proficiency depends on sustained exposure and systematic implementation (Van Mensel et al., 2020). Apparent CLIL advantages in observational studies may be inflated by selection effects (Bruton, 2011). In Indonesian higher education specifically, rigorous evidence remains limited (Khoiriyah, 2021; Norhasanah & Setiawan, 2023). Recent work in this sector remains largely descriptive or evaluative: a curriculum review at a medical faculty identified lecturers' language and content competencies, differentiated instruction, and institutional support as preconditions for sustainable CLIL implementation (Sujana et al., 2023); a mixed-methods study of online medical-CLIL classes reported high student satisfaction driven by the design of learning materials, flexibility, and lecturer support (Khoiriyah et al., 2025); and, at the postgraduate level, Talaue & Hendijani (2025) proposed a holistic CLIL-based support framework for business students whose academic writing is impeded by anxiety, underscoring the salience of affective barriers to content–language integration in Indonesian universities.

As affective barriers are the focal point of the local problem, the present intervention integrates CLIL with Flipgrid, an asynchronous video-discussion platform that enables students to record, upload, and respond to short videos without the pressure of synchronous performance. Asynchronous video discussion has been associated with lower speaking anxiety, greater willingness to communicate, and a stronger social presence in language classrooms (Bartlett, 2018; Green & Green, 2018; Stoszkowski, 2018). Research shows that such platforms offer genuine speaking and listening practice and a psychologically safer environment for oral production, since learners can rehearse and re-record before submission (Lowenthal & Moore, 2020; Mango, 2021; Tuyet & Khang, 2020). Video-response features have also been associated with social learning and community building (Bartlett, 2018; Stoszkowski, 2018). These benefits are tempered, however, by technological obstacles and uneven digital literacy (Miskam et al., 2019).

Multicultural education is a suitable site for such content–language integration. It encompasses teaching practices that incorporate the experiences and perspectives of diverse cultural backgrounds (Nieto & Bode, 2012), and Banks (1993) delineated its five dimensions: content integration, knowledge construction, prejudice reduction, equity pedagogy, and an empowering school culture. Because language and culture are inextricably linked (Dimici & Başbay, 2021), integrating multicultural and intercultural content throughout English-medium teaching has been shown to reinforce both language and cultural competence (Dimici & Başbay, 2021; Wei, 2018), and comparable benefits have been documented in Indonesian higher education (Nurbatra, 2022).

While CLIL, asynchronous video technology, and multicultural education are each well established in the literature, little research has examined how the three domains intersect as a unit in Indonesian teacher education. Most studies to date treat CLIL or technology integration in isolation, and their potentially synergistic functioning is seldom examined. Research on CLIL in Indonesia has grown rapidly but remains concentrated at the primary and secondary school levels and dominated by descriptive or case-study designs (Khoiriyah, 2021; Kholili et al., 2024; Norhasanah & Setiawan, 2023; Setyaningrum et al., 2022); the few experimental studies available, such as Idris et al. (2025) in vocational secondary education, do not address teacher preparation, so designs that restrict causal inference still prevail in this sector. Teacher-focused Indonesian studies show the same pattern: recent investigations have documented in-service teachers' beliefs and practices in a CLIL primary school (Kurniawati et al., 2025) and teachers' professional learning through action research on designing CLIL-based reading activities (Widiati et al., 2026), yet both rely on qualitative, non-controlled designs and neither addresses pre-service teacher preparation. The development and

validation of comprehensive learning tools that embed CLIL principles, an asynchronous video platform, and multicultural content therefore remains underexplored.

This study addresses that gap by developing, validating, and evaluating CLIL-based learning tools supported by Flipgrid for a multicultural education course, using a quasi-experimental design with a control group. It contributes (a) an integrated CLIL–video instructional model designed for Indonesian bilingual teacher education, (b) controlled quasi-experimental evidence from a comparison with conventional instruction, and (c) validated, adaptable instruments. The conceptual premise is that combining structured content–language integration (CLIL) with technology-mediated anxiety reduction (asynchronous video) may yield complementary benefits for content mastery and language development (Figure 1); the present design evaluates the combined package rather than the separate contribution of each component. Three research questions guide the study: (1) What is the validity of the CLIL-based learning tools supported by Flipgrid? (2) How effective are the tools in improving learning outcomes relative to conventional instruction? (3) How practical are the tools as perceived by students?.

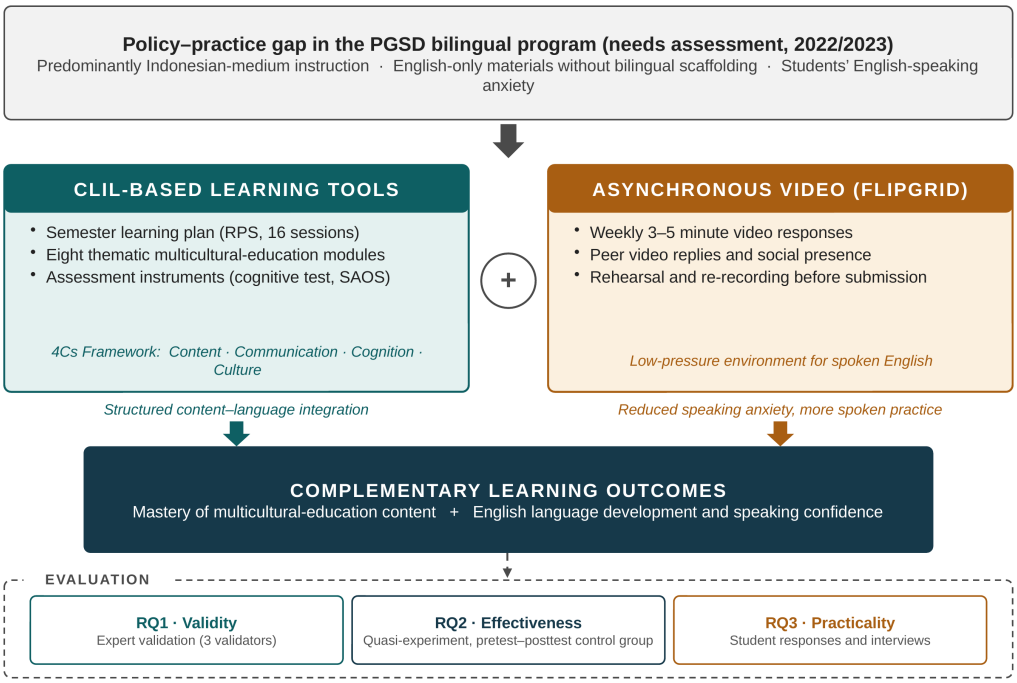

Figure 1. Conceptual framework of the CLIL-based learning tools mediated by asynchronous video

METHOD

Research design

This study used a mixed-methods design that combined research and development (R&D) with a quasi-experimental evaluation. The R&D phase followed the Four-D model (Thiagarajan et al., 1974), with each phase producing specified outputs. In the *Define* phase, a needs assessment was conducted through analysis of course syllabi and lesson plans, structured classroom observations, and interviews with six lecturers, purposively selected from the 20 who taught in the bilingual program, and eight students. Its outputs were a specification of the policy–practice gap, an analysis of learner characteristics and entry proficiency, and integrated content–language objectives for the multicultural education course. In the *Design* phase, these specifications were translated into prototypes: the blueprint of a 16-session semester learning plan (RPS) mapping the 4Cs onto each session, the template and draft content of eight thematic teaching modules, the blueprints of the cognitive test and the observation sheet, and the selection of Flipgrid as the asynchronous video platform. In the *Develop* phase, the prototypes underwent iterative expert validation by three validators, revision, pilot testing with 12 students from a non-participating cohort, and final refinement. In the original Four-D model the fourth phase is *Disseminate*; in this study that phase was deliberately modified into a limited implementation-and-evaluation stage within the target program, because large-scale dissemination lay beyond the study's scope and resources. The evaluation employed a quasi-experimental pretest–posttest control-group design over

16 weeks (one semester), chosen to strengthen causal inference relative to the pre-experimental designs common in comparable studies (Shadish et al., 2002). Because institutional scheduling did not permit randomization, intact class sections (one per condition) were assigned to the experimental and control conditions; baseline comparability was examined statistically, and the resulting class-level (cluster) assignment is acknowledged as a limitation on causal inference.

Participants

An a priori power analysis using G*Power 3.1 (Faul et al., 2007) for an independent-samples comparison indicated that 72 participants (36 per group) would provide 80% power to detect a medium-to-large effect ($d = 0.67$) at $\alpha = .05$ (two-tailed). The final sample of 74 students (37 per group) therefore slightly exceeded the minimum required. The final sample comprised: 53 women and 21 men, aged 19–22 years (M = 20.4, SD = 0.89). Baseline English proficiency was predominantly intermediate (B1–B2 on the CEFR), with a minority at level A2 (nine of 37 students in the experimental group); pre-test scores did not differ significantly between groups, $t(72) = 0.31$, $p = .76$ (standardized mean difference = 0.07). The learning tools were evaluated by three validators, selected purposively on the basis of (a) a doctorate in a field matching a component of the tools (English language education, elementary education, and teacher preparation, respectively); (b) experience in instructional design or instrument validation; and (c) no other involvement in the study.

Intervention conditions

Experimental group. Over 16 weekly 100-minute sessions, the experimental group received CLIL-based instruction with asynchronous video support. Each session was structured around the 4Cs: content was presented primarily in English with Indonesian scaffolding as needed (approximately 70% English / 30% Indonesian); communication occurred through structured discussion and weekly Flipgrid video responses (3–5 minutes) with peer video replies; cognition tasks required higher-order thinking (analysis, evaluation, creation); and culture was addressed by connecting multicultural themes to local Indonesian contexts.

Control group. Over the same period and contact time, the control group received conventional bilingual instruction using existing Indonesian-language materials with English supplements. Instruction was primarily in Indonesian (approximately 80%). English was used mainly for key terminology and brief explanations. Materials carried English glossaries and assignments were in Indonesian. No video platform was used and discussion occurred in traditional in-class formats.

Instructor. The same instructor, who had eight years of experience in bilingual education and appropriate certification, taught both conditions. In order to minimize experimenter bias, the instructor adhered to an identical lesson-plan structure in both conditions. An independent rater blind to group assignment scored all assessments. Inter-rater reliability between the instructor and the blind rater was high (ICC = .82). The implications of this single-instructor arrangement are considered further in the Limitations.

Learning tools developed

The tools developed include: a) a Semester Learning Plan (RPS) incorporating CLIL principles in 16 sessions with traceable 4 C's; b) learning materials arranged in eight thematic modules in diversity themes, cultural identity, prejudice reduction, equity pedagogy, and culturally responsive teaching; and c) assessment instruments, with a cognitive test aligned with the revised Bloom's taxonomy, as well as a Student Activity Observation Sheet (SAOS) with eight indicators derived from CLIL. The communication component predominantly took place on the asynchronous video platform, through weekly video responses and peer interactions.

Instruments and validation

The cognitive test was developed and validated in several steps. Content validity was established through review by the three validators, who rated the alignment of each item with the learning objectives using a structured rubric. Construct validity was examined through a separate pilot administration to 50 students from a different cohort. Exploratory factor analysis and item analysis were then conducted: five items with low discrimination (< .20) were removed from an initial pool of 35, and the retained 30 items loaded on three interpretable factors content knowledge (12 items), language application (10 items), and cultural understanding (8 items) accounting for 62.4% of the variance, with a final reliability of Cronbach's $\alpha = .84$. Discriminant validity was supported by a low correlation between the cognitive test and a general English-proficiency measure ($r = .31$), indicating that the test captured content-specific learning rather than general language ability. The SAOS was validated through expert review and pilot testing; two observers trained on the rubric independently coded 10 video-recorded sessions, achieving Cohen's $\kappa = .81$, and any disagreements during data collection were resolved by discussion to consensus.

Data collection

Multiple instruments were used to support triangulation. Validity data came from expert-validation questionnaires. Effectiveness data comprised (a) pre- and post-tests (30 items; Cronbach's $\alpha = .84$) administered to both groups at meetings 1 and 16, and (b) systematic observation of the experimental group at meetings 2, 9, and 15 using the SAOS ($\kappa = .81$). Practicality data came from a student response questionnaire (14 items; Cronbach's $\alpha = .87$); the questionnaire also included items on technical difficulties encountered with the video platform, and these questionnaire reports were the source of the technical-difficulty counts analysed in the Results. Semi-structured interviews with 12 students (eight experimental, four control), selected purposively to represent high, medium, and low performers, provided supplementary qualitative data.

Data analysis

Quantitative analysis combined descriptive and inferential statistics. Validity data were analysed by summing each validator's item ratings for each component; the resulting validator totals and their mean were then converted into four ordered categories (from "not valid" to "very valid") using the ideal-score interval conversion conventional in Indonesian educational R&D, based on each instrument's ideal mean, Mi = (maximum + minimum)/2, and ideal standard deviation, SDi = (maximum-minimum)/6, applied to that instrument's score range. A four-category conversion was adopted so that the number of interpretive categories matches the four-point response scale, which offers no neutral midpoint; under this conversion, "very valid" corresponds to $X > Mi + 1.5\ SDi$, "valid" to $Mi < X \leq Mi + 1.5\ SDi$, "less valid" to $Mi - 1.5\ SDi < X \leq Mi$, and "not valid" to $X \leq Mi - 1.5\ SDi$. The criterion score (Fs) reported in Table 1 marks the minimum of the highest ("very valid") category for each instrument. Inter-validator agreement was quantified with the intraclass correlation coefficient (ICC; two-way random-effects, absolute-agreement, average-measures model), computed on component totals rescaled to the proportion of each instrument's maximum so that the components share a common metric, and reported with its 95% confidence interval. Baseline comparability was examined with an independent-samples *t* test and described with the standardized mean difference (Cohen's d), because a non-significant test does not by itself establish equivalence. Effectiveness was assessed using (a) analysis of covariance (ANCOVA) with pre-test scores as the covariate to adjust for baseline differences and increase precision (Field, 2018); (b) Hake (1998) normalized gain, g = (post − pre) / (maximum-pre), interpreted as high ($g \geq 0.70$), medium ($0.30 \leq g < 0.70$), or low ($g < 0.30$); and (c) partial eta-squared as the ANCOVA effect size, reported with its 90% confidence interval. The tenability of ANCOVA assumptions was evaluated through inspection of pre–post scatterplots (linearity), Levene's test (homogeneity of variance), the group × pre-test interaction term (homogeneity of regression slopes), and the Shapiro-Wilk test on model residuals (normality); the results of these checks are reported alongside the main analysis, and where a check indicated a violation, a sensitivity ANCOVA on rank-transformed scores (Conover & Iman, 1981) was conducted to verify the robustness of the conclusion. Normalized gain was computed from group and subgroup mean scores using the class-average formula with the test maximum of 100, g = (Mpost-Mpre)/(100-Mpre). Mastery attainment, the proportion of students reaching the minimum mastery criterion (KKM ≥ 70), termed "classical completeness" in Indonesian criterion-referenced practice, was benchmarked at 75% of the class; group proportions were compared with a two-proportion z test, a Newcombe (1998) hybrid score 95% confidence interval for the difference in proportions (chosen for consistency with the score-based test), and a risk ratio with its 95% confidence interval. Exploratory profiling of students who did not attain mastery was conducted post hoc and analysed descriptively only, without inferential testing. Qualitative data were analysed thematically following Braun & Clarke (2006) six phases: familiarization, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the report; the coding process and the resulting themes are reported in the Results.

Ethical considerations

Research involving human beings was conducted in accordance with established ethical standards. Permission to conduct the research in the PGSD bilingual classes was granted by the study program. All participants received written information about the purpose and procedures of the study and gave informed consent. Participation was voluntary, declining had no academic consequences, and all data were anonymized before analysis. Because the control condition consisted of the standard curriculum, no student was deprived of regular instruction, and the validated learning materials were made available to the control group after the post-test.

RESULTS

This section reports the findings for the three research questions (validity, effectiveness, and practicality) in the order of the research objectives. Their interpretation, together with the study's contributions and limitations, is presented separately in the Discussion.

Validity of the learning tools

Three validators rated the tools on content accuracy, pedagogical appropriateness, integration of CLIL principles, language quality, and alignment with objectives. As Table 1 shows, every component exceeded the minimum score of the highest interval, so all four components (the RPS, the learning materials, the cognitive test, and the observation sheet) were classified "very valid"; the learning materials mean (20.3) exceeded the "very valid" boundary (> 19.5), as did each individual validator's total for every component. Inter-validator agreement, computed on component totals rescaled to the proportion of each instrument's maximum so that the four components share a common metric, was high (ICC = .86, 95% CI [.30, .99]; two-way random-effects, absolute-agreement, average-measures model), the wide interval reflecting the small number of rated components; in absolute terms, the validators' totals differed by no more than one point on any component. The RPS received a high mean rating (M = 43.7 of a possible 48). Initial disagreements among validators concerned the sufficiency of scaffolding for lower-proficiency students and the clarity of some rubric criteria; these were resolved through discussion to consensus against the rubric standards, and the materials were revised accordingly before piloting.

Table 1. Expert validation results

Component	Criterion	V1	V2	V3	Mean
Semester Learning Plan (RPS)	Fs > 39	44	43	44	43.7
Learning materials	Fs > 19.5	21	20	20	20.3
Cognitive test	Fs > 13	15	14	15	14.7
Activity observation sheet	Fs > 19.5	22	22	22	22.0

Note. V1 = English language education; V2 = elementary education; V3 = teacher preparation. Scores are totals of each validator's ratings per component. Fs = minimum criterion score of the highest ("very valid") category under the four-category ideal-score interval conversion defined in the Data analysis section (X > Mi + 1.5 SDi), applied to each instrument's score range. All four components exceeded this criterion and were classified "very valid". Validation-sheet lengths and possible total-score ranges: RPS, 12 items, range 12-48; learning materials, 6 items, range 6-24; cognitive test, 4 items, range 4-16; activity observation sheet, 6 items, range 6-24. Item counts refer to the expert-validation rating sheets, not to the instruments they assess (the cognitive test itself comprises 30 items).

Effectiveness: cognitive learning outcomes

Pre-test scores were closely similar across the groups, experimental M = 42.0 (SD = 10.8) versus control M = 41.2 (SD = 11.4), $t(72) = 0.31$, $p = .76$, Cohen's $d = 0.07$. Because a non-significant test does not by itself establish equivalence, the negligible standardized mean difference is the more informative indicator of baseline comparability, and pre-test scores were additionally controlled as a covariate in the primary analysis. Post-test results (Table 2; Figure 2) favoured the experimental group, and ANCOVA controlling for the pre-test (Table 3) showed a significant group difference, $F(1, 71) = 24.36$, $p < .001$, partial $\eta^2 = .26$, 90% CI [.12, .38]. Covariate-adjusted means were 70.8 (SE = 1.7) for the experimental group and 58.6 (SE = 1.7) for the control group, an adjusted mean difference of 12.14 points, 95% CI [7.23, 17.04]. The covariate (pre-test) was significantly related to post-test scores, $F(1, 71) = 22.40$, $p < .001$, and the full model accounted for $R^2 = .41$ (adjusted $R^2 = .39$) of post-test variance (error MS = 111.76). Because mastery-related clustering left control-group scores more dispersed, Levene's test indicated greater variance in the control group, $F(1, 72) = 6.12$, $p = .016$; the group × pre-test interaction was non-significant, $F(1, 70) = 0.14$, $p = .71$, supporting homogeneity of regression slopes; and model residuals were approximately normal, Shapiro–Wilk $W = .98$, $p = .40$. With equal group sizes, the F test is robust to variance heterogeneity of this form (Field, 2018), and a sensitivity ANCOVA on rank-transformed scores reproduced the group effect, $F(1, 71) = 17.94$, $p < .001$ (Conover & Iman, 1981). The experimental group achieved a medium class-average normalized gain (g = 0.50) versus a low gain for controls (g = 0.29). Mastery attainment was 75.7% (28 of 37 students) in the experimental group, exceeding the 75% benchmark by the equivalent of a single student, compared with 54.1% (20 of 37) in the control group; the difference of 21.6 percentage points, 95% CI [−0.1, 40.7] (Newcombe, 1998, hybrid score method), did not reach conventional significance, z = 1.95, $p = .051$; risk ratio = 1.40, 95% CI [0.99, 1.98]. Consistent with the marginal test, both interval estimates for this comparison include the null value. Given this and the cut-off dependence of the dichotomized

indicator, the continuous outcomes (the adjusted mean difference and the normalized gain) constitute the primary evidence of effectiveness.

Table 2. Cognitive learning outcomes for the experimental and control groups

Measure	Exp. pre	Exp. post	Ctrl. pre	Ctrl. post
Minimum score	25	52	23	40
Maximum score	63	89	65	78
Mean (M)	42.0	71.0	41.2	58.4
SD	10.8	9.7	11.4	14.0
Students reaching KKM	0 (0%)	28 (76%)	0 (0%)	20 (54%)
N-gain (g)	-	0.50 (medium)	-	0.29 (low)

∫ ANCOVA results and mastery-attainment comparison (post-test; pre-test as covariate)

Source / statistic	Value	p	Interpretation
Group effect (ANCOVA, covariate: pre-test)	$F(1, 71) = 24.36$	< .001	Significant
Covariate (pre-test)	$F(1, 71) = 22.40$	< .001	Covariate effect on post-test
Error term and model fit	MS error = 111.76	—	$R^2 = .41$; adjusted $R^2 = .39$
Adjusted mean, experimental group	70.8 (SE = 1.7)	—	Adjusted to grand pre-test mean
Adjusted mean, control group	58.6 (SE = 1.7)	—	Adjusted to grand pre-test mean
Partial η^2	0.26	—	Large effect; 90% CI [.12, .38]
Adjusted mean difference	12.14	—	95% CI [7.23, 17.04]
Mastery attainment (KKM ≥ 70)	75.7% vs. 54.1%	.051	Δ = 21.6 pp, 95% CI [−0.1, 40.7]; RR = 1.40, 95% CI [0.99, 1.98]

Note. Adjusted means are evaluated at the grand pre-test mean (41.6); standard errors are model-based. The mastery-attainment comparison uses a two-proportion z test; Δ = difference in proportions in percentage points (pp); RR = risk ratio. The 95% CI for Δ uses the Newcombe hybrid score method, chosen for consistency with the score-based z test. η^2 CI at 90% as conventional for F-based effect sizes.

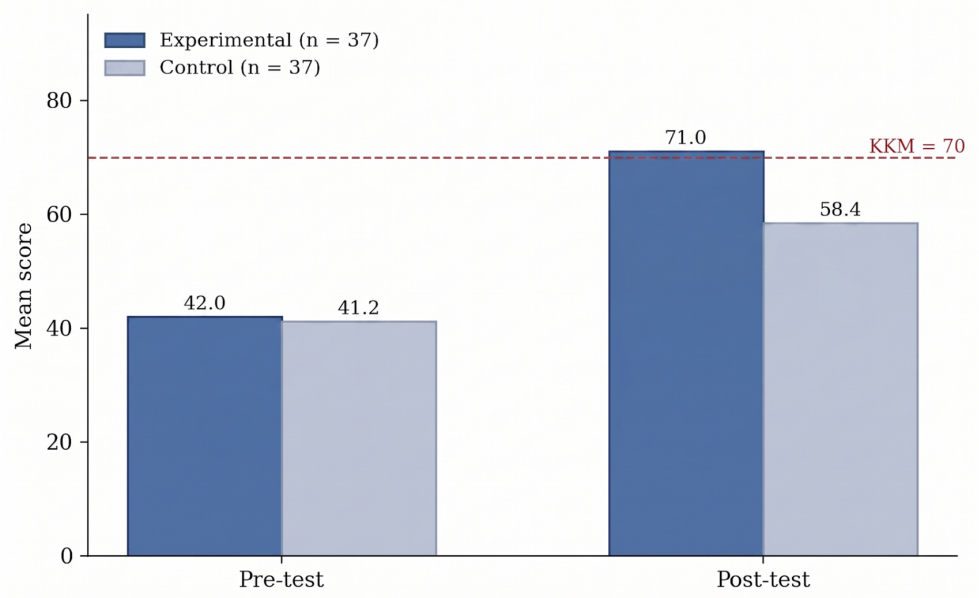

Figure 2. Pre-test and post-test mean scores of the experimental and control groups (dashed line marks the mastery criterion, KKM = 70)

Exploratory profile of students not attaining mastery (post hoc)

Nine students in the experimental group (24%) did not reach the mastery criterion (score < 70). Table 4 profiles this subgroup; because the comparison was defined after outcomes were inspected, it is post hoc and exploratory, is reported descriptively without inferential tests, and does not identify predictors or moderators of mastery. Seven of the nine (78%) had entered at CEFR level A2 and two at B1–B2, whereas 26 of the 28 mastery students (93%) had entered at B1–B2 and two at A2. The non-mastery subgroup's pre-test mean (31.2) was 14.3 points below that of the mastery students (45.5). These students nonetheless made substantial progress (subgroup class-average $g = 0.40$), with a mean absolute gain of 27.5 points, larger than the control group's mean gain of 17.2 points. On the end-of-course response questionnaire, six of the nine (67%) reported technical difficulties with uploading or platform access, versus four of the 28 mastery students (14%); interviews with a subset of these students elaborated on the reports, describing persistent difficulty with academic vocabulary, longer processing time for English-medium content, and increased cognitive load during video production. Because lower entry proficiency and technical difficulty co-occurred within the same students, their separate contributions to non-mastery cannot be distinguished.

Table 4. Profile of non-mastery and mastery students (experimental group)

Characteristic	Non-mastery (n = 9)	Mastery (n = 28)
Baseline English proficiency	A2: 7; B1–B2: 2	B1–B2: 26; A2: 2
Pre-test M (SD)	31.2 (5.4)	45.5 (9.8)
Post-test M (SD)	58.7 (4.8)	74.9 (7.2)
N-gain (g)	0.40 (medium)	0.54 (medium)
Video completion rate	69% (11/16)	94% (15/16)
Reported technical difficulties (questionnaire)	6 of 9 (67%)	4 of 28 (14%)

Note. Post hoc, exploratory, and descriptive comparison; no inferential tests were conducted. Technical-difficulty counts derive from the end-of-course response questionnaire administered to all experimental-group students; interviews with a subset elaborated these reports. N-gain uses the class-average formula applied to subgroup means. Video completion rates are means of weekly submissions out of 16.

Student engagement and activity progression

Structured observation with the SAOS was conducted in the experimental group only at meetings 2, 9, and 15. The instrument comprises eight CLIL-derived indicators, each rated on a four-point scale (total range 8–32); totals were expressed as percentages of the maximum score and categorized as low (< 65%), medium (65–79%), or high (≥ 80%). Mean activity rose from 19.6 (low) at meeting 2 to 25.8 (high) at meeting 15, an increase of 31.6% across the three observation points (Table 5; Figure 3). Indicator-level analysis showed the largest increase in speaking (+1.1), followed by listening (+0.9) and writing (+0.8). Because no comparable observations were collected in the control group and only three sessions were sampled, these data describe within-group progression and do not constitute comparative evidence of an intervention effect on engagement.

Table 5. Student activity progression (experimental group)

Observation	Meeting	M	SD	Category	Change
Initial	2	19.6	3.8	Low	—
Mid-point	9	22.4	3.4	Medium	+2.8 (14.3%)
Final	15	25.8	2.9	High	+6.2 (31.6%)

Note. SAOS totals range from 8 to 32 (eight indicators, each rated 1–4); categories reflect the percentage of the maximum score: low < 65%, medium 65–79%, high ≥ 80%. Observations were conducted in the experimental group only, so values describe within-group progression.

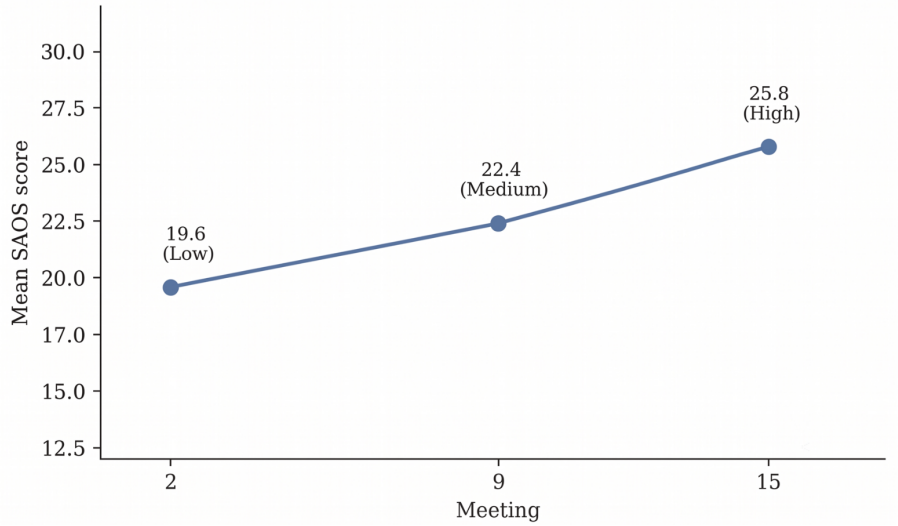

Figure 3. Progression of student activity scores (SAOS) in the experimental group across observations

Practicality and student responses

On the 14-item response questionnaire (five-point Likert scale; possible total range 14–70), the experimental group's mean score (n = 37) was 48.2 (SD = 6.8), equivalent to 68.9% of the maximum and classified as "positive" under the instrument's percentage-of-maximum intervals (very positive > 80%; positive 61–80%; neutral 41–60%; negative 21–40%; very negative ≤ 20%; the scale minimum of 14 equals 20% of the maximum, so the lowest band coincides with the scale floor); 32 students (86.5%) obtained totals in the positive or very positive categories, and the remaining five (13.5%) responded neutrally or negatively, citing workload and technical difficulties. Practicality is therefore operationalized here as favourable student perception of the tools in use, not as direct evidence that implementation is unproblematic under all conditions.

Thematic analysis of the twelve interviews followed the phases described in the Data analysis section. Open coding of the transcripts produced an initial set of codes (e.g., "re-recording until satisfied", "fear of making mistakes in front of the class", "topics match future teaching", "long preparation time", "upload failure"), which were iteratively grouped, reviewed against the full data set, and consolidated into four themes: (a) reduced speaking anxiety through asynchronicity, (b) perceived relevance to professional practice, (c) workload pressure, and (d) technical barriers.

The first two themes were voiced mainly by experimental-group students. One student explained, "I feel braver speaking English in the video because I can repeat it until I am satisfied; in class I am afraid of making mistakes in front of my friends." Another connected the course content to her future work: "The topics about culture are exactly what I will face as a teacher, so practising them in English felt directly useful." The third and fourth themes capture the costs of the approach. As one student noted, "The video assignment was useful but time-consuming. At times, I ended up recording many times because I was not satisfied with my English." Lower-proficiency interviewees specifically linked workload to language demands, while several others described upload failures or unstable connections that delayed submissions, aligning with the technical-difficulty pattern in Table 4. Because speaking anxiety was not measured with a standardized pre–post instrument, the anxiety-related theme reflects students' self-reports and is treated as interpretive rather than confirmatory evidence.

DISCUSSION

This section interprets the findings in relation to the three research questions and the wider CLIL literature, then considers the integrated nature of the intervention, its theoretical contributions, its limitations, and its implications.

Validity of the tools

The consistently high validation ratings, with every component reaching the "very valid" category, corroborate the theoretical coherence and pedagogical soundness of the tools and support the systematic logic of the Four-D model (Thiagarajan et al., 1974). Validators particularly noted how clearly the 4Cs were articulated across sessions in the RPS. This explicit integration addresses an ongoing challenge in CLIL implementation: ensuring that language and content goals are genuinely integrated rather than merely co-present (Coyle et al., 2010). Notably, the validators' initial

concerns regarding scaffolding for lower-proficiency students and rubric clarity anticipated the very patterns that later appeared in the effectiveness data, underscoring the diagnostic value of rigorous expert review.

Effectiveness and the magnitude of learning gains

The effectiveness results indicate that instruction with the integrated package was associated with practically meaningful gains beyond those observed under conventional instruction. Because the control group also improved (g = 0.29), the experimental group's advantage is not reducible to simple exposure to the course material. It cannot, however, be attributed specifically to the integration of CLIL and video. In a non-randomized comparison for a multi-component package, the estimated advantage reflects the package as a whole together with any accompanying differences. These differences include greater English exposure, novel materials and task formats, and potential novelty effects that the design cannot separate (see Limitations). The observed effect (partial η^2 = .26, 90% CI [.12, .38], large by Cohen (2013), benchmark of .14) lies toward the upper range of meta-analytic estimates for CLIL instruction (Ho et al., 2023), and the ANCOVA-based estimate benefits from the added precision that this analysis affords in quasi-experimental comparisons (Field, 2018). One caveat deserves emphasis: mastery attainment exceeded the 75% benchmark by only the equivalent of a single student, and the group difference in attainment was marginal (p = .051). Therefore, this categorical indicator should not be over-interpreted. The stronger evidence of effectiveness lies in the continuous outcomes, namely the adjusted mean difference and the normalized gain, which do not depend on where the mastery cut-off happens to fall.

Baseline proficiency and mastery attainment: an exploratory pattern

Lower baseline proficiency was descriptively associated with non-mastery, a pattern consistent with warnings that CLIL outcomes are conditioned by entry-level proficiency and require sustained, systematic support (Kovacikova, 2019; Van Mensel et al., 2020). Because this association emerged from a post hoc profile rather than from regression or interaction analyses, it identifies a hypothesis about who benefits least under the standard implementation, not an established predictor or moderator; the co-occurrence of low proficiency and technical difficulty within the same students further complicates attribution. For A2-level students, a single semester may simply provide insufficient exposure to offset the double challenge of comprehending complex content and producing the language. Crucially, non-mastery should not be equated with non-learning: the mastery threshold (≥ 70) is a local standard rather than an absolute measure of progress, and these students achieved gains exceeding the control-group average. The pattern nonetheless points to concrete refinements worth testing, including pre-teaching of academic vocabulary, sustained first-language scaffolding in initial tasks, flexible video length or alternative production formats, dedicated technology support with backup submission options, and extended preparation phases. Reporting these outcomes transparently strengthens confidence in the findings by showing that the 76% attainment rate reflects genuine individual and contextual constraints rather than a best-case scenario.

Engagement and oral-communication development

The rise in observed activity within the experimental group is consistent with the interpretation that the asynchronous video format supported oral communication by pairing CLIL's communication goals with the rehearsal and refinement benefits of video responses; the largest indicator-level improvement occurred in speaking, echoing findings that recorded video tasks lower speaking barriers and build willingness to communicate over time (Mango, 2021; Tuyet & Khang, 2020).

Practicality: viable but demanding

The largely favourable student responses, together with the candid dissenting voices, suggest that the model is viable but demanding. The recurring mention of re-recording is revealing: the platform appears to have functioned not merely as a delivery space but as a self-managed rehearsal space that students experienced as reducing anxiety (Lowenthal & Moore, 2020; Tuyet & Khang, 2020). Because anxiety was not measured with a pre–post instrument, this interpretation rests on interview accounts and is advanced as a hypothesis for future research. At the same time, the workload and technical concerns echo evidence that the benefits of video-based learning can be blunted by digital-literacy and access barriers (Miskam et al., 2019). This profile converges with recent Indonesian evidence on online CLIL in higher education: Khoiriyah et al. (2025) found that medical students' satisfaction was driven primarily by the design of learning materials, flexibility, and lecturer helpfulness, whereas online interaction was the lowest-rated aspect mirroring the combination of favorable overall responses and interaction- and technology-related frictions observed here.

The CLIL–video combination: an integrated package, not a demonstrated synergy

A central interpretive question is whether CLIL and the asynchronous video component contributed separately or synergistically. The present design cannot answer it: a two-group comparison of a combined package estimates the relative effectiveness of the whole, not the effect of each component or their interaction, which would require factorial (CLIL × video) or mediation designs; any account of synergy from these data is therefore inferential rather than demonstrated. Three observations are nonetheless consistent with an integrated reading and are registered here as hypotheses. First, the largest observed gains occurred in speaking, where CLIL's communication objectives meet the platform's low-stakes production opportunities (Coyle et al., 2010; Tuyet & Khang, 2020). Second, students' repeated re-recording suggests that the platform's affordances actively supported the communication component rather than simply accompanying it. Third, students whose video component was disrupted technically attained weaker overall outcomes, although this pattern is confounded with baseline proficiency and cannot bear causal weight. Pending component-level evidence, the defensible claim is that the tools were designed, validated, and implemented as a single package, and the present results speak to the package as implemented, not to its parts.

A note on platform availability and transferability

One contextual development requires acknowledgement. The data were collected while Flipgrid (later renamed Flip) was fully operational; Microsoft retired the standalone platform on 30 September 2024 and migrated a subset of its features into Teams for Education (Ascode, 2024; Microsoft, 2024). This does not invalidate the findings, but it clarifies their referent: the design relied on generic affordances of asynchronous video discussion low-stakes recording, the option to re-record before posting, and peer viewing and response that several currently available tools share, including video-assignment features in learning-management systems and Microsoft Teams. The study is therefore best read as evidence concerning asynchronous video-mediated communication within a CLIL framework rather than as an endorsement of a single product, and adopters should select platforms that preserve these affordances.

Situating the findings amid mixed evidence

The results should also be read against divergent findings rather than in isolation. In Indonesian vocational secondary education, Idris et al. (2025) reported minimal to non-significant between-group differences for CLIL, whereas the present study estimated a large adjusted difference on a course-specific cognitive test. Several design features plausibly bear on this contrast and merit systematic comparison in future work: intervention duration and intensity (a full 16-week course here), the nature of the materials (purpose-built, expert-validated tools versus adapted materials), participant characteristics (pre-service teachers with predominantly B1–B2 proficiency versus vocational-school learners), the type of outcome measured (integrated content–language–culture knowledge versus a specific language skill), and implementation fidelity (a single trained instructor following structured lesson plans). Rather than treating either result as definitive, the discrepancy underscores that CLIL effects are conditional on design and context (Bruton, 2011; Ho et al., 2023).

Theoretical contributions

The study's contribution is best characterized as preliminary, context-specific evidence rather than causal confirmation of a mechanism. It extends CLIL scholarship to Indonesian pre-service teacher education, where controlled evidence has been scarce (Khoiriyah, 2021; Norhasanah & Setiawan, 2023), and shows that the 4Cs Framework can be operationalized in validated, coherent tools for multicultural education. It also advances, as a working model to be tested rather than a demonstrated mechanism, the proposition that technology-mediated, low-stakes oral production can complement CLIL where language anxiety is salient. Finally, the exploratory subgroup pattern converges with prior reports that entry-level proficiency conditions CLIL outcomes (Kovacikova, 2019; Van Mensel et al., 2020) and with Indonesian evidence that practitioners differentiate language demands for lower-proficiency learners (Kurniawati et al., 2025; Sujana et al., 2023), generating hypotheses about differentiation rather than confirming moderation.

Limitations and methodological considerations

Several limitations qualify these conclusions. First, and most importantly, both conditions were taught by the same instructor, who therefore had a vested interest in the outcome. While standardized lesson plans were implemented and all tests were scored by a blind independent rater (instructor–rater ICC = .82), these safeguards are intended to address the issue of measurement objectivity, not the fact that instruction could have been delivered with differential effort across conditions; further research will need to use multiple instructors or independent review of recorded lessons to directly assess the fidelity of the instruction delivered. Second, although the control-group design bolsters inference, intact sections prohibited randomization: assignment was effectively at the class level (one section per condition), so class-level (cluster) effects cannot be separated from the treatment effect and the individual-level analysis does not

model this clustering; while baseline comparability was supported (d = 0.07) and ANCOVA controlled for pre-test differences, unmeasured variables (e.g., motivation, prior tech experience) cannot be entirely ruled out (Shadish et al., 2002). Third, experimental-group students were aware they were receiving an innovative procedure, suggesting this difference may have been due to Hawthorne and novelty effects (Parsons, 1974); further research including an active control with similar novelty may aid in isolating instructional effects. Fourth, although the sample size was adequate to detect medium effects, the single-institution sample (N = 74) limits generalizability, warranting multi-site replication. Fifth, retention and transfer to actual teaching practice have not been investigated, as only immediate outcomes were measured. Sixth, as indicated above, the design does not allow statistical separation of the CLIL and video components. Seventh, engagement was observed in the experimental group only and at three sessions, and speaking anxiety was not measured pre–post, so claims about engagement growth and anxiety reduction are descriptive or hypothesis-level; relatedly, the subgroup profile in Table 4 is post hoc and exploratory.

Practical implications

Several implications follow, conditional on the preliminary nature of the evidence. The validated tools offer an adaptable framework for bilingual teacher education where the integrated design is preserved. Implementation should begin with diagnostic assessment of entry proficiency and build in differentiated scaffolding, because lower-proficiency students were over-represented among those not attaining mastery. Reliable technical support, backup submission options, and attention to unequal access should be in place before implementation, since technical difficulties clustered among the most linguistically vulnerable learners. Workload expectations should be made explicit and calibrated to students' capacity. Consistent with Indonesian evidence that CLIL practitioners require sustained professional learning and institutional support to balance content and language objectives (Kurniawati et al., 2025; Sujana et al., 2023; Widiati et al., 2026), programmes adopting these tools should pair implementation with structured professional development for teacher educators.

Future research directions

Future work should pursue multi-site replication with larger and more diverse samples; longitudinal designs that follow graduates into teaching practice; direct pre–post measurement of speaking anxiety to test the hypothesized affective mechanism; factorial or component-isolation designs, together with analyses that model class-level clustering, to disentangle the contributions of CLIL and asynchronous video; studies of differentiated support for lower-proficiency learners; designs using multiple instructors with independent fidelity checks; comparisons of alternative asynchronous video platforms within CLIL; and qualitative work on how students experience content–language–technology integration.

CONCLUSION

This study set out to narrow the gap between bilingual policy and instructional practice in Indonesian elementary teacher preparation by developing, validating, and evaluating a set of CLIL-based learning tools mediated by asynchronous video. Three conclusions can be drawn in relation to the research questions, stated at the strength the design supports. First, the tools are valid: every component met the expert-validation criterion, with all four components in the highest ("very valid") category and high inter-validator agreement, indicating that the 4Cs Framework can serve as a coherent scaffold for integrated content–language materials in multicultural education. Second, regarding effectiveness: in this sample and context, students taught with the CLIL–asynchronous-video package achieved significantly higher covariate-adjusted post-test scores than peers receiving conventional bilingual instruction (adjusted mean difference = 12.14, 95% CI [7.23, 17.04], partial η^2 = .26), with a higher class-average normalized gain; observed engagement rose descriptively within the experimental group (the only group observed) so no comparative claim about engagement is warranted, and mastery attainment was descriptively associated with baseline English proficiency. Language development and anxiety reduction were not directly measured and therefore remain hypotheses arising from qualitative data rather than findings. Third, regarding practicality: most students (86.5%) perceived the tools positively while candidly identifying workload and technical demands, so practicality here denotes favourable perception under supported conditions rather than unconditional feasibility. Conceptually, the study offers preliminary, context-specific evidence for an instructional package in which structured content–language integration and low-stakes asynchronous video communication were designed to operate together; the two-group design evaluates the package as a whole and cannot confirm component synergy or the hypothesized anxiety-reduction mechanism. The affordances identified (low-stakes, repeatable, and peer-responsive video communication) are available on platforms beyond the specific tool used here, which supports transferability as a design argument. These

conclusions are bounded by the single-institution, non-randomized, single-instructor design with intact classes and an immediate outcome window; generalization to other programmes and to graduates' classroom practice awaits multi-site, longitudinal replication.

REFERENCES

Ascode. (2024). *Microsoft Flip has moved into Teams for education*. ASC Office of Distance Education. https://ascode.osu.edu/news/microsoft-flip-has-moved-teams-education

Banks, J. A. (1993). Multicultural education: Historical development, dimensions, and practice. *Review of Research in Education*, *19*(1), 3–49. https://doi.org/10.3102/0091732X019001003

Bartlett, M. (2018). Using Flipgrid to increase students' connectedness in an online class. *ELearn*, *2018*(12). https://doi.org/10.1145/3302261.3236703

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Bruton, A. (2011). Is CLIL so beneficial, or just selective? Re-evaluating some of the research. *System*, *39*(4), 523–532. https://doi.org/10.1016/j.system.2011.08.002

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge. https://doi.org/10.4324/9780203771587

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, *35*(3), 124–129. https://doi.org/10.1080/00031305.1981.10479327

Coyle, D. (2008). CLIL---A pedagogical approach from the European perspective. In N. H. Hornberger (Ed.), *Encyclopedia of Language and Education* (pp. 1200–1214). Springer US. https://doi.org/10.1007/978-0-387-30424-3_92

Coyle, D., Hood, P., & Marsh, D. (2010). *CLIL: Content and language integrated learning*. Cambridge University Press & Assessment.

Cummins, J. (1992). Bilingual education and English immersion: The Ramírez report in theoretical perspective. *Bilingual Research Journal*, *16*(1–2), 1–62. https://doi.org/10.1080/15235882.1992.10162628

Dalton-Puffer, C. (2011). Content-and-language integrated learning: From practice to principles? *Annual Review of Applied Linguistics*, *31*, 182–204. https://doi.org/10.1017/S0267190511000092

Dimici, K., & Başbay, A. (2021). Integrating multicultural education into English language teaching: Practical examples for language teachers. *Trends and Developments for the Future of Language Education in Higher Education*, 17–40. https://doi.org/10.4018/978-1-7998-7226-9.CH002

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

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE Publications.

Garcia, O., & Wei, L. (2014). *Translanguaging: Language, bilingualism and education*. Palgrave Macmillan. https://doi.org/10.1057/9781137385765

Goris, J., Denessen, E., & Verhoeven, L. (2019). The contribution of CLIL to learners' international orientation and EFL confidence. *Language Learning Journal*, *47*(2), 246–256. https://doi.org/10.1080/09571736.2016.1275034

Green, T., & Green, J. (2018). Flipgrid: Adding voice and video to online discussions. *TechTrends*, *62*(1), 128–130. https://doi.org/10.1007/S11528-017-0241-X

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*(1), 64–74. https://doi.org/10.1119/1.18809

Hidalgo, D. R., & Ortega-Sánchez, D. (2023). CLIL (content and language integrated learning) methodological approach in the bilingual classroom: A systematic review. *International Journal of Instruction*, *16*(3), 915–934. https://doi.org/10.29333/iji.2023.16349a

Ho, L. J., Lee, H., & Lo, Y. Y. (2023). Effects of EMI-CLIL on secondary-level students' English learning: A multilevel meta-analysis. *Studies in Second Language Learning and Teaching*, *13*(2), 317–345. https://doi.org/10.14746/ssllt.38277

Horwitz, E. (2001). Language anxiety and achievement. *Annual Review of Applied Linguistics*, *21*, 112–126. https://doi.org/10.1017/S0267190501000071

Idris, F., Said, N. E. M., & Sulaiman, N. A. (2025). The effects of CLIL on enhancing EFL writing accuracy and

language learning motivation in an Indonesian vocational high school setting. *Forum for Linguistic Studies*, *7*(6), 122–148. https://doi.org/10.30564/fls.v7i6.9281

Khoiriyah, K. (2021). Content and language integrated learning (CLIL) in Indonesian context: An overview. *Jurnal Pendidikan Progresif*, *11*(3), 587–601. https://doi.org/10.23960/jpp.v11.i3.202110

Khoiriyah, Roy, S., & Hasan, M. K. (2025). Exploring the implementation of online learning at medical-CLIL classes: EFL students' satisfaction. *Journal of Education and E-Learning Research*, *12*(4), 588–598. https://doi.org/10.20448/jeelr.v12i4.7798

Kholili, M., Rachmajanti, S., & Wulyani, A. N. (2024). CLIL in senior high school: Empowering English language teaching and learning in Merdeka Curriculum. *Didaktika: Jurnal Kependidikan*, *13*(4), 5739–5756. https://doi.org/10.58230/27454312.1091

Kovacikova, E. (2019). Development of speaking at primary schools through CLIL. *XLinguae*, *12*(2), 17–26. https://doi.org/10.18355/XL.2019.12.02.02

Kurniawati, A., Eko, A., & Atmojo, P. (2025). Teachers' beliefs and practices in implementing CLIL in an Indonesian primary school. *Journal of English Language Pedagogy, Literature, and Culture*, *10*(1), 28–39. https://doi.org/10.35974/acuity.v10i1.3222

Lowenthal, P. R., & Moore, R. L. (2020). Exploring student perceptions of flipgrid in online courses. *Online Learning Journal*, *24*(4), 28–41. https://doi.org/10.24059/OLJ.V24I4.2335

Mango, O. (2021). Flipgrid: Students' perceptions of its advantages and disadvantages in the language classroom. *International Journal of Technology in Education and Science*, *5*(3), 277–287. https://doi.org/10.46328/IJTES.195

Microsoft. (2024). *Using Flip on Teams - Microsoft Q&A*. https://learn.microsoft.com/en-us/answers/questions/4419083/using-flip-on-teams

Miskam, N. N., Aminabibi, A., & Saidalvi, S. (2019). The use of flipgrid for teaching oral presentation skills to engineering students. *International Journal of Recent Technology and Engineering*, *8*(1C2), 636–641. https://www.ijrte.org/portfolio-item/a10880581c219/

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, *17*(8), 873–890. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I

Nieto, S., & Bode, P. (2012). *Affirming diversity: The sociopolitical context of multicultural education* (6th ed.). Pearson.

Norhasanah, & Setiawan, R. (2023). Content and language integrated learning (CLIL) implementation for Indonesian EFL learners: A case study. *SAGA: Journal of English Language Teaching and Applied Linguistics*, *4*(2), 89–100. https://doi.org/10.21460/saga.2023.42.165

Nurbatra, L. H. (2022). Infusing culturally responsive teaching in higher education: Insights for multicultural education in Indonesia. *Journal of Innovation in Educational and Cultural Research*, *3*(4), 722–730. https://doi.org/10.46843/jiecr.v3i4.321

Parsons, H. M. (1974). What happened at Hawthorne? *Science*, *183*(4128), 922–932. https://doi.org/10.1126/science.183.4128.922

Setyaningrum, R. W., Setiawan, S., & Anam, S. (2022). Translanguaging as a scaffolded practice in a primary school content and language integrated learning context during the Covid-19 pandemic. *European Journal of Educational Research*, *11*(4), 2043–2055. https://doi.org/10.12973/eu-jer.11.4.2043

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Stoszkowski, J. (2018). Using Flipgrid to develop social learning. *Compass: Journal of Learning and Teaching*, *11*(2). https://doi.org/10.21100/COMPASS.V11I2.786

Sujana, I. M., Waluyo, U., Fitriana, E., & Sudiarta, I. W. (2023). The potentials and limitations of applying content and language integrated learning (CLIL) approach to English teaching for medical students. *World Journal of English Language*, *13*(2), 331–335. https://doi.org/10.5430/wjel.v13n2p331

Talaue, F. G., & Hendijani, R. B. (2025). Holistic pedagogical support to address postgraduate students' challenges in academic publishing: The case of a business school in Indonesia. *Asia-Pacific Education Researcher*, *34*(2), 565–573. https://doi.org/10.1007/s40299-024-00878-7

Thiagarajan, S., Semmel, D. S., & Semmel, M. I. (1974). *Instructional development for training teachers of exceptional children: A sourcebook*. A joint publication of the Leadership Training Institute/Special Education, University of Minnesota; The Center for Innovation in Teaching the Handicapped (CITH), Indiana University.

Tuyet, T. T. B., & Khang, N. D. (2020). The influences of the Flipgrid app on Vietnamese EFL high school learners' speaking anxiety. *European Journal of Foreign Language Teaching*, *5*(1), 128–149. https://doi.org/10.46827/EJFL.V5I1.3264

Van Mensel, L., Hiligsmann, P., Mettewie, L., & Galand, B. (2020). CLIL, an elitist language learning approach? A background analysis of English and Dutch CLIL pupils in French-speaking Belgium. *Language, Culture and Curriculum*, *33*(1), 1–14. https://doi.org/10.1080/07908318.2019.1571078

Wei, L. (2018). Translanguaging as a practical theory of language. *Applied Linguistics*, *39*(1), 9–30. https://doi.org/10.1093/applin/amx039

Widiati, U., Muniroh, S., El Khoiri, N., Metaria, M., Murtafi'ah, M., Hanifiyah, L., & Sharif, T. I. S. T. (2026). Enhancing Indonesian EFL teachers' literacy teaching strategies and professional learning through designing CLIL-based reading activities. *Studies in English Language and Education*, *13*(1), 189–210. https://doi.org/10.24815/siele.v13i1.153

Wright, W. E., & Baker, C. (2017). Key concepts in bilingual education. In O. García, A. M. Y. Lin, & S. May (Eds.), *Bilingual and Multilingual Education* (pp. 65–79). Springer International Publishing. https://doi.org/10.1007/978-3-319-02258-1_2

ACKNOWLEDGEMENT

The authors thank Universitas Negeri Semarang for institutional support and for permission to conduct the research in the PGSD FIPP UNNES bilingual classes.

AUTHOR CONTRIBUTION STATEMENT

Conceptualization, N.S. and Z.X.; methodology, N.S. and Z.X.; software, N.S.; validation, N.S. and Z.X.; formal analysis, N.S. and Z.X.; investigation, N.S.; resources, N.S.; data curation, N.S.; writing-original draft preparation, N.S.; writing, review and editing, N.S. and Z.X. All authors have read and agreed to the published version of the manuscript.

AI DISCLOSURE STATEMENT

The authors declare that the generative artificial intelligence (AI) tool, Gemini, was used exclusively for language editing and/or grammatical improvement. The use of AI did not influence the scientific content, study design, data analysis, data interpretation, results, or conclusions of the manuscript. Full responsibility for the content remains with the authors.

***Novi Setyasto (Corresponding Author)**
Department of Primary School Teacher Education, Faculty of Education and Psychology,
Universitas Negeri Semarang,
UNNES Sekaran Campus, Gunungpati, Semarang 50229, Central Java, Indonesia
Email: novisetyasto@mail.unnes.ac.id

Zhao Xinzhe
Department of Normal Education,
Liuzhou City Vocational College,
No.1 Wenyuan Road, Guantang Avenue, Yufeng, Liuzhou, 545036, Guangxi, China
Email: 1605657412@qq.com