Effects of flipped language classrooms on learning outcomes in higher education: A Bayesian meta-analysis

Xieling Chen
Guangzhou University, China

Di Zou
The Education University of Hong Kong

Gary Cheng
The Education University of Hong Kong

Haoran Xie
Lingnan University, Hong Kong

Fan Su
The Education University of Hong Kong

Despite accumulated evidence demonstrating the effectiveness of flipped language classrooms in higher education, there is no quantitative examination of the extant empirical studies to draw a general conclusion. Based on Bayesian methodologies and 26 effect sizes, this study quantitatively examines empirical studies that investigated flipped language classrooms' effects on learning outcomes in higher education. Our results indicate a large overall effect in favour of the effectiveness of flipped language classrooms. Subgroup analyses indicated that intervention duration, target languages, outcome types, allocation, and school locations were significantly related to the variability in language learning outcomes. A low risk of publication bias was identified. This study concluded that the flipped language classroom was a promising pedagogical approach to promoting language learning. Findings provided insights into an evidence-informed application of flipped language classrooms, for example: (1) sufficient face-to-face time to maximise the effectiveness of flipped language classrooms; (2) making flipped design adjustments based on student responses during long-term intervention; (3) giving students pre-training of flipped language classrooms and showing them the underlying benefits; (4) flipping basic contents of language learning and teaching complex contents face-to-face; and (5) adopting scaffolding strategies like code-switching to scaffold lower achievers.

Implications for practice or policy:

- Instructors should flip writing and speaking courses with enough face-to-face time and technical support being provided to students.
- Instructors should consider time variance's effects on learning performance and seek ways to maintain learners' interest.
- Instructors should pre-train learners of flipped learning before implementation.
- Instructors should include practices, quizzes, and asynchronous online interaction tools in pre-class activities to check learners' understandings and promote interaction and feedback provision.

Keywords: flipped classroom, higher education, language learning, learning outcomes, Bayesian meta-analysis

Introduction

As a novel instructional strategy, flipped classrooms minimise direct lecturing, maximise interaction and collaboration, and enhance social interaction, teamwork, and cultural diversity (Tan et al., 2017). In

flipped classrooms, homework and teaching are swapped to enable students to learn new knowledge through videos before class and actively participate in in-class activities (Nwosisi et al., 2016). Since the implementation of a flipped chemistry course in 2007, flipping learning has been popularly applied in higher education for learning a wide range of subjects, including foreign/second languages (Bergmann & Sams, 2012).

Although research on flipped language classrooms in higher education is flourishing (e.g., Bezzazi, 2019; Haghighi et al., 2019; Lin et al., 2018; Özkurkudis & Bümen, 2019; Vaezi et al., 2019; Yang et al., 2018), the field is relatively nascent. There is a need for an examination of flipped language classrooms' effects on learning outcomes in higher education, compared with traditional instruction approaches, to draw a general conclusion. Accordingly, this study presents the first in-depth meta-analysis designed to address this need.

Flipped language classrooms

Language acquisition requires time and practice (Turan & Akdag-Cimen, 2020). Students need to participate in adequate activities to master a new language. However, students usually have limited opportunities for practice in traditional classrooms. The flipped language classrooms can provide students with rich opportunities to practice listening, speaking, and writing through group activities in class as they restrict the instruction to outside the classrooms.

According to Amiryousefi (2019), flipped classrooms benefit foreign language learners by promoting student-centred learning and autonomy. Basal (2015)'s analysis of 47 pre-service language instructors' perspectives suggested that flipped language classrooms promote students' learning at their paces, enhance their preparation and engagement, and remove time-relevant issues in the classrooms. Chen Hsieh et al. (2017)'s mixed-methods study with 48 second-year student participants suggested flipped language classrooms had positive effects on idiomatic knowledge acquisition and learner engagement and motivation.

Regarding the skills involved in second language learning, researchers have assessed the flipped approaches' effects on different language learning outcomes, for example, flipped interventions targeting writing performance contribute to increased writing achievement and student engagement (e.g., Afrilyasanti et al., 2017; Leis et al., 2015). Regarding speaking courses, flipped approaches have been found to improve second language oral proficiency and cultivate autonomous learners to gain an in-depth understanding of course content (e.g., Amiryousefi, 2019; Wang et al., 2018). Researcher focusing on flipped grammar learning (e.g., Thaichay & Sitthitikul, 2016; Webb & Doman, 2016) has highlighted flipped approaches' potential for promoting grammar performances and making students feel comfortable and confident about second language grammar use. Several studies focusing on flipped vocabulary learning concluded that flipped language classrooms motivate learners to develop both receptive and productive vocabularies effectively for communication interaction (e.g., Arslan, 2020; Kirmizi & Kömeç, 2019; Zhang et al., 2016)

The results of these studies demonstrate flipped language classrooms' significance in encouraging learning at one's own pace by taking self-learning responsibilities (Amiryousefi, 2019). Flipped language classrooms' potential for promoting learners' language learning achievement and engagement, lowering their cognitive load by virtue of flexible time and dynamic and interactive learning environments, and facilitating learners' in-depth understanding of concepts (Amiryousefi, 2019), has been well evidenced in the literature. Given the flexibility of the approach, flipped language classrooms have gained increasing attention in recent years (Bergmann & Sams, 2012) and have been applied predominantly in higher education to promote in-depth discussions and knowledge applications (Lundin et al., 2018).

Literature review

There have been several systematic reviews on flipped classrooms (e.g., Akçayır & Akçayır, 2018; Lo & Hew, 2017; Lundin et al., 2018). Several meta-analysis studies (e.g., Låg & Sæle, 2019; Shi et al., 2020; Strelan et al., 2020; van Alten et al., 2019) have demonstrated flipped learning's effectiveness in enhancing learning outcomes compared to conventional classrooms. These reviews focused on the general field of flipped learning (e.g., Chen et al., 2018; Cheng et al., 2019) or its applications in specific educational areas such as mathematics (Lo et al., 2017), health (Hew & Lo, 2018), nursing (Tan et al., 2017; Xu et al., 2019), and engineering (Lo & Hew, 2019).

In the field of flipped language classrooms, there have been reviews presenting systematic overviews and comprehensive perspectives of flipped language classrooms (e.g., Arslan, 2020; Jiang et al., 2022; Turan & Akdag-Cimen, 2020; Zou et al., 2020), as shown in Table 1. For example, Turan and Akdag-Cimen (2020) systematically reviewed the trends and the major findings of 43 studies focusing on flipped English language teaching. Their review indicated that the flipped English language teaching gained popularity after 2014. Additionally, mixed and quantitative research methods have been widely adopted, with speaking and writing abilities studied the most. Jiang et al. (2022) examined 33 flipped language classroom studies in the Social Sciences Citation Index up to 2018. Four issues were investigated: trends and publication features, research topics, roles of technologies, and integrated second/foreign language learning theories, models, and strategies. Jiang et al.'s (2022) study revealed a bias towards outcome-driven quantitative over process-driven qualitative studies. Arslan (2020) discussed the benefits and challenges of flipped English language teaching by systematically reviewing 78 studies. The review indicated increasing interest in the investigated field. Furthermore, most of the studies had college students as their participants. Additionally, flipped instruction's positive effects on learners' language learning, such as enhancement of writing and speaking skills, were commonly reported. Within the above-mentioned reviews, some issues have been commonly investigated, for example, annual distribution, prolific countries in publishing flipped language classroom studies, the educational levels of participants, and research methods. However, a quantitative examination of the effectiveness of flipped foreign language education is lacking.

A meta-analysis by Vitta and Al-Hoorie (2020) focused on flipped language classrooms. They revealed a large overall effect size of 0.99 favouring flipped language classrooms, however, the Trim and Fill approach with an effect size shrinking to 0.58 indicated publication bias (Duval & Tweedie, 2000a; 2000b). One potential cause for the large publication bias in the study of Vitta and Al-Hoorie was because they included a large number of non-indexed journal articles, which, according to Paiva et al. (2017), are likely to be conducted by researchers lacking experience in well-established interventions. Thus, these studies should be included with caution. Another possible explanation for the large publication bias was because Vitta and Al-Hoorie used traditional meta-analysis methodologies that are more likely to induce publication bias resulting from small-study effects. Advanced meta-analysis techniques that can overcome this drawback appear essential. Second, in subgroup analysis, Vitta and Al-Hoorie (2020) only considered participant- and publication-level factors. However, the inclusion of factors related to treatment and design characteristics can draw attention to the effective design of flipped language classrooms. Furthermore, Vitta and Al-Hoorie (2020) only included studies published before August 2019. With recent advances in technology, the inclusion of more up-to-date studies can provide a more complete overview of flipped language classroom implementations and strategies. Additionally, Vitta and Al-Hoorie (2020) specifically focused on flipped language classrooms in all education levels. As flipped learning dominates the higher education sector (Lundin et al., 2018), there is a need to narrow the focus to only cover certain aspects of higher education to understand the effectiveness of flipped language classrooms in higher education.

Table 1
Examples of recent reviews on flipped learning and its relevant topics

Dimension	Reviewer(s) and year	Methods	Subjects	Educational level	Number of articles	Reviewed period
Flipped classrooms	Akçayır & Akçayır (2018)	Systematic analysis	Cross-disciplinary	Cross-levels	71	2000–2016
	Lo & Hew (2017)	Systematic analysis	Cross-disciplinary	K-12 education	15	2013–2016
	Lundin et al. (2018)	Systematic analysis	Cross-disciplinary	Higher education	31	2010–2015
	Chen et al. (2018)	Meta-analyses	Cross-disciplinary	Cross-levels	46	2012–2016
	Cheng et al. (2019)	Meta-analyses	Cross-disciplinary	Cross-levels	55	2013–2016
	van Alten et al. (2019)	Meta-analyses	Cross-disciplinary	Cross-levels	114	2010–2016
	Strelan et al. (2020)	Meta-analyses	Cross-disciplinary	Cross-levels	198	2012–2018
	Låg & Sæle (2019)	Meta-analyses	Cross-disciplinary	Cross-levels	271	2010–2017
	Shi et al. (2020)	Meta-analyses	Cross-disciplinary	Higher education	33	2013–2017
Flipped learning in subject areas	Tan et al. (2017)	Systematic analysis	Nursing	Cross-levels	29	2015–2016
	Lo et al. (2017)	Systematic analysis	Mathematics	Cross-levels	21	2014–2016
	Xu et al. (2019)	Meta-analyses	Nursing	Cross-levels	22	2015–2018
	Lo & Hew (2019)	Meta-analyses	Engineering	Cross-levels	31	2008–2017
	Hew & Lo (2018)	Meta-analyses	Health	Cross-levels	28	2012–2017
Foreign language classrooms	Arslan (2020)	Systematic analysis	Language	Cross-levels	78	2014–2018
	Jiang et al. (2022)	Systematic analysis	Language	Cross-levels	33	2010–2018
	Turan & Akdag-Cimen (2020)	Systematic analysis	Language	Cross-levels	43	2016–2017
	Zou et al. (2020)	Systematic analysis	Language	Cross-levels	34	2015–2019
	Vitta & Al-Hoorie (2020)	Meta-analyses	Language	Cross-levels	56	Till August 2019

Purpose of the study and research questions

Empirical studies have evaluated flipped learning classrooms' effectiveness in higher educational contexts by comparing them with traditional classrooms, with differing results. Past reviews on flipped language classrooms commonly adopted systematic analysis methodologies (e.g., Arslan, 2020; Jiang et al., 2022; Turan & Akdag-Cimen, 2020). Thus, the conclusions could not be generalised due to the lack of comparison with non-flipped learning. These reviews also failed to provide "the magnitude of an effect, the strength of the relationship, or the importance of a finding observed within a group of studies" (Norris

& Ortega, 2000, p. 425). In other words, these reviews cannot present the cause-effect relationship. The only meta-analysis study similar to this study was conducted in a broad educational context (Vitta & Al-Hoorie, 2020). A systematic meta-analysis on flipped learning classrooms in higher education has yet to be published. Flipped learning classroom is more suitable in higher education because college students commonly have higher self-regulation abilities needed for flipped learning activities compared to K-12 students (Rodrigues et al., 2016; Tomas et al., 2019). This study sought to present an up-to-date quantitative assessment of the overall effect size of the flipped learning classrooms on learning outcomes in higher education based on cause-effect relationship identification between flipped instructions and language learning outcomes. In addition to the overall analysis, we also conducted moderator analyses based on several instructional design characteristics (e.g., quizzes and face-to-face time) to identify potential differential effects of flipped learning classrooms. Our findings may be beneficial to various stakeholders, including teachers, policymakers, and researchers interested in flipped learning classrooms. From the methodological perspective, this study adopted advanced Bayesian meta-analysis to overcome the limitations found in Vitta and Al-Hoorie's (2020) work, where results were potentially biased due to small-study effects in their traditional meta-analysis methodology. This study was guided by three research questions:

1. What were the overall quantitative characteristics of flipped learning classroom studies?
    a. What were the flipped learning classrooms' overall effects on learning outcomes compared to non-flipped instruction?
    b. How heterogeneous were the effects of the flipped learning classroom studies?
2. Did characteristics (intervention duration, publication type, target language, outcome type, regions, teachers, group equivalence tests, allocation, classroom time, and quizzes) moderate the flipped learning classrooms' effect?
3. What was the potential risk of publication bias?

Methods

Literature search and selection

Figure 1 depicts the steps of data search and selection. The data search involved three steps.

1. Identification of search terms. Three groups of search terms (Table A1 in the Appendix) related to flipped learning, language learning, and empirical research design, respectively, were identified based on previous studies (e.g., Chen, Zou, Xie & Cheng, 2021; 2022; Chen, Zou, Xie & Su, 2021; Cheng et al., 2019; van Alten et al., 2019; Wang et al., 2020).
2. Conduct of a database search. Search terms in Table A1 were adopted to retrieve flipped learning classroom studies from academic databases, including Educational Resources Center, ProQuest Digital Dissertations, Web of Science, Scopus, Linguistics and Language Behavior Abstracts, OpenDissertations, and Academic Search Premier. Journal articles, dissertations/theses, and conference papers were considered. The search period started in 2000 when the concept of flipped classroom emerged (Baker, 2000).
3. Conduct of a data search in academic journals. Academic journals focusing on educational technology and second/foreign language learning are essential in publishing flipped learning classroom studies. In addition to the 31 academic journals that are considered prominent academic journals in the field of English as a second/foreign language (Lin & Lin, 2019), we included eight further academic journals: RELC Journal, AsiaTEFL, ELT Journal, Interactive Learning Environments, Australasian Journal of Educational Technology, English for Academic Purposes, Language Teaching, and Language Learning. These are also important journals in the fields of language learning and educational technology.

Our search generated 1476 records. A total of 440 duplicates were removed. When screening titles and abstracts of the remaining 1036 studies, we eliminated another 396 studies due to their irrelevance to

flipped learning classrooms. To ensure screening reliability, two researchers examined each study, leading to an inter-reliability of 94%, with inconsistencies being addressed through discussions.

In the following steps, full texts of the remaining 640 studies were examined for eligibility based on the listed criteria in Appendix A, Table A2. Studies were eliminated if they: (1) were not written in English; (2) did not measure flipped learning classrooms' quantitative effects in higher education; and (3) did not involve non-flipped classrooms as a control group. Additionally, studies measuring other dependent variables than language outcomes were eliminated; for example, self-regulation and critical thinking. Finally, 26 studies remained for further meta-analysis.

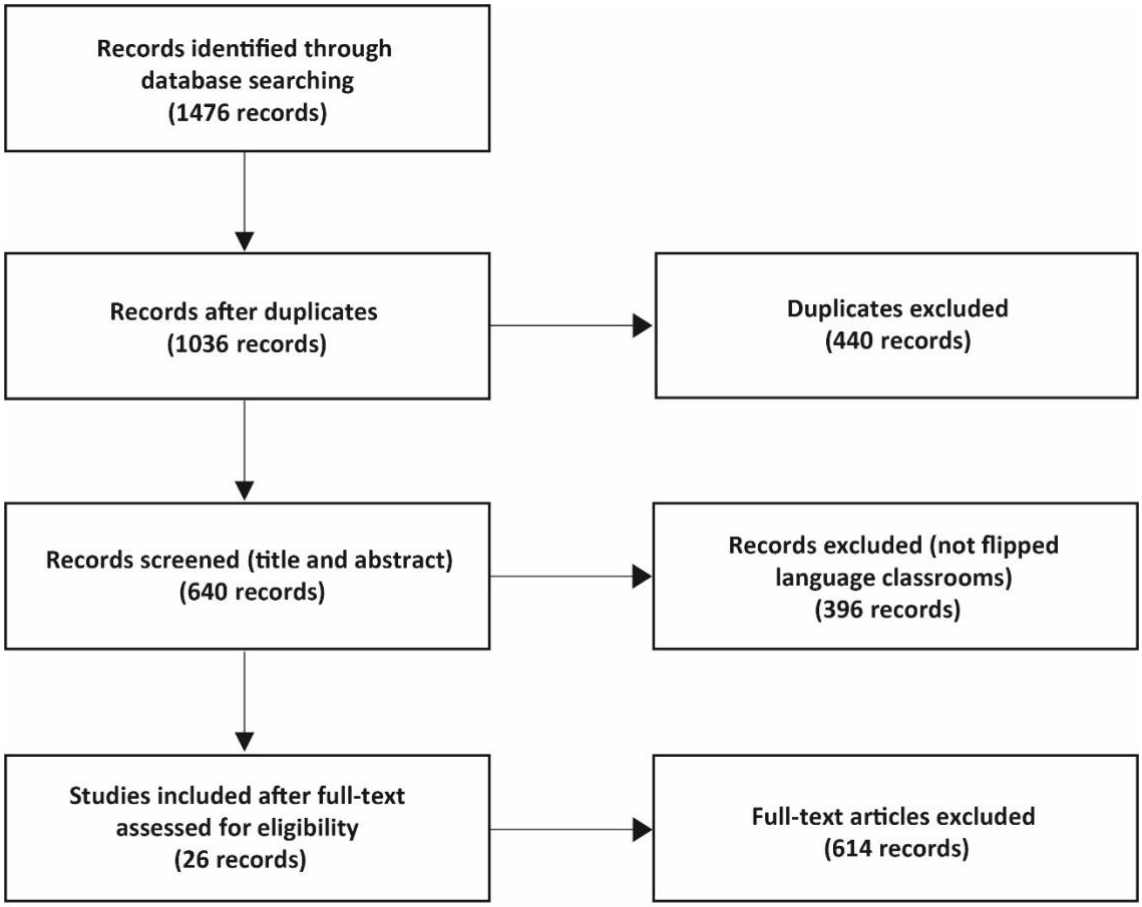

Figure 1. Literature search and selection

Coding rubrics

A coding rubric (Appendix A, Table A3) was proposed based on relevant meta-analyses (e.g., van Alten et al., 2019; Zhang et al., 2021) to code key characteristics of the studies, including five treatment-related factors (intervention duration, allocation, teachers, group equivalence tests, and language outcomes), two participant-related factors (target languages and school locations), one publication-related factor (publication types), and two design characteristics (face-to-face time and quizzes). To clearly understand the features of allocation, group equivalence tests, language learning outcomes, face-to-face time, and quizzes, we provide specific explanations as follows.

- Allocation: Allocation (van Alten et al., 2019) indicated how participants were allocated to experimental and control groups, including random allocation at the individual level, at the pre-existing group level (e.g., one of the two classes was randomly selected as an experiment group and other as a control group), and no randomisation.

- Group equivalence test: This indicated how the prior knowledge of participants in the flipped and control groups was evaluated before interventions. If the statistical test was used in the pre-test, the study was coded as "tested, equal" or "tested, not equal", depending on the test result. That is, a study by Özkurkudis and Bümen (2019) was coded as "tested, equal" as no statistical difference in the pre-test was found. On the contrary, a study was coded as "tested, not equal" if a significant difference was found. For instance, in Leis et al. (2015), a significant difference in the pre-test between experimental and control groups was identified using a one-way analysis of covariance. Additionally, some studies (e.g., Bicen & Beheshti, 2022) specified equivalence with a simple statement about similar initial language levels among participants in the control and experimental groups without using statistical analysis. These studies were coded as "not tested, descriptive statement". Furthermore, some studies (e.g., Hung, 2017; Salem, 2018) were coded as "not tested, no descriptive statement" because they did not mention experimental and control groups' equivalence either through a statistical test or a descriptive statement.
- Language learning outcomes: the language learning outcomes reported were categorised into writing, speaking, vocabulary/grammar, listening, and multiple (with more than one skill being reported).
- Face-to-face time: Some researchers argued that students spent more learning time before class at the expense of less face-to-face time (Baepler et al., 2014). Other researchers (e.g., Heyma et al., 2015) argued that reducing face-to-face time in flipped classrooms might result in worse learning outcomes. This study examined the influence of face-to-face time on flipped learning classrooms' effect variance. Specifically, if face-to-face time was equal for both experiment and control groups, the study was coded as FC (flipped classroom) = TC (traditional classroom), otherwise, FC < TC.
- Quizzes: We considered quizzes as formative assessments, with scores (both before and in class during interventions) evaluating student comprehension and language learning performance. In contrast, practices without scores were not considered.

Two researchers independently coded the 26 flipped learning classroom studies, with the inter-rater reliability (Cohen's κ) of 0.84. During coding, frequent discussions were conducted to achieve agreement and reliability. The coding results are presented in Appendix, Table A4.

Effect sizes and effect-size homogeneity

The effect sizes of the 26 studies were computed following van Alten et al.'s (2019) order of preference to maximise meta-analysis precision. The most favoured approach was to compute effect sizes using adjusted post-test means to consider possible pre-treatment differences. The second was the use of pre-test-post-test-control group designs. Third, when the above two approaches could not be achieved, effect sizes were calculated based on raw means and standard deviations in post-tests. However, such an approach was only applicable to the cases showing not much difference in pre-tests. When the above approaches were impossible, effect size calculation based on inferential statistics (e.g., *t*- or *F*-test) was the final choice. The effect-size homogeneity refers to the agreement/disagreement of flipped instruction effectiveness compared to non-flipped instruction on language learning outcomes. *Q* statistic (Hedges, 1982) and I^2 index (Higgins et al., 2003) were adopted to explore the effect-size homogeneity.

The standardised mean difference for the k^{th} of $K = 26$ effect sizes was calculated using Equation (1), in which $\bar{Y}_k^T$ and $\bar{Y}_k^C$ were mean language learning outcomes for the experimental and control groups, separately. S_k^P was the pooled standard deviation, and n_k^T and n_k^C were flipped and non-flipped sample sizes, respectively. $d_k>0$ indicates a mean difference preferring the flipped group, whereas $d_k<0$ prefers the non-flipped group.

$$d_k = \left(1 - \frac{3}{4\left(n_k^T + n_k^C - 2\right) - 1}\right) \frac{\bar{Y}_k^T - \bar{Y}_k^C}{S_k^P} \qquad (1)$$

The computation of d_k using t and F statistics are as Equations (2) and (3), in which t was the t-statistic testing and F_k was the F statistics for one-way analysis of variance.

$$d_k = \left(1 - \frac{3}{4\left(n_k^T + n_k^C - 2\right) - 1}\right) t \sqrt{\frac{n_k^T + n_k^C}{n_k^T n_k^C}} \quad (2)$$

$$d_k = \left(1 - \frac{3}{4\left(n_k^T + n_k^C - 2\right) - 1}\right) \left(\pm \sqrt{\frac{F_k\left(n_k^T + n_k^C\right)}{n_k^T n_k^C}}\right) \quad (3)$$

After computing effect sizes, the sample effect-size variances were calculated using Equation (4).

$$v_k = \left(1 - \frac{3}{4\left(n_k^T + n_k^C - 2\right) - 1}\right)^2 \left(\frac{n_k^T + n_k^C}{n_k^T n_k^C} + \frac{d_k^2}{2\left(n_k^T + n_k^C\right)}\right) \quad (4)$$

When calculating effect sizes, this study included only one effect size for a single study. According to Lipsey and Wilson (2001), the inclusion of over one effect size per study leads to statistical dependence, generating a biased overall effect size. To achieve reliability, we adopted two rules to decide which effect size to be included. If over one outcome variable was adopted to evaluate the same construct (e.g., Hamdani, 2019; Lin & Hwang, 2018), we averaged the effect sizes. For instance, Hamdani (2019) used listening, writing, speaking, and reading tests to measure students' learning outcomes. When a study contained multiple treatments, we choose the effect size related to our research target. For instance, for the study by Hung (2015), which used flipped group treatment, semi-flipped group treatment, and non-flip group treatment, we selected flipped group and non-flipped group treatments.

The effect-size homogeneity refers to the agreement (or disagreement) of flipped instruction effectiveness in comparison to non-flipped instruction on language learning outcomes. To explore the effect-size homogeneity, we adopted Q statistic and I^2 index. Following Chen and Yang (2019), the between-class variance component Q_B was computed by Q_T minus Q_W, where Q_T was the variance component of the effect size and Q_W was the within-class variance component. I^2 was used to measure the remained percentage of effect-size variation with the consideration of sampling error. The Q and I^2 were calculated using Equations (5) and (6), where $\bar{d}_{IV}$ was a weighted mean via inverse-variance weighting and $Q \sim \chi^2(k - 1)$. The larger the Q, the higher the disagreement or heterogeneity among effect sizes. I^2 equaling to 0, 0.25, 0.5, and 0.75 showed no, low, moderate, and high variations, separately.

$$Q = \sum_{k=1}^{K} v_k^{-1} \left(d_k - \bar{d}_{IV}\right)^2 \quad (5)$$

$$I^2 = \frac{Q - K + 1}{Q} \times 100\% \quad (6)$$

Bayesian meta-analysis

This study used Bayesian meta-analysis approaches to address potential limitations of traditional meta-analysis approaches, including publication bias caused by small-study effects and the restricted number of variables to be included due to the small sample size. Bayesian meta-analysis approaches assume parameters coming from a super-population, representing an infinite population of abstractions with distinctive population characteristics, and the finite population itself is a sample of a super-population (Royall, 1970).

With a set of sample effect sizes T and variances V, overall mean μ and between-studies standard deviation τ can be estimated as follows. A proportional statistical approach combining likelihood $p(T|\mu, \tau, \vartheta, V)$ and prior $p(\mu, \tau, \vartheta)$ information was used to calculate posterior distribution $p(\mu, \tau, \vartheta|T, V)$, with ϑ being the real effect sizes.

This study adopted a random-effect approach to measuring the mean effect size and variability among effect sizes. The hierarchical distributional form of the random-effects model is expressed as follows:

$d_k \sim N(\delta_k, v_k)$
$\delta_k \sim N(\mu, \tau^2)$
$\mu \sim N(0, 100^2)$
$\tau \sim DM(s_0)$

δ_k represented the real value of the k^{th} effect size. DM(s_0) was a DuMouchel prior distribution (DuMouchel, 1994). We compared prior distributions such as uniform and square root for τ to resolve sensitivity. Because of no significant differences, this study presented results using DuMouchel for τ. Additionally, we computed Bayes factors to provide additional parameter information.

This study graphically and quantitatively assessed μ and τ estimates through marginal posterior distributions. Specifically, the marginal posterior was plotted; the median, mean, and standard deviation, and the 95% highest posterior density intervals (HPDI) were reported. We also provided Bayes factor results, the numerical tests with null models (no overall effect, = μ_0 and no between-studies variability, = τ_0).

For subgroup analysis, we considered 10 moderators (intervention duration, publication type, language learned, language outcome type, school location, teacher, group equivalence test, allocation, face-to-face time, and quizzes) as categorical variables. We conducted subgroup analyses by examining each of the 10 moderators separately. For each category within a moderator, an individual Bayesian meta-analysis was performed. For instance, for moderator intervention duration containing two subgroups (> 10 weeks or 1 – 10 weeks), two Bayesian meta-analyses were performed, with one examining the intervention over 10 weeks and the other examining the intervention shorter than 10 weeks. As per the overall analyses, we examined mean effect sizes and between-studies standard deviation, respectively.

Publication bias

Funnel plots were produced to assess potential asymmetry and heterogeneity. Begg's rank correlation test, Egger's regression test, Trim-and-Fill test, and Vevea and Hedges weighted function were also applied to detect potential publication bias. Data analysis in this study was conducted based on R packages bayesmeta, weightr, metafor, esc, and meta.

Results

Overall analyses

The overall analysis of mean effect size (Appendix A, Table A5) and between-studies standard deviation (Appendix, Table A6) together with a forest plot (Figure 2) indicated positive effect size point estimates, with 21 out of 26 differing from zero. As for marginal posterior distribution (Figure 3), in comparison to the non-flipped group, the mean effect of the flipped group on language learning outcomes was larger, with $\bar{d}_{overall}$ = 1.096 (HPDI = [0.887, 1.307]) and a true value differing from zero. This indicated a true standardised mean difference between flipped and non-flipped groups lying between 0.887 and 1.307 with 95% likelihood and a low Bayes factor < 0.001.

The two non-Bayesian heterogeneity assessments (i.e., $Q_T(25) = 74.796$, $p < 0.001$, and $I^2 = 67.2\%$) demonstrated a possible variability among effect sizes. The marginal posterior distribution mean of $\hat{\tau}_{overall}$ = 0.427 (HPDI = [0.241, 0.627]) and a Bayes factor < 0.001 indicated disagreements among the 26 effect sizes. The causes of the disagreements could be explained by the results of the subgroup analysis.

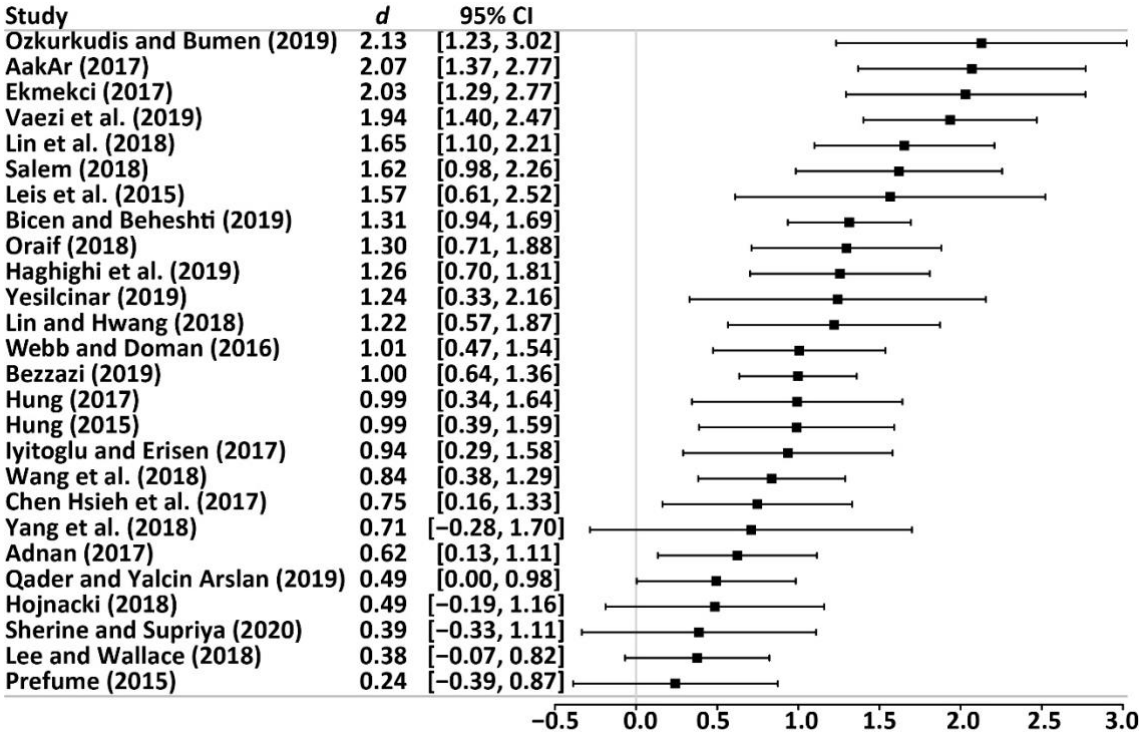

Figure 2. Forest plot of effect sizes (CI = confidence interval)

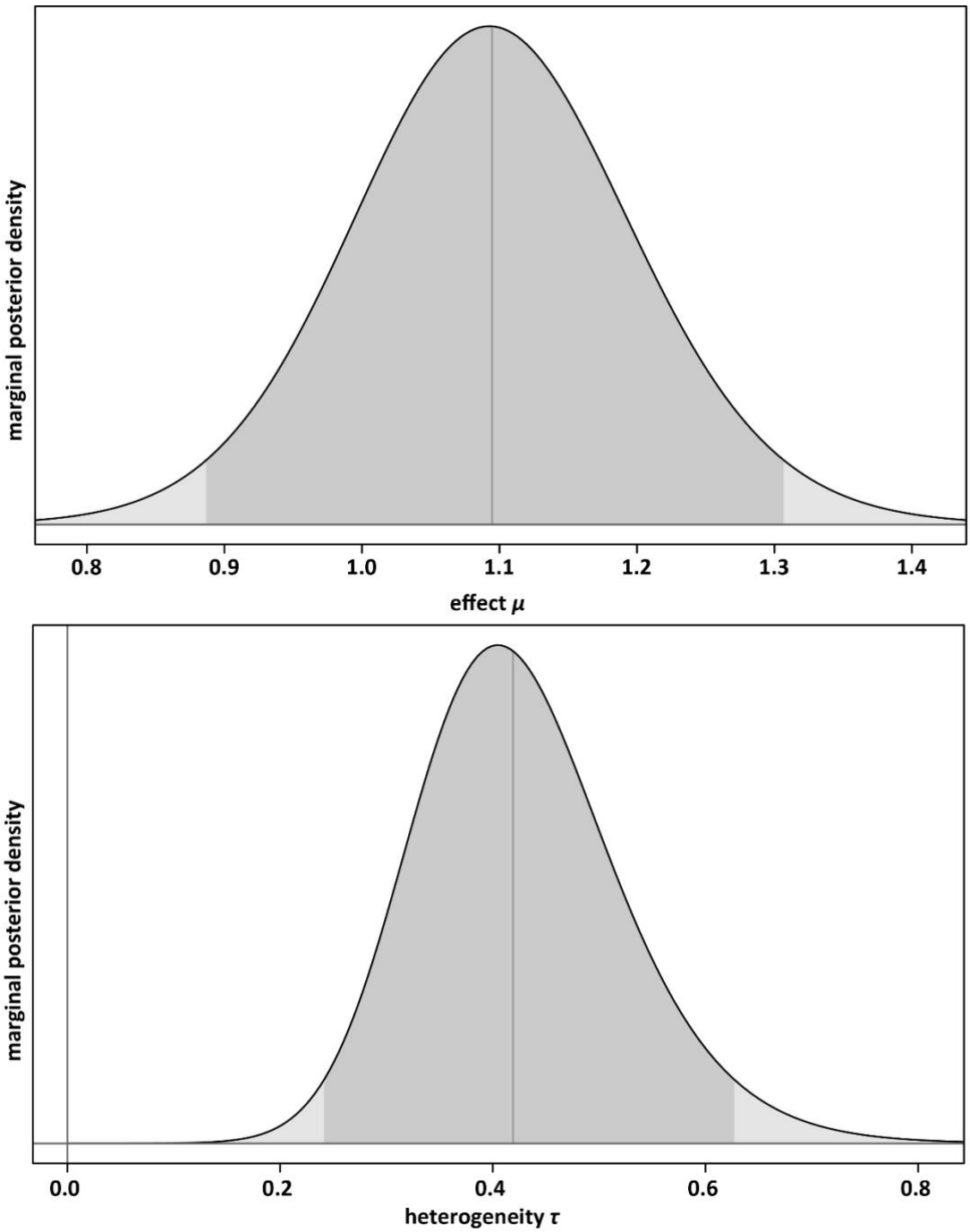

Figure 3. Overall mean and between-studies standard deviation marginal posterior distributions

Subgroup analyses

We investigated the effects of 10 moderators on flipped learning classrooms to explain the effect-size variability. For each moderator, effect sizes were grouped into pre-defined levels. Quantitative results are depicted in Appendix A, Tables A5 and A6. Figures 4 to 8 depict marginal posterior densities for each moderator. Referring to Q_B in Appendix A, Table A5, intervention duration, the language learned, language outcome, allocation, and school location, were significantly related to language learning outcome variability.

Intervention duration

We divided studies into two intervention duration groups: more than 10 weeks ($K_{>10weeks}$ = 13) or less than 10 weeks ($K_{1-10weeks}$ = 13). For both duration groups, participants in flipped learning classrooms significantly outperformed their counterparts in the non-flipped condition. Moreover, studies with less than 10-week intervention ($\hat{d}_{1-10weeks}$ = 1.252, HPDI = [0.954, 1.561]) had larger mean effect than those with over 10-week intervention ($\hat{d}_{>10weeks}$ = 0.932, HPDI = [0.645, 1.223]). Variability of effects in the two groups were quite similar, that is, $\hat{\tau}_{1-10weeks}$ = 0.416 (HPDI = [0.108, 0.754]) and $\hat{\tau}_{>10weeks}$ = 0.392 (HPDI = [0.132, 0.686]). The marginal posterior densities shape (Figure 4) indicated slight differences.

Language learned

In terms of language learned, we focused on English ($K_{English}$ = 22) and Chinese ($K_{Chinese}$ = 2). The mean effect for the English group was larger than that for the Chinese group, with $\hat{d}_{English}$ = 1.183 (HPDI = [0.960, 1.410]) and $\hat{d}_{Chinese}$ = 0.803 (HPDI = [-0.020, 1.605]), respectively. Although both groups showed a positive mean effect favouring flipped learning classrooms, the HPDI interval for the Chinese group included zero and was wider, potentially resulting from a low sample size. Compared to the Chinese group ($\hat{\tau}_{Chinese}$ = 0.334, HPDI = [0, 1.117]), the between-studies variability for the English group ($\hat{\tau}_{English}$ = 0.417, HPDI = [0.221, 0.631]) was slightly larger. Additionally, the Chinese group had HPDIs starting with zero (Figure 5).

Language outcomes

We partitioned the language outcome moderator into multiple ($K_{multiple}$ = 7), writing ($K_{writing}$ = 8), speaking ($K_{speaking}$ = 8), and vocabulary/grammar ($K_{vocabulary/grammar}$ = 2). The posterior mean effect for writing was the largest with $\hat{d}_{writing}$ = 1.361 (HPDI = [0.897, 1.847]) , followed by speaking with $\hat{d}_{speaking}$ = 1.043 (HPDI = [0.686, 1.412]) , multiple with $\hat{d}_{multiple}$ = 0.866 (HPDI = [0.546, 1.177]) , and vocabulary/grammar with $\hat{d}_{vocabulary/grammar}$ = 0.806 (HPDI = [-0.144, 1.615]). Between-studies variability demonstrated the largest variability for writing ($\hat{\tau}_{writing}$ = 0.534 (HPDI = [0.160, 0.987])), followed by vocabulary/grammar ($\hat{\tau}_{vocabulary/grammar}$ = 0.391 (HPDI = [0, 1.285])), speaking ($\hat{\tau}_{speaking}$ = 0.336 (HPDI = [0, 0.715]), and multiple ($\hat{\tau}_{multiple}$ = 0.250 (HPDI = [0, 0.546]). Except for writing, the HPDI intervals for the other three outcome groups included zero. Furthermore, the writing group was comparatively more normally distributed than the other groups (Figure 6).

School location

The school location moderator was divided into Asia (K_{Asia} = 14), Europe (K_{Europe} = 7), and North America ($K_{North\ America}$ = 2). Q_B was significant for the school location. The European group has the largest mean effect of the marginal posterior distribution ($\hat{d}_{Europe}$ = 1.415 (HPDI = [0.937, 1.925])), followed by the Asian group ($\hat{d}_{Asia}$ = 1.033 (HPDI = [0.776, 1.293])) and the North American group ($\hat{d}_{North\ America}$ = 0.359 (HPDI = [-0.467, 1.188])). Although all three groups showed a positive mean effect favouring flipped learning classrooms, zero was included in the HPDI interval for the North American group, and its interval was wider, which may be caused by the small sample size. The between-studies variability for the European group was larger than others, with $\hat{\tau}_{Europe}$ = 0.489 (HPDI = [0, 0.941]) . Although variability for the Asian and North American groups showed many similarities, with $\hat{\tau}_{Asia}$ = 0.366 (HPDI = [0.128, 0.634]) and $\hat{\tau}_{North\ America}$ = 0.349 (HPDI = [0, 1.163]) , respectively, the North American group included zero in HPDIs (Figure 7).

Allocation

Considering the allocation of experimental and control groups, we divided studies into randomised pre-existing groups ($K_{randomised\ groups}$ = 13), randomised on an individual level ($K_{randomised\ individual}$ = 10), and non-random allocation ($K_{non\text{-}random}$ = 2). The posterior mean effect for the individual randomisation

($\widehat{d}_{\text{randomised individual}}$ = 1.247 (HPDI = [0.959, 1.532])) was slightly higher than that for the randomised pre-existing group allocation ($\widehat{d}_{\text{randomised groups}}$ = 1.036 (HPDI = [0.756, 1.323])). The variability for randomised pre-existing groups and randomised on individual level did not differ significantly with $\hat{\tau}_{\text{randomised groups}}$ = 0.375 (HPDI = [0.048, 0.688]) and $\hat{\tau}_{\text{randomised individual}}$ = 0.284 (HPDI = [0, 0.583]), respectively (Figure 8).

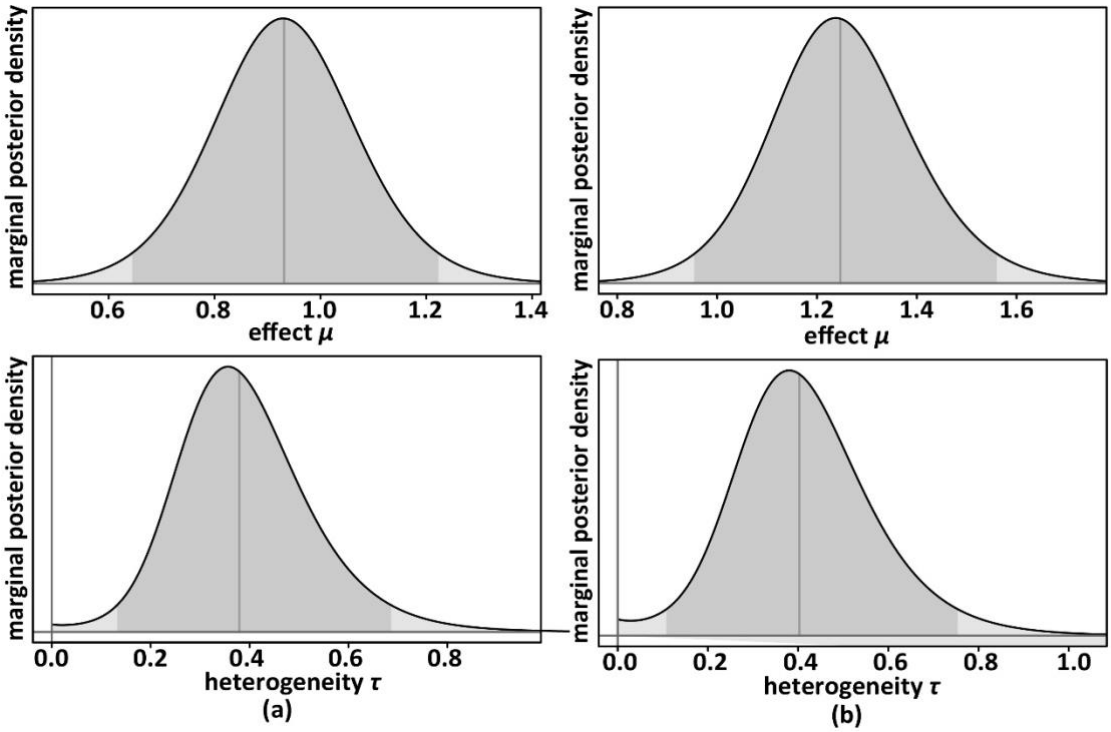

Figure 4. Mean and between-studies standard deviation marginal posterior distributions for intervention duration moderator with (a) > 10 weeks, and (b) 1 – 10 weeks

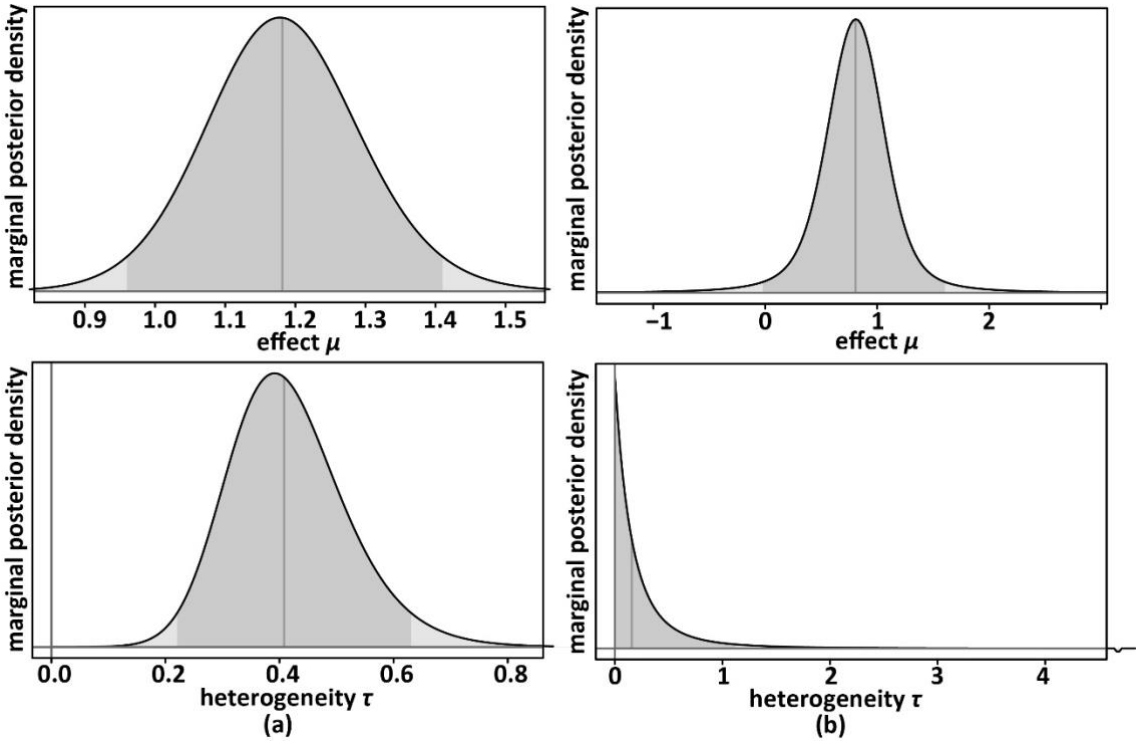

Figure 5. Mean and between-studies standard deviation marginal posterior distributions for learned language moderator with (a) English and (b) Chinese

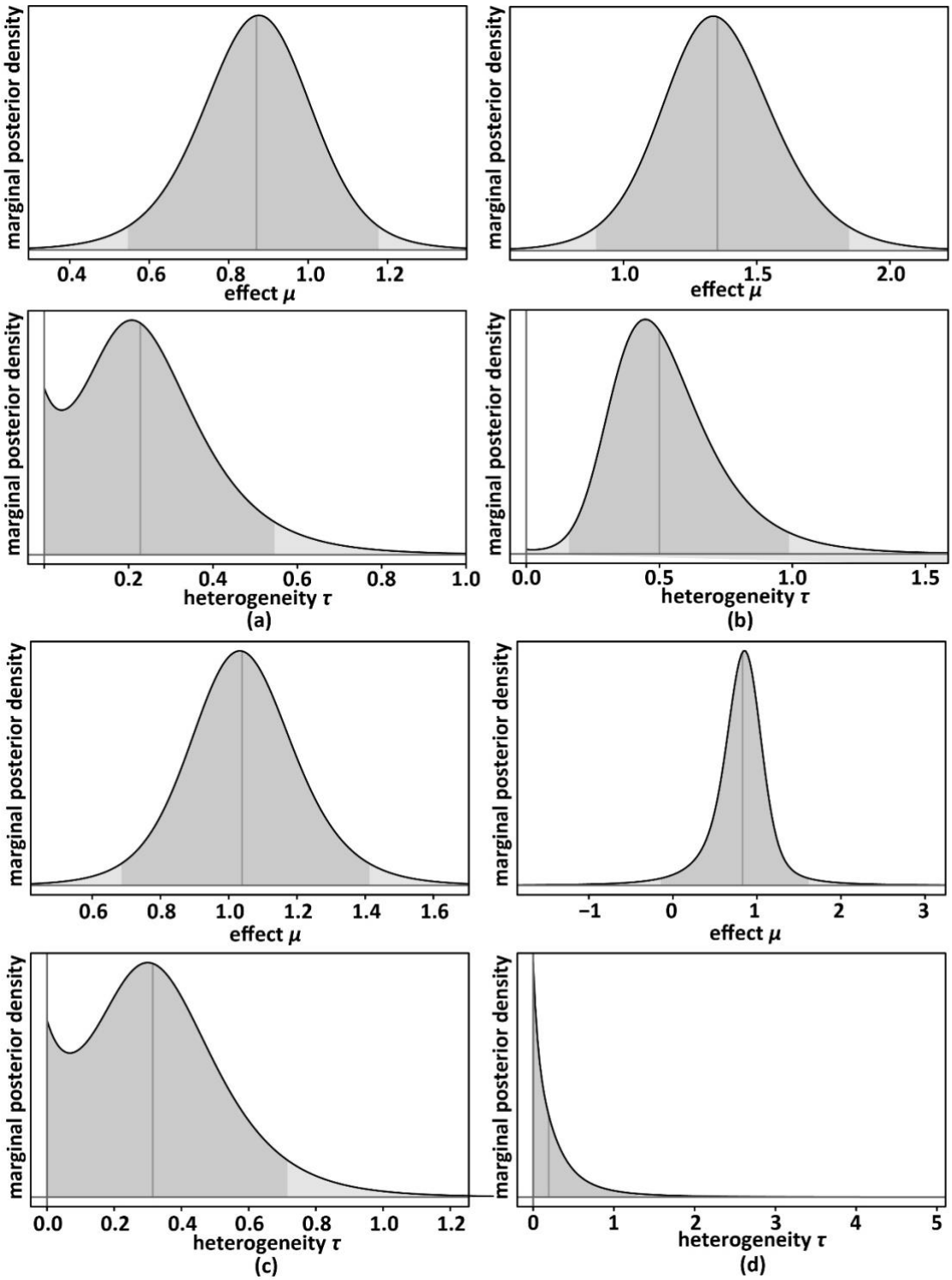

Figure 6. Mean and between-studies standard deviation marginal posterior distributions for language outcome moderator with (a) multiple, (b) writing, (c) speaking, and (d) vocabulary/grammar

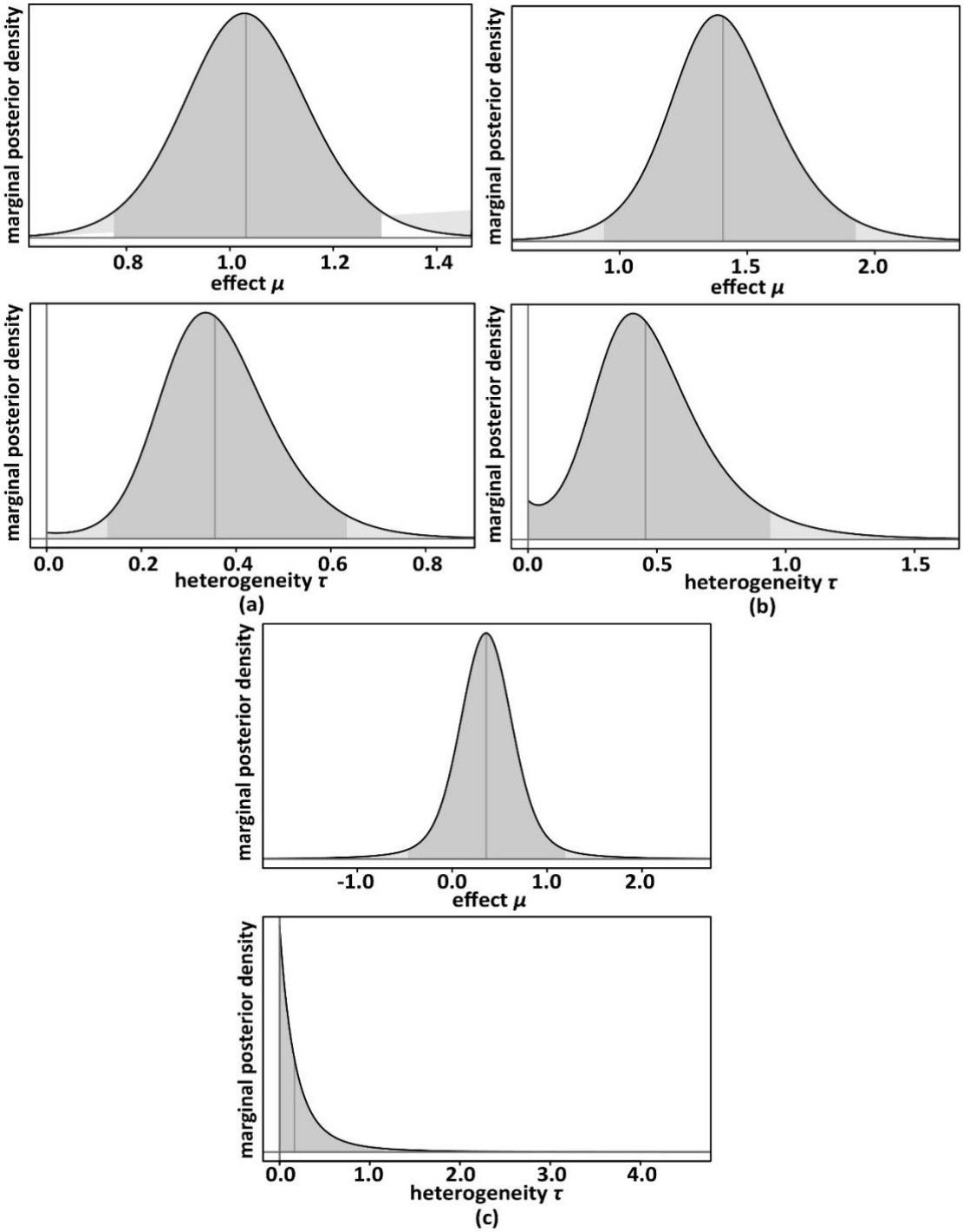

Figure 7. Mean and between-studies standard deviation marginal posterior distributions for school location moderator with (a) Asia, (b) Europe, and (c) North America

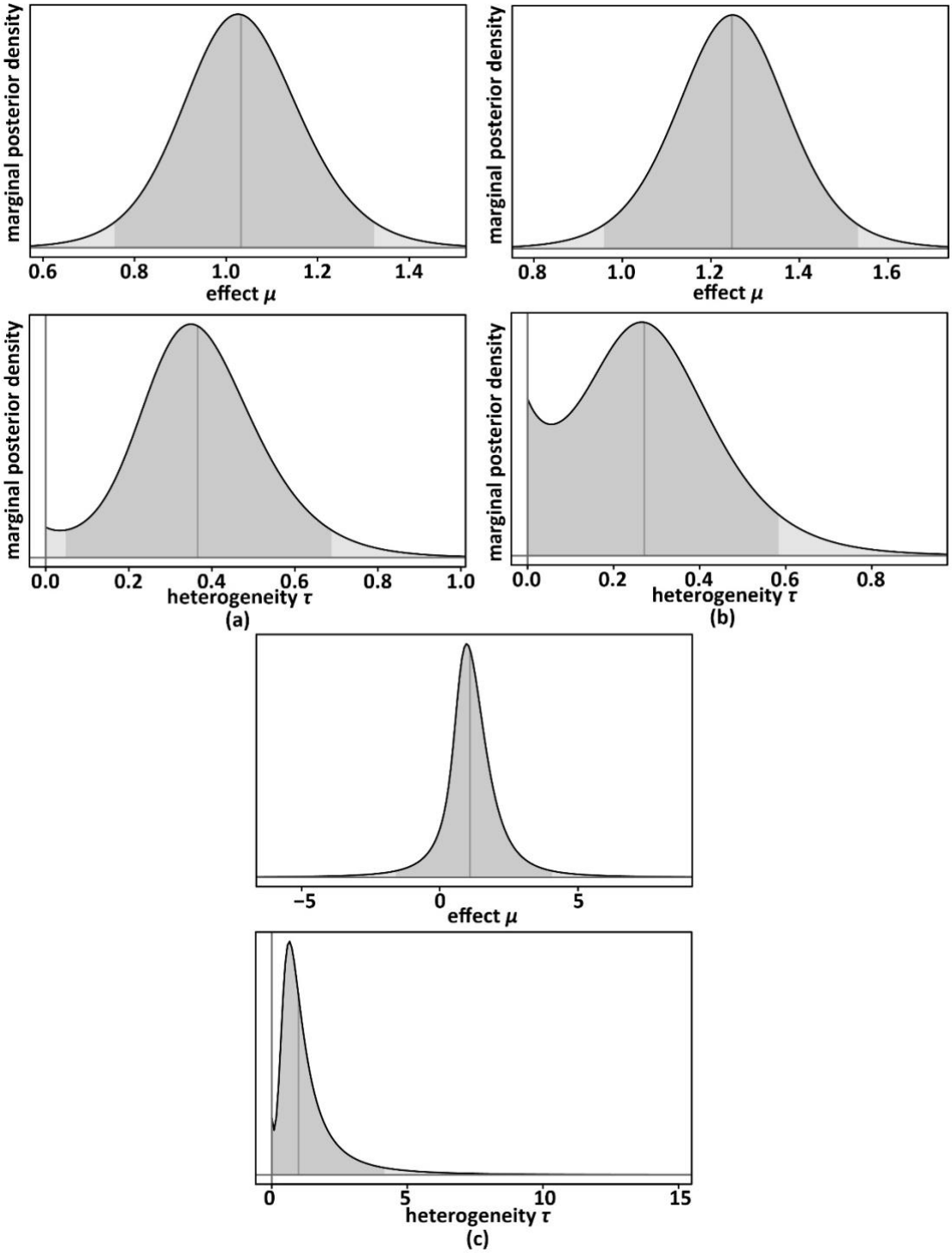

Figure 8. Mean and between-studies standard deviation marginal posterior distributions for allocation moderator with (a) randomised pre-existing groups, (b) randomised on the individual level, and (c) no randomisation.

Other moderators

In addition to moderators that showed a significant relation to the variability of language learning outcomes, other moderators' effect sizes, including publications, instructor, face-to-face time, group equivalence test, and quizzes, are elaborated as follows. Considering the publication type, there were two types of publications involved in the 26 studies, that is, journal article ($K_{\text{journal article}} = 23$) and doctoral

dissertation ($K_{\text{doctoral dissertation}}$ = 3). We found a larger posterior mean effect in the journal article group ($\hat{d}_{\text{journal article}}$ = 1.146, HPDI = [0.929, 1.368]) than in the doctoral dissertation group with $\hat{d}_{\text{doctoral dissertation}}$ = 0.695 (HPDI = [-0.129, 1.498]).

According to the teacher moderator, studies were divided into three types, including the same instructors for the experimental and control groups ($K_{\text{same instructor}}$ = 19), different instructors ($K_{\text{different instructors}}$ = 4), or unspecified ($K_{\text{unspecified}}$ = 3). Similarities were shown in the mean effects for groups with the same and different instructors, with $\hat{d}_{\text{same instructor}}$ = 1.047 (HPDI = [0.784, 1.313]) and $\hat{d}_{\text{different instructors}}$ = 1.159 (HPDI = [0.517, 1.850]), respectively. Though the two groups demonstrated a moderate mean effect favouring flipped learning classrooms, the HPDI interval of the different instructor group was wider than that of the same instructor group. This difference was a possible result of the small sample size of $K_{\text{different instructors}}$ = 4.

Considering the moderator of face-to-face time, there were six studies where face-to-face time was less in flipped learning classrooms than that in traditional classrooms ($K_{\text{FC < TC}}$ = 6), 17 studies where face-to-face time was equal to that in traditional classrooms ($K_{\text{FC = TC}}$ = 17), and three studies without specification about face-to-face time ($K_{\text{unspecified}}$ =3). The results showed that the studies shortening face-to-face time flipped learning classrooms had a smaller average effect than those offering equal face-to-face time to flipped learning classrooms and traditional classrooms, with $\hat{d}_{\text{FC < TC}}$ = 0.874 (HPDI = [0.364, 1.385]) and $\hat{d}_{\text{FC = TC}}$ = 1.174 (HPDI = [0.897, 1.457]), respectively.

Considering the equivalence test between intervention and control groups, we coded the studies as not tested, descriptive statement ($K_{\text{descriptive statement}}$ = 8), tested, equal ($K_{\text{tested, equal}}$ = 12), not tested, no descriptive statement (K_{No} = 5), and tested, not equal ($K_{\text{tested, not equal}}$ = 1). The mean effects for the tested, equal group was slightly larger than that for the descriptive statement group, with $\hat{d}_{\text{tested, equal}}$ = 1.100 (HPDI = [0.733, 1.477]) and $\hat{d}_{\text{descriptive statement}}$ = 1.084 (HPDI = [0.661, 1.509]), respectively.

Lastly, we categorised the studies according to the use of quizzes in flipped conditions. In the context of flipped classrooms, the mean effects showed that the studies using quizzes had a slightly larger average effect than studies without quizzes, with $\hat{d}_{\text{quizzes in FC}}$ = 1.108 (HPDI = [0.830, 1.389]) and $\hat{d}_{\text{no quizzes in FC}}$ = 1.061 (HPDI = [0.745, 1.388]), respectively.

Publication bias

There was a low risk of bias. Although the funnel plot (Figure 9) showed a certainly visible asymmetry, it was not substantial enough to draw a large publication bias conclusion. Furthermore, zero imputed effect was found through the Trim-and-Fill test, indicating no extra imputed effect in the funnel plot. Begg's rank correlation test (Kendall's tau = 0.1541, p = 0.270), Egger's regression test (Z = 1.310, p = 0.190), and Vevea and Hedges likelihood ratio test ($\chi^2(1)$ =0 .003, p = 0.956) were not statistically significant.

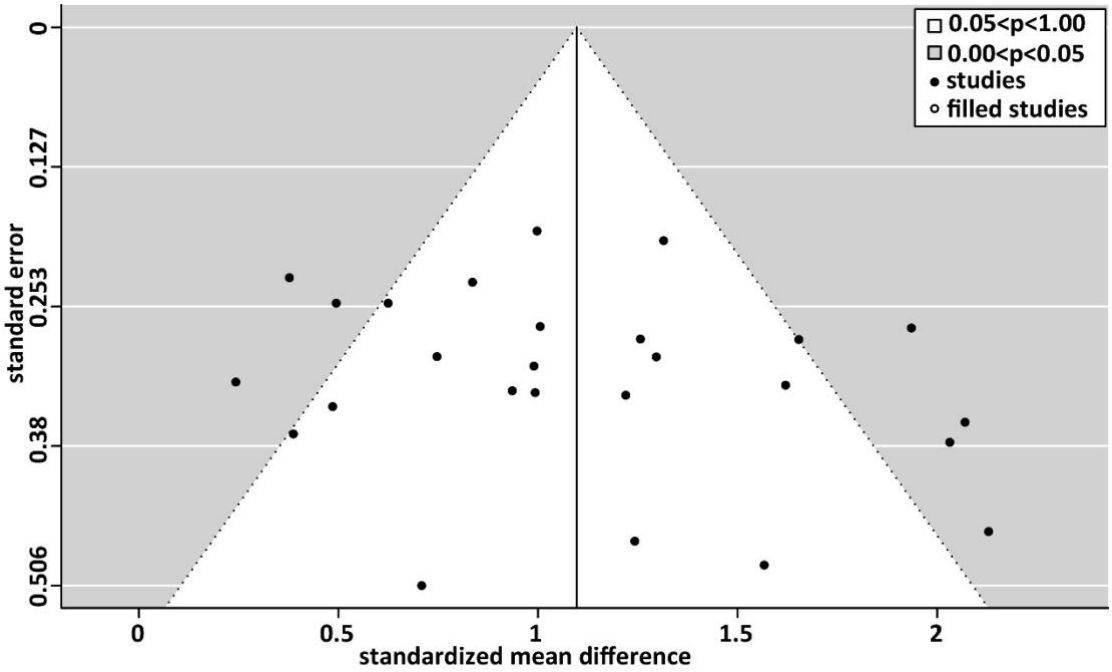

Figure 9. Funnel plot of effect sizes (95% CI = confidence interval)

Discussion

Overall and sub-group analyses

The Bayesian meta-analysis of 26 flipped learning classroom studies reported a mean effect size of 1.096 (95% CI as [0.887, 1.307]), indicating a positive and large effect of flipped-based instruction on language learning outcomes in general. As this meta-analysis was conducted by comparing flipped learning classroom groups with control groups, the above findings indicated that students participating in flipped learning classrooms outperformed non-flipped classrooms in higher education. Subgroup analyses presented moderators' effects on effect sizes. The first consideration was the intervention duration. Studies with a longer flipped learning classroom intervention duration (more than 10 weeks) showed smaller effect sizes than those with a shorter duration (less than ten weeks). This finding was consistent with prior meta-analyses on flipped classrooms (e.g., Cheng et al., 2019), implying that flipped learning classrooms' effectiveness might be associated with learners' curiosity. That is, learners had a higher curiosity about the first-time flipped instruction and thus a greater learning motivation. However, the motivation tended to decrease as time went by, thereby lowering the effectiveness of flipped instruction. In practice, flipped learning classrooms' effectiveness depends not only on curiosity but also on practical design. Thus, future research should consider both learners' psychological states and flipped learning classrooms' effective design.

The language learning outcome was the second moderating factor. Generally, compared to non-flipped instructions, flipped instructions are effective in promoting various types of language skills. However, flipped instruction was particularly effective in promoting writing and speaking, while its effectiveness for vocabulary/grammar learning was less significant. However, as only two studies focused on vocabulary/grammar learning, such a conclusion might be less convincing. There is also a call for further research on how and to what extent different types of language skills benefit from flipped instructions.

When considering school locations, there was a smaller effect size for flipped learning classroom studies implemented in North America than in Asia or Europe. This might be explained by the larger effect size of English learning and its prevalence in Asian and European studies. However, such results may be restricted by the small sample size of North American studies. Additionally, European flipped learning classroom studies presented more significant effects than the effects in Asia due to the following reasons. First,

compared with Asian students, students in Europe experienced less pressure or school competition, and teachers had more time to design and implement flipped classrooms (Chen & Yang, 2019). Second, English learning is prevalent in Asian and European flipped learning classroom studies, which brings another explanation for the larger effect sizes of European studies. Specifically, European students were more motivated and active in learning English as a prerequisite for future development (e.g., joining international business or European Union institutions). Moreover, European schools emphasise communication, which benefits language learning, and non-native European learners are more fluent and confident in speaking compared to Asian learners. Also, similarities between English and most European languages (e.g., French and German) perhaps help European students learn well. Thus, teachers and researchers are advised to motivate students by tailoring learning goals and activities and adjusting teaching guidance according to student backgrounds and target language features.

Our study found that the additional quiz in flipped learning classrooms was not a significant moderator. A possible explanation was that we did not differentiate different types of quizzes (i.e., in-class or take-home) because of the limited descriptions of quizzes in the studies. According to Christiansen et al. (2017), students who finished open-book take-home quizzes without time limitation obtained lower grades in the final tests than those who finished in-class quizzes. Additionally, students were reported to prefer in-class quizzes to take-home ones, and take-home quizzes demotivated their attendance of pre-class video watching. Thus, we suggest flipped learning classroom researchers provide details about quizzes used in their experiments to allow the exploration of their effects on learning outcomes in flipped learning classrooms. Additionally, although quizzes were not a significant moderator, this had no negative moderating effect, indicating that the addition of quizzes in flipped learning classrooms did not hinder learning.

When it came to face-to-face time, participants in flipped learning classrooms with equal face-to-face time with control groups outperformed participants in flipped learning classrooms with reduced face-to-face time. Such a finding contradicted the conclusion by Baepler et al. (2014) that less face-to-face time in flipped classrooms showed at least equal effectiveness to conventional classrooms. Nevertheless, our results aligned with some other studies. For instance, as van Alten et al. (2019) suggest, participants in flipped classrooms without cutting down the face-to-face time significantly outperformed those with less time. In Heyma et al. (2015), participants in flipped classrooms with reduced face-to-face time did not perform better, which was probably due to the inability of participants to take responsibility for self-regulated learning. We thus highlight sustaining face-to-face time as an important factor for flipped learning classrooms' success. Though it is commonly argued that the implementation of flipped classrooms is to reduce face-to-face teaching time (Asef‐Vaziri, 2015), we suggest exploring this idea with caution. Teachers are not encouraged to reduce face-to-face time intentionally during flipped learning classroom implementation.

In terms of allocation, previous studies (e.g., Sala & Gobet, 2017) indicated smaller effect sizes in random allocation. However, in our results, with only one non-randomised allocation study, such a conclusion was not convincing. Among the 23 flipped learning classroom studies with randomised allocation, participants showed improvement in language learning outcomes. Participants randomised on individual levels showed slightly higher performance in language outcomes than those randomised based on pre-existing groups. A possible explanation is that compared to randomisation on pre-existing groups, randomisation on individual levels ensures more comparability of the experimental and control groups. Another explanation may be that due to familiarity, pre-existing groups where participants are more likely to have off-task behaviours may focus less on task competition as compared to randomisation on individual levels where participants commonly do not know each other very well and can focus more on task competition (Mozaffari, 2017).

Lastly, the group equivalence test was examined. In practice, the statistical proficiency test is regarded as the prerequisite of an empirical study because initial proficiency could influence post-test scores (Leis et al., 2015). However, our findings did not support this suggestion, as we found that group equivalence was not significantly related to language learning outcome variability. A possible explanation is some

difficulties in correctly assigning studies into the four categories (not tested/descriptive statement, tested/equal, not tested/no descriptive statement, and tested/not equal) due to the insufficient or inconsistent description in the studies analysed. For example, some studies (e.g., Haghighi et al., 2019) were coded as not tested due to the lack of statistical evidence, however, the scores in the pre-test for the experiment and control groups were almost equal. Thus, we suggest that future research provides explicit details on their tests to allow exploration of the effects of group equivalence approaches on language learning outcomes.

Pedagogical implications and suggestions for future flipped language classroom research

Our findings provide implications for language instructors who have flipped their classes or are thinking of doing so. These implications are listed below:

1. Instructors are suggested to flipped writing and speaking courses with enough face-to-face time and necessary technical support being provided to students to maximise flipped learning classroom effectiveness.
2. Instructors need to consider time variance's effects on learning performance and seek ways to maintain learners' interest in long-term interventions.
3. Instructors should gather information about how to flip a foreign language classroom and pre-train learners of flipped learning before starting flipped learning classroom implementation.
4. During flipped course design, instructors are suggested to listen to students' opinions, predict possible difficulties, and prepare solutions accordingly.
5. Instructors should include practices and quizzes in pre-class activities to constantly check learners' understanding of learning contents, with asynchronous online interaction tools being used for promoting interaction and feedback provision.
6. Instructors can scaffold lower-achievers' learning by providing pre-class videos in first languages at the beginning of flipped learning classroom implementation and using second-language videos step by step.

More specific descriptions of these implications are as followings.

From the perspective of the flipped learning classroom application, suggestions are as follows. Firstly, the larger effect size of writing and speaking outcomes indicated the significance of integrating flipped learning classrooms into the language curriculum to facilitate second/foreign language writing and speaking. Secondly, considering that participants in flipped learning classrooms with equal face-to-face classroom time significantly outperformed their counterparts with reduced face-to-face time, we suggest teachers provide enough face-to-face time to maximise the effectiveness of flipped learning classrooms. Furthermore, the short-time treatment duration (5 to 10 weeks) generated a larger effect size in flipped learning classrooms. Thus, for long-term intervention practice, teachers should focus on ways to maintain flipped learning classrooms' effectiveness through flipped design adjustment. Accordingly, further investigation into time variance's effects on learning performance is needed.

In the digital era, teacher roles have evolved from information providers to guiders, mentors, and facilitators. Accordingly, teachers' responsibilities shift to adaptation assistance and awareness-raising (Khvilon & Patru, 2002). Thus, teachers need to learn technical knowledge about technology integration into teaching (Adnan, 2017; Bezzazi, 2019). They should also give students pre-training about flipped learning to prepare students before flipped learning classroom implementation and show them the underlying flipped learning classroom benefits to increase their motivation (Chen Hsieh et al., 2017). Additionally, technological support for students is needed to resolve unpredicted difficulties during instructions (Yang et al., 2018).

In terms of course design, teachers are suggested to flip courses that can benefit most from flipping by identifying modules where "online instruction would help to save class time for the application of skills gained after instruction" (Webb & Doman, 2016, p. 57). For example, teachers and curriculum designers may consider flipping basic language learning content and teaching complex grammatical structures face-

to-face (Hojnacki, 2018). To successfully implement flipped learning classrooms, teachers should consider teaching plans, learning materials, and evaluation criteria beforehand (Lee & Wallace, 2018).. In addition to teachers' preparation, student opinions are important in the flipped course design. For example, teachers could collect student ideas when preparing videos and texts (Yesilçinar, 2019) and even give students opportunities to create videos based on learning objectives and interactive principles (Hojnacki, 2018; Oraif, 2018). Furthermore, teachers are recommended to provide videos from sources that students trust to motivate them to watch and cultivate their sense of responsibility (Oraif, 2018). Teachers should also be able to predict possible difficulties that students may encounter during flipped learning and provide solutions accordingly (Yang et al., 2018). Teachers should also check learners' understanding of learning contents constantly since this understanding may affect their determination of video-watching (Oraif, 2018). In class, teachers could design activities to engage students to participate in group discussions, conduct gamified activities, raise questions, give prompt feedback, and address any misconceptions (Adnan, 2017; Bezzazi, 2019).

To enhance interaction, teachers can provide an asynchronous online forum where learners can exchange learning experiences, which helps to develop their deduction and induction abilities and critical thinking skills and facilitate instructor-learner and learner-learner interactions in an online community (Lin et al., 2018). Additionally, interaction-based activities and materials should be provided to learners to help them receive feedback outside the classroom (Yesilçinar, 2019).

To facilitate learners' understanding of flipped learning, teachers could adjust their instruction language to follow standards or traditions within a given context to assist students' concept learning (Lin et al., 2018). Weekly practices and quizzes, as formative assessments, are necessary for students to review the newly learned knowledge and deepen their understanding of unfamiliar knowledge (Webb & Doman, 2016). For lower-achievers, teachers could record lectures in students' first languages in the beginning stage. When students become familiar with flipped learning classrooms, learning videos can be shot in the target language with subtitles in the students' first language (Lee & Wallace, 2018). Lastly, teachers should always adapt lessons based on student responses and reactions and provide scaffolding when necessary (Chen Hsieh et al., 2017).

Potential directions for future flipped learning classroom research were included in our analyses, including:

(1) How learners with different proficiency levels or personal characteristics benefit from flipped learning classrooms (Yesilçinar, 2019).
(2) How flipped classrooms motivate students with low learning motivation (Leis et al., 2015).
(3) How to use different flipped pedagogical approaches (e.g., planning, reflecting, and outside-class and in-class activities) to promote learners' engagement in and perceptions about Fflipped learning classrooms (Lin & Hwang, 2018).
(4) Seamless combination of classroom activities with online activities and learners' self-learning ability development in flipped learning classrooms (Adnan, 2017).
(5) Effects of interactive behaviours on learners' academic performance with varied flipped strategies (Lin & Hwang, 2018).
(6) Integrating cutting-edge technologies into flipped learning classrooms, for example, mobile-based contextual games for flipped learning (Lin et al., 2018).
(7) Flipped learning classrooms' effects on collaborative/communicative relations between learners and instructors (Lin et al., 2018).
(8) Using equation modelling to analyse the direct/indirect effects of intermediate flipped variables on language learning outcomes and to identify relationships between variables through a mediation effect (Oraif, 2018).
(9) Flipped learning classrooms' long-term effects (Salem, 2018).
(10) (10) Using between and within-subject research design by switching both flipped and traditional instructions (several times if possible) between independent groups of learners.
(11) The costs (e.g., time and technological cost) of flipped learning classroom implementation.

Conclusion

This study quantitatively examined 26 studies that investigated flipped learning classrooms' effects on learning outcomes in higher education. The overall analysis indicated that students in flipped learning classrooms performed significantly better than those in conventional classrooms regarding learning outcomes, particularly in writing and speaking. Results of subgroup analyses and the examination of each study content provided insights into the effective design of flipped learning classrooms, including: (1) sufficient face-to-face time to maximise the effectiveness of flipped learning classrooms; (2) instructors' focus on ways to improve flipped learning classrooms' effectiveness through flipped design adjustment during long-term intervention; (3) giving students pre-training of flipped learning before flipped learning classroom implementation and showing them the underlying benefits to activate their motivation; (4) flipping basic contents of language learning and teaching complex contents face-to-face; (5) listening to students' opinions during flipped course design and adapting lessons based on student responses and reactions during flipped learning classroom implementation; (6) checking learners' understanding of learning contents from time to time during flipped learning, particularly by using in-class quizzes; and (7) adopting scaffolding strategies such as code-switching (interchangeably using learners' first and target languages) to scaffold lower achievers' learning in flipped learning classrooms.

References

Adnan, M. (2017). Perceptions of senior-year ELT students for flipped classroom: A materials development course. *Computer Assisted Language Learning*, *30*(3–4), 204–222. https://doi.org/10.1080/09588221.2017.1301958

Afrilyasanti, R., Cahyono, B. Y., & Astuti, U. P. (2017). Indonesian EFL students' perceptions on the implementation of flipped classroom model. *Journal of Language Teaching and Research*, *8*(3), 476–484. http://dx.doi.org/10.17507/jltr.0803.05

Akçayır, G., & Akçayır, M. (2018). The flipped classroom: A review of its advantages and challenges. *Computers & Education*, *126*, 334–345. https://doi.org/10.1016/j.compedu.2018.07.021

Amiryousefi, M. (2019). The incorporation of flipped learning into conventional classes to enhance EFL learners' L2 speaking, L2 listening, and engagement. *Innovation in Language Learning and Teaching*, *13*(2), 147–161. https://doi.org/10.1080/17501229.2017.1394307

Arslan, A. (2020). A systematic review on flipped learning in teaching English as a foreign or second language. *Dil ve Dilbilimi Çalışmaları Dergisi*, *16*(2), 775–797. https://doi.org/10.17263/jlls.759300

Asef-Vaziri, A. (2015). The flipped classroom of operations management: A not-for-cost-reduction platform. *Decision Sciences Journal of Innovative Education*, *13*(1), 71–89. https://doi.org/10.1111/dsji.12054

Baepler, P., Walker, J. D., & Driessen, M. (2014). It's not about seat time: Blending, flipping, and efficiency in active learning classrooms. *Computers & Education*, *78*, 227–236. https://doi.org/10.1016/j.compedu.2014.06.006

Baker, J. W. (2000). The classroom flip: Using web course management tools to become the guide by the side. *Paper presented at the 11th International Conference on College Teaching and Learning*. Jacksonville, FL. https://works.bepress.com/j_wesley_baker/15/

Basal, A. (2015). The implementation of a flipped classroom in foreign language teaching. *Turkish Online Journal of Distance Education*, *16*(4), 28–37. https://dergipark.org.tr/tr/pub/tojde/issue/16949/176933

Bergmann, J., & Sams, A. (2012). *Flip your classroom: Reach every student in every class every day*. International Society for Technology in Education. https://www.rcboe.org/cms/lib/GA01903614/Centricity/Domain/15451/Flip_Your_Classroom.pdf

Bezzazi, R. (2019). Learning English grammar through flipped learning. *The Asian Journal of Applied Linguistics*, *6*(2), 170–184. https://caes.hku.hk/ajal/index.php/ajal/article/view/597

Bicen, H., & Beheshti, M. (2022). Assessing perceptions and evaluating achievements of ESL students with the usage of infographics in a flipped classroom learning environment. *Interactive Learning Environments*, *30*(3), 498–526. https://doi.org/10.1080/10494820.2019.1666285

Chen, C.-H., & Yang, Y.-C. (2019). Revisiting the effects of project-based learning on students' academic achievement: A meta-analysis investigating moderators. *Educational Research Review*, *26*, 71–81. https://doi.org/10.1016/j.edurev.2018.11.001

Chen, K., Monrouxe, L., Lu, Y., Jenq, C., Chang, Y., Chang, Y., & Chai, P. Y. (2018). Academic outcomes of flipped classroom learning: A meta‐analysis. *Medical Education*, *52*(9), 910–924. https://doi.org/10.1111/medu.13616

Chen, X., Zou, D., Xie, H., & Cheng, G. (2021). Twenty years of personalized language learning: Topic modeling and knowledge mapping. *Educational Technology & Society*, *24*(1), 205-222. https://www.jstor.org/stable/26977868

Chen, X., Zou, D., Xie, H., & Cheng, G. (2022). Socialization in language learning: Topic modeling and bibliometric analysis. In J. Colpaert, & G. Stockwell (Eds.), *Smart CALL: Personalization, contextualization, & socialization* (pp. 151-183). Castledown Publishers. https://doi.org/10.29140/9781914291012-8

Chen, X., Zou, D., Xie, H., & Su, F. (2021). Twenty-five years of computer-assisted language learning: A topic modeling analysis. *Language Learning & Technology*, *25*(3), 151-185. http://hdl.handle.net/10125/73454

Cheng, L., Ritzhaupt, A. D., & Antonenko, P. (2019). Effects of the flipped classroom instructional strategy on students' learning outcomes: A meta-analysis. *Educational Technology Research and Development*, *67*(4), 793–824. https://doi.org/10.1007/s11423-018-9633-7

Chen Hsieh, J. S., Wu, W.-C. V., & Marek, M. W. (2017). Using the flipped classroom to enhance EFL learning. *Computer Assisted Language Learning*, *30*(1–2), 1–21. https://doi.org/10.1080/09588221.2015.1111910

Christiansen, M. A., Lambert, A. M., Nadelson, L. S., Dupree, K. M., & Kingsford, T. A. (2017). In-class versus at-home quizzes: Which is better? A flipped learning study in a two-site synchronously broadcast organic chemistry course. *Journal of Chemical Education*, *94*(2), 157–163. https://doi.org/10.1021/acs.jchemed.6b00370

DuMouchel, W. (1994) Hierarchical Bayes linear models for meta-analysis. *Technical Report 27*. National Institute of Statistical Sciences. https://people.eecs.berkeley.edu/~russell/classes/cs294/f05/papers/dumouchel-1994.pdf

Duval, S., & Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*(449), 89–98. https://doi.org/10.1080/01621459.2000.10473905

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel‐plot–based method of testing and adjusting for publication bias in meta‐analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Haghighi, H., Jafarigohar, M., Khoshsima, H., & Vahdany, F. (2019). Impact of flipped classroom on EFL learners' appropriate use of refusal: achievement, participation, perception. *Computer Assisted Language Learning*, *32*(3), 261–293. https://doi.org/10.1080/09588221.2018.1504083

Hamdani, M. (2019). Effectiveness of flipped classroom (FC) method on the development of English language learning of the high school students in Ahwaz. *International Journal of Applied Linguistics and English Literature*, *8*(2), 12–20. http://journals.aiac.org.au/index.php/IJALEL/article/view/5600

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, *92*(2), 490–499. https://doi.org/10.1037/0033-2909.92.2.490

Hew, K. F., & Lo, C. K. (2018). Flipped classroom improves student learning in health professions education: A meta-analysis. *BMC Medical Education*, *18*(1), 1–12. https://doi.org/10.1186/s12909-018-1144-z

Heyma, A., Bisschop, P., van den Berg, E., Wartenbergh-Cras, F., Kurver, B., Muskens, M., & Spanjers, I. (2015). *Effectmeting InnovatieImpuls Onderwijs: eindrapport*. SEO Economisch Onderzoek. https://pure.uva.nl/ws/files/2673585/172138_514504.pdf.

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

Hojnacki, S. (2018). *The flipped classroom in introductory foreign language classes*. Michigan State University. https://www.proquest.com/openview/994f3b2db287a41b19b2d69418736e8f/1?pq-origsite=gscholar&cbl=18750

Hung, H.-T. (2015). Flipping the classroom for English language learners to foster active learning. *Computer Assisted Language Learning*, *28*(1), 81–96. https://doi.org/10.1080/09588221.2014.967701

Hung, H.-T. (2017). Design-based research: Redesign of an English language course using a flipped classroom approach. *Tesol Quarterly*, *51*(1), 180–192. https://doi.org/10.1002/tesq.328

Jiang, M. Y., Jong, M. S., Lau, W. W., Chai, C., Liu, K. S., & Park, M. (2022). A scoping review on flipped classroom approach in language education: challenges, implications and an interaction model. *Computer Assisted Language Learning*, *35*(5–6), 1218–1249. https://doi.org/10.1080/09588221.2020.1789171

Khvilon, E., & Patru, M. (2002). *Information and communication technologies in teacher education: a planning guide*. UNESCO. https://tinyurl.com/y4hmmqre

Kirmizi, Ö., & Kömeç, F. (2019). The impact of the flipped classroom on receptive and productive vocabulary learning. *Dil ve Dilbilimi Çalışmaları Dergisi*, *15*(2), 437–449. https://doi.org/10.17263/jlls.586096

Låg, T., & Sæle, R. G. (2019). Does the flipped classroom improve student learning and satisfaction? A systematic review and meta-analysis. *American Educational Research Association Open*, *5*(3), 1–17. https://doi.org/10.1177/2332858419870489.

Lee, G., & Wallace, A. (2018). Flipped learning in the English as a foreign language classroom: Outcomes and perceptions. *Tesol Quarterly*, *52*(1), 62–84. https://doi.org/10.1002/tesq.372

Leis, A., Cooke, S., & Tohei, A. (2015). The effects of flipped classrooms on English composition writing in an EFL environment. *International Journal of Computer-Assisted Language Learning and Teaching*, *5*(4), 37–51. https://doi.org/10.4018/IJCALLT.2015100103

Lin, C.-J., & Hwang, G.-J. (2018). A learning analytics approach to investigating factors affecting EFL students' oral performance in a flipped classroom. *Journal of Educational Technology & Society*, *21*(2), 205–219. https://www.jstor.org/stable/26388398

Lin, C.-J., Hwang, G.-J., Fu, Q.-K., & Chen, J.-F. (2018). A flipped contextual game-based learning approach to enhancing EFL students' English business writing performance and reflective behaviors. *Journal of Educational Technology & Society*, *21*(3), 117–131. https://www.jstor.org/stable/26458512

Lin, J.-J., & Lin, H. (2019). Mobile-assisted ESL/EFL vocabulary learning: A systematic review and meta-analysis. *Computer Assisted Language Learning*, *32*(8), 878–919. https://doi.org/10.1080/09588221.2018.1541359

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Sage Publications.

Lo, C. K., & Hew, K. F. (2017). A critical review of flipped classroom challenges in K-12 education: Possible solutions and recommendations for future research. *Research and Practice in Technology Enhanced Learning*, *12*(1), 1–22. https://doi.org/10.1186/s41039-016-0044-2

Lo, C. K., & Hew, K. F. (2019). The impact of flipped classrooms on student achievement in engineering education: A meta-analysis of 10 years of research. *Journal of Engineering Education*, *108*(4), 523–546. https://doi.org/10.1002/jee.20293

Lo, C. K., Hew, K. F., & Chen, G. (2017). Toward a set of design principles for mathematics flipped classrooms: A synthesis of research in mathematics education. *Educational Research Review*, *22*, 50–73. https://doi.org/10.1016/j.edurev.2017.08.002

Lundin, M., Rensfeldt, A. B., Hillman, T., Lantz-Andersson, A., & Peterson, L. (2018). Higher education dominance and siloed knowledge: A systematic review of flipped classroom research. *International Journal of Educational Technology in Higher Education*, *15*(1), 1–30. https://doi.org/10.1186/s41239-018-0101-6

Mozaffari, S. H. (2017). Comparing student-selected and teacher-assigned pairs on collaborative writing. *Language Teaching Research*, *21* (4): 496–516. https://doi.org/10.1177/1362168816641703

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417–528. https://doi.org/10.1111/0023-8333.00136

Nwosisi, C., Ferreira, A., Rosenberg, W., & Walsh, K. (2016). A study of the flipped classroom and its effectiveness in flipping thirty percent of the course content. *International Journal of Information and Education Technology*, *6*(5), 348–351. https://doi.org/10.7763/IJIET.2016.V6.712

Oraif, I. M. K. (2018). *An investigation into the impact of the flipped classroom on intrinsic motivation (IM) and learning outcomes on an EFL writing course at a university in Saudi Arabia based on self-determination theory (SDT)* (Doctoral dissertation), The University of Leicester, England. https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.745826

Özkurkudis, M. J., & Bümen, N. T. (2019). Flipping the writing classroom: Using grammar videos to enhance writing. *Journal of Education and Future*, *15*, 1–16. https://doi.org/10.30786/jef.425632

Paiva, C. E., Araujo, R. L., Paiva, B. S. R., de Pádua Souza, C., Cárcano, F. M., Costa, M. M., Serrano, S. V., & Lima, J. P. N. (2017). What are the personal and professional characteristics that distinguish the researchers who publish in high-and low-impact journals? A multi-national web-based survey. *Ecancermedicalscience, 11*, 718–735. https://doi.org/10.3332/ecancer.2017.718

Rodrigues, R., Silva, J., Ramos, J. L. C., de Souza, F. da F., & Gomes, A. S. (2016). Uma Abordagem de Regressão Múltipla para validação de variáveis de Autorregulação da aprendizagem em ambientes de LMS. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática Na Educação-SBIE)*, *27*(1), 916–925. https://doi.org/10.5753/cbie.sbie.2016.916

Royall, R. M. (1970). Finite population sampling - On labels in estimation. *The Annals of Mathematical Statistics*, *41*(5), 1774–1779. https://www.jstor.org/stable/2239887

Sala, G., & Gobet, F. (2017). Does far transfer exist? Negative evidence from chess, music, and working memory training. *Current Directions in Psychological Science*, *26*(6), 515–520. https://journals.sagepub.com/doi/pdf/10.1177/0963721417712760Salem, A. A. (2018). Engaging ESP university students in flipped classrooms for developing functional writing skills, HOTs, and eliminating writer's block. *English Language Teaching*, *11*(12), 177–198. https://doi.org/10.5539/elt.v11n12p177

Shi, Y., Ma, Y., MacLeod, J., & Yang, H. H. (2020). College students' cognitive learning outcomes in flipped classroom instruction: A meta-analysis of the empirical literature. *Journal of Computers in Education*, *7*(1), 79–103. https://doi.org/10.1007/s40692-019-00142-8

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

Tan, C., Yue, W.-G., & Fu, Y. (2017). Effectiveness of flipped classrooms in nursing education: Systematic review and meta-analysis. *Chinese Nursing Research*, *4*(4), 192–200. https://doi.org/10.1016/j.cnre.2017.10.006

Thaichay, T., & Sitthitikul, P. (2016). Effects of the flipped classroom instruction on language accuracy and learning environment: A case study of Thai EFL upper-secondary school students. *Rangsit Journal of Educational Studies*, *3*(2), 35–63. https://doi.org/10.14456/rjes.2016.10

Tomas, L., Evans, N. S., Doyle, T., & Skamp, K. (2019). Are first year students ready for a flipped classroom? A case for a flipped learning continuum. *International Journal of Educational Technology in Higher Education*, *16*(1), 1–22. https://doi.org/10.1186/s41239-019-0135-4

Turan, Z., & Akdag-Cimen, B. (2020). Flipped classroom in English language teaching: a systematic review. *Computer Assisted Language Learning*, *33*(5–6), 590–606. https://doi.org/10.1080/09588221.2019.1584117

Vaezi, R., Afghari, A., & Lotfi, A. (2019). Investigating listening comprehension through flipped classroom approach: Does authenticity matter. *CALL-EJ*, *20*(1), 178–208. http://www.callej.org/journal/20-1/Vaezi-Afghari-Lotfi2019.pdf

van Alten, D. C. D., Phielix, C., Janssen, J., & Kester, L. (2019). Effects of flipping the classroom on learning outcomes and satisfaction: A meta-analysis. *Educational Research Review*, *28*, 100281. https://doi.org/10.1016/j.edurev.2019.05.003

Vitta, J. P., & Al-Hoorie, A. H. (2020). The flipped classroom in second language learning: A meta-analysis. *Language Teaching Research*. https://doi.org/10.1177/1362168820981403.

Wang, C., Lan, Y.-J., Tseng, W.-T., Lin, Y.-T. R., & Gupta, K. C.-L. (2020). On the effects of 3D virtual worlds in language learning–A meta-analysis. *Computer Assisted Language Learning*, *33*(8), 891–915. https://doi.org/10.1080/09588221.2019.1598444

Wang, J., An, N., & Wright, C. (2018). Enhancing beginner learners' oral proficiency in a flipped Chinese foreign language classroom. *Computer Assisted Language Learning*, *31*(5–6), 490–521. https://doi.org/10.1080/09588221.2017.1417872

Webb, M., & Doman, E. (2016). Does the flipped classroom lead to increased gains on learning outcomes in ESL/EFL contexts? *CATESOL Journal*, *28*(1), 39–67. Retrieved from https://files.eric.ed.gov/fulltext/EJ1111606.pdf

Xu, P., Chen, Y., Nie, W., Wang, Y., Song, T., Li, H., Li, J., Yi, J., & Zhao, L. (2019). The effectiveness of a flipped classroom on the development of Chinese nursing students' skill competence: A systematic review and meta-analysis. *Nurse Education Today*, *80*, 67–77. https://doi.org/10.1016/j.nedt.2019.06.005

Yang, J., Yin, C., & Wang, W. (2018). Flipping the classoom in teaching Chinese as a foreign language. *Language Learning & Technology*, *22*(1), 16–26. https://dx.doi.org/10125/44575

Yesilçinar, S. (2019). Using the flipped classroom to enhance adult EFL learners' speaking skills. *PASAA: Journal of Language Teaching and Learning in Thailand*, *58*, 206–234. https://files.eric.ed.gov/fulltext/EJ1227386.pdf

Zhang, H., Li, J., Jiao, L., Ma, W., & Guan, C. (2016). The adjustment and effects of vocabulary teaching strategies in flipped classroom. *Creative Education*, *7*(14), 1966-1973. https://doi.org/10.4236/ce.2016.714199

Zhang, S., Xu, H., & Zhang, X. (2021). The effects of dictionary use on second language vocabulary acquisition: A meta-analysis. *International Journal of Lexicography*, *34*(1), 1–38. https://doi.org/10.1093/ijl/ecaa010

Zou, D., Luo, S., Xie, H., & Hwang, G.-J. (2020). A systematic review of research on flipped language classrooms: Theoretical foundations, learning activities, tools, research topics and findings. *Computer Assisted Language Learning*, 1–27. https://doi.org/10.1080/09588221.2020.1839502

Corresponding author: Di Zou, dizoudaisy@gmail.com

Please cite as: Chen, X., Zou, D., Cheng, G., Xie, H., & Su, F. (2023). Effects of flipped language classrooms on learning outcomes in higher education: A Bayesian meta-analysis. *Australasian Journal of Educational Technology, 39*(2), 65-97. https://doi.org/10.14742/ajet.8019

Appendix A

Table A1
Search terms

Topic	Search terms
Flipped learning	(Flip* OR invert* OR revers*) AND (class* OR course* OR learning OR teaching OR instruction OR lecture*)
Language learning	("intercultural exchange" OR "online communication" OR "virtual exchange" OR "online interaction and exchange" OR "computer-mediated communication" OR "foreign language" OR "second language" OR "bilingual" OR "language education" OR "language learning")
Empirical study	("treatment" OR "control" OR "experimental" OR "control group" OR "experimental group" OR "pre-test" OR "post-test" OR "experimental design" OR "dependent variable*" OR "independent variable*" OR "mixed" OR "design-based" OR "quantitative" OR "qualitative" OR "empirical")

Table A2
Inclusion and exclusion criteria

Criteria	Inclusion	Exclusion
Publication date	Published from 2000 to 13 June 2020	Published outside the period
Peer-reviewed	Peer-reviewed journal articles, theses/dissertations, and conference papers	Literature of other types
Language	Written in English	Not written in English
Study type	Empirical study	Studies of other types
Availability	Full text available	Not available
Language learning	It must be about language learning	Has nothing to do with language learning
Impact study	Studies examining the effectiveness of flipped language learning	Other non-impact studies
Experimental design	Experimental or quasi-experimental studies comparing flipped-based and non-flipped-based instructions	Non-experimental studies
Dependent variable	Studies examining language learning outcomes	Studies examining other dependent variables
Learning context	Studies focusing on flipped learning classrooms in higher education	Studies focusing on flipped learning classrooms at other educational levels
Statistical information	Studies with sufficient statistical information reported for effect size calculation	Studies with insufficient statistical information reported

Table A3
Coding rubric and moderator variables

Variable types	Variables	Coded as	Reference
Publication-related	ID of the study	The order of the study	Zhang et al. (2021)
	Author(s) of the study	Author(s) names	
	Publication year	The year the study was published, e.g., 2020	
	Publication type	Journal article, thesis/dissertation, or conference paper	
	Sample size	The same size involved in the study, e.g., 10, 20	
Participant-related variables (*n* = 2)	Language learned	English, Chinese, Germany, Germany	
	School location	Asia, Europe, North America, Africa, Multiple, unspecified	van Alten et al. (2019)
Treatment-related factors (*n* = 5)	Intervention duration	> 10 weeks or 1 - 10 weeks	
	Teacher	Same for both conditions, different for both conditions, unspecified	
	Allocation	Randomised pre-existing groups, randomised on the individual level, no randomisation	
	Group equivalence test	Not tested, descriptive statement, tested, equal, not tested, no descriptive statement, tested, not equal	
	Language learning outcome	Multiple, writing, speaking, vocabulary/grammar,lListening	Self-developed
Design characteristics (*n* = 3)	Face-to-face time	FC = TC, FC < TC	van Alten et al. (2019)
	Quizzes	Addition in FC, no addition in FC	

Table A4

Study characteristics of the 26 flipped learning classroom *studies*

Publication	Grade	Duration	Type	Language	Outcome	Allocation	Region	Teacher	Intervention	Face	Group	Quizzes
Özkurkudis and Bümen (2019)	College	1 - 10 weeks	Journal Article	English	Writing	No randomisation	Europe	Different for both conditions	Tested, equal	FC = TC	Unspecified	Addition in FC
Ã akÄ±r (2017)	College	1 - 10 weeks	Journal Article	English	Speaking	Randomised pre-existing groups	Europe	Same for both conditions	Tested, equal	FC < TC	Addition in FC	No
Ekmekci (2017)	College	1 - 10 weeks	Journal Article	English	Writing	Randomised pre-existing groups	Europe	Same for both conditions	Tested, equal	FC = TC	Addition in FC	Addition in FC
Vaezi et al. (2019)	College	> 10 weeks	Journal Article	English	Listening	Randomised on individual level	Asia	Same for both conditions	Not tested, descriptive statement	FC = TC	Unspecified	Addition in FC
Lin et al. (2018)	College	1 - 10 weeks	Journal Article	English	Writing	Randomised pre-existing groups	Asia	Same for both conditions	Not tested, descriptive statement	FC = TC	Unspecified	Addition in FC
Salem (2018)	College	> 10 weeks	Journal Article	English	Writing	Randomised on individual level	Africa	Unspecified	Not tested, no descriptive statement	FC = TC	Addition in FC	Addition in FC
Leis et al. (2015)	College	1 - 10 weeks	Journal Article	English	Writing	Randomised on individual level	Asia	Same for both conditions	Tested, not equal	FC = TC	No	Addition in FC
Bicen and Beheshti (2022)	College	1 - 10 weeks	Journal Article	English	Multiple	Randomised on individual level	Europe	Unspecified	Not tested, no descriptive statement	FC = TC	Addition in FC	Addition in FC
Oraif (2018)	College	1 - 10 weeks	Doctoral dissertation	English	Writing	Randomised on individual level	Asia	Different for both conditions	Tested, equal	Unspecified	Addition in FC	No
Haghighi et al. (2019)	College	1 - 10 weeks	Journal Article	English	Speaking	Randomised on individual level	Asia	Same for both conditions	Not tested, descriptive statement	FC = TC	Addition in FC	No
Yesilçinar (2019)	College	> 10 weeks	Journal Article	English	Speaking	Randomised on individual level	Europe	Same for both conditions	Tested, equal	FC = TC	Addition in FC	Addition in FC
Lin and Hwang (2018)	College	> 10 weeks	Journal Article	English	Speaking	Randomised pre-existing groups	Asia	Same for both conditions	Not tested, descriptive statement	FC = TC	Addition in FC	Addition in FC
Webb and Doman (2016)	College	> 10 weeks	Journal Article	English	Multiple	Randomised on individual level	Multiple	Different for both conditions	Not tested, descriptive statement	Unspecified	No	Addition in FC
Bezzazi (2019)	College	1 - 10 weeks	Journal Article	English	Vocabulary / Grammar	Randomised pre-existing groups	Asia	Same for both conditions	Tested, equal	FC = TC	Addition in FC	Addition in FC
Hung (2017)	College	> 10 weeks	Journal Article	English	Speaking	Randomised pre-existing groups	Asia	Same for both conditions	Not tested, no descriptive statement	FC = TC	Addition in FC	No

Hung (2015)	College	1 - 10 weeks	Journal Article	English	Multiple	Randomised on the individual level	Asia	Unspecified	Tested, equal	FC = TC	Addition in FC	No
Iyitoğlu and Erişen (2017)	College	> 10 weeks	Journal Article	English	Multiple	Randomised pre-existing groups	Europe	Same for both conditions	Tested, equal	Unspecified	No	No
Wang et al. (2018)	College	> 10 weeks	Journal Article	Chinese	Speaking	Randomised pre-existing groups	Asia	Same for both conditions	Not tested, descriptive statement	FC < TC	Addition in FC	Addition in FC
Chen Hsieh (2017)	College	1 - 10 weeks	Journal Article	English	Speaking	Randomised pre-existing groups	Asia	Same for both conditions	Not tested, no descriptive statement	FC < TC	Addition in FC	No
Yang et al. (2018)	College	> 10 weeks	Journal Article	Chinese	Multiple	Randomised pre-existing groups	Not specified	Same for both conditions	Tested, equal	FC < TC	No	Addition in FC
Adnan (2017)	College	> 10 weeks	Journal Article	English	Writing	Randomised pre-existing groups	Europe	Same for both conditions	Not tested, no descriptive statement	FC = TC	Addition in FC	Addition in FC
Qader and Yalcin Arslan (2019)	College	1 - 10 weeks	Journal Article	English	Writing	Randomised pre-existing groups	Asia	Same for both conditions	Tested, equal	FC = TC	Addition in FC	No
Hojnacki (2018)	College	> 10 weeks	Doctoral dissertation	Germany	Multiple	Randomised pre-existing groups	North America	Different for both conditions	Tested, equal	FC < TC	Addition in FC	Addition in FC
Sherine and MJ (2020)	College	1 - 10 weeks	Journal Article	English	Vocabulary /Grammar	Unspecified	Asia	Same for both conditions	Not tested, descriptive statement	FC < TC	No	Addition in FC
Lee and Wallace (2018)	College	> 10 weeks	Journal Article	English	Multiple	No randomisation	Asia	Same for both conditions	Not tested, descriptive statement	FC = TC	Addition in FC	Addition in FC
Prefume (2015)	College	> 10 weeks	Doctoral dissertation	Japanese	Speaking	Randomised on the individual level	North America	Same for both conditions	Tested, equal	FC = TC	Addition in FC	Addition in FC

Table A5

Descriptive statistics of marginal posterior distributions: Effect-size mean

Model	*QB*	*K*	Median	Mean	*SD*	HPDI: LB	HPDI: UB	I^2	*QW*	Bayes factor
Overall		26	1.095	1.096	0.106	0.887	1.307	0.672	74.796*	0.000
Intervention duration	6.0430*	26								
> 10 weeks		13	0.931	0.932	0.145	0.645	1.223	0.612	33.410*	0.013
1 - 10 weeks		13	1.247	1.252	0.152	0.954	1.561	0.662	35.343*	0.001
Study type	3.824	26								
Journal article		23	1.145	1.146	0.111	0.929	1.368	0.661	64.579*	0.000
Doctoral dissertation		3	0.700	0.695	0.441	-0.129	1.498	0.533	6.393	28.676
Language learned	11.831*	24								
English		22	1.181	1.183	0.114	0.960	1.410	0.664	62.913*	0.000
Chinese		2	0.807	0.803	0.592	-0.020	1.605	0.138	0.052	15.582
Outcome category	16.507*	25								
Multiple		7	0.870	0.866	0.159	0.546	1.177	0.393	11.743	0.348
Writing		8	1.353	1.361	0.239	0.897	1.847	0.712	27.140*	0.082
Speaking		8	1.039	1.043	0.181	0.686	1.412	0.500	17.212*	0.155
Vocabulary/ grammar		2	0.830	0.806	0.644	-0.144	1.615	0.304	2.194	18.545
Allocation	11.093*	25								
Randomised pre-existing groups		13	1.033	1.036	0.142	0.756	1.323	0.614	32.400*	0.004
Randomised on the individual level		10	1.247	1.247	0.143	0.959	1.532	0.454	19.523*	0.003
No random-isation		2	1.096	1.152	1.905	-1.587	4.054	0.882	11.780*	29.618
School location	18.001*	23								
Asia		14	1.031	1.033	0.130	0.776	1.293	0.621	36.304*	0.001
Europe		7	1.405	1.415	0.247	0.937	1.925	0.676	20.225*	0.124
North America		2	0.358	0.359	0.621	-0.467	1.188	0.198	0.266	143.916
Teacher	3.775	26								
Same for both conditions		19	1.046	1.047	0.133	0.784	1.313	0.705	60.214*	0.000
Different for both conditions		4	1.147	1.159	0.340	0.517	1.850	0.503	8.806	2.957
Unspecified		3	1.303	1.303	0.247	0.860	1.747	0.165	2.001	0.754
Group equivalence test	1.198	25								
Not tested, descriptive statement		8	1.084	1.084	0.212	0.661	1.509	0.734	28.784*	0.299
Tested, equal		12	1.096	1.100	0.187	0.733	1.477	0.704	35.939*	0.033
Not tested, no descriptive statement		5	1.064	1.063	0.194	0.676	1.444	0.384	8.875	0.559
Face-to-face time	2.577	26								
FC = TC		17	1.172	1.174	0.141	0.897	1.457	0.722	56.926*	0.000
FC < TC		6	0.873	0.874	0.254	0.364	1.385	0.581	14.500*	4.152
Unspecified		3	1.082	1.082	0.251	0.625	1.538	0.131	0.793	1.351
Quizzes	0.061	26								
Addition in FC		18	1.107	1.108	0.141	0.830	1.389	0.725	59.257*	0.000

No		8	1.058	1.061	0.161	0.745	1.388	0.417	15.478*	0.057

Note. HPDI = highest posterior density interval; LB = lower bound; UB = upper bound. *p < .05.

Table A6

Descriptive statistics of marginal posterior distributions: Between-studies standard deviation

Model	*K*	Median	Mean	*SD*	HPDI: LB	HPDI: UB	Bayes factor
Overall	26	0.419	0.427	0.099	0.241	0.627	0.000
Intervention duration	26						
>10 weeks	13	0.379	0.392	0.141	0.132	0.686	0.027
1-10 weeks	13	0.403	0.416	0.162	0.108	0.754	0.049
Study type	26						
Journal article	23	0.405	0.414	0.104	0.220	0.624	0.001
Doctoral dissertation	3	0.342	0.466	0.546	0.000	1.293	0.643
Language learned	24						
English	22	0.408	0.417	0.106	0.221	0.631	0.000
Chinese	2	0.157	0.334	0.939	0.000	1.117	1.472
Outcome category	25						
Multiple	7	0.228	0.250	0.165	0.000	0.546	0.538
Writing	8	0.500	0.534	0.221	0.160	0.987	0.013
Speaking	8	0.314	0.336	0.213	0.000	0.715	0.475
Vocabulary/grammar	2	0.193	0.391	1.033	0.000	1.285	1.054
Allocation	25						
Randomised pre-existing groups	13	0.366	0.375	0.158	0.048	0.688	0.105
Randomised on individual level	10	0.272	0.284	0.170	0.000	0.583	0.485
No randomisation	2	0.984	1.517	2.731	0.000	4.149	0.052
School location	23						
Asia	14	0.355	0.366	0.128	0.128	0.634	0.026
Europe	7	0.455	0.489	0.253	0.000	0.941	0.103
North America	2	0.165	0.349	0.978	0.000	1.163	1.539
Teacher	26						
Same for both conditions	19	0.457	0.468	0.121	0.248	0.712	0.000
Different for both conditions	4	0.332	0.427	0.414	0.000	1.148	0.742
Unspecified	3	0.120	0.202	0.290	0.000	0.645	1.572
Group equivalence test	25						
Not tested, descriptive statement	8	0.458	0.489	0.199	0.154	0.895	0.010
Tested, equal	12	0.491	0.510	0.184	0.174	0.894	0.019
Not tested, no descriptive statement	5	0.213	0.258	0.221	0.000	0.657	0.821
Face-to-face time	26						
FC = TC	17	0.458	0.472	0.127	0.242	0.727	0.000
FC < TC	6	0.390	0.429	0.292	0.000	0.949	0.403
Unspecified	3	0.116	0.191	0.273	0.000	0.604	1.824
Quizzes	26						
Addition in FC	18	0.471	0.485	0.129	0.250	0.745	0.000
No	8	0.256	0.279	0.189	0.000	0.618	0.636

Note. HPDI = highest posterior density interval; LB = lower bound; UB = upper bound