Evaluating the effectiveness of learning activities in a flipped classroom: A case study of an English-as-a-foreign-language (EFL) class in a Chinese university

Yan Shen

Higher Education: Research Evaluation and Enhancement, PhD

September 2022

This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Department of Educational Research

Lancaster University

UK

Abstract

This single-site case study aims to evaluate the effectiveness of the learning activities in a flipped EFL class so that critical pedagogical insights can be generated to enhance the design and implementation of flipped classrooms. Student engagement in learning activities was explored and examined to understand how the flipped classroom (FC) activities enhanced learning. A convergent parallel mixed-methods design with nested samples was adopted to collect data over a 15-week semester. Participants were from a flipped EFL class (n = 25) and a blended EFL class (n = 28). Data were collected from focus group discussions, participant observation, the learning management system (LMS) and iWrite. An adapted four-level Kirkpatrick Model was employed as the evaluation framework to explore student engagement in the learning activities and measure the learning outcomes. Kahu's engagement framework was employed to discuss the influencing factors. The findings of this project showed that the flipped design enhanced student performance in the learning process though there was no statistically significant difference in student final examination grades between the two classes. The findings further indicated that student engagement in learning activities was influenced by the interplay of teaching practices and student characteristics. Meanwhile, it was mediated by the assessment policy and practice and shaped by significant socio-cultural influences of China's educational system and social conventions. Evidence-based reflections were then made to improve the design, implementation and evaluation of flipped classrooms. This project fills the current knowledge gap in the learning process of the flipping pedagogy and

furthers our understanding of the mechanisms contributing to student engagement in flipped learning.

Table of Contents

List of Figures and Tables

Acknowledgements

Above all, I would like to express my deepest gratitude to my supervisor, Dr Kyungmee Lee, who guided me patiently through the dissertation process. She read my drafts carefully and provided valuable feedback to improve them. She is always supportive and responsive. Most importantly, she has been an encouraging model of a diligent academic and loving mother, inspiring me to make it through my doctoral years while striving for a balance between life and work.

I am deeply indebted to the academic staff on the Doctoral Programme in Higher Education Research, Evaluation and Enhancement. Professor Paul Trowler introduced me to the field of educational research. Under his guidance, I did the first mixed-methods research in my academic life. Professor Murray Saunders led me to educational evaluative studies, which resulted in this dissertation evaluating the flipped classroom. Dr Gemma Derrick read my assignment paper on flipped classrooms meticulously and offered constructive revision feedback. It was a great help in developing my understanding of the focus of my dissertation. Dr Janja Komljenovic coached us on how to do robust educational research. Without them, this dissertation would not have come into existence. My heartfelt thanks also go to Dr Rebeca Marsden, who quickly came to my rescue whenever I needed technical support and Mrs Kathryn Doherty, who made me stay connected with the Department and feel supported even though I was away from campus.

My heart also overflows with gratitude to my colleagues and students at Shanghai University of Political Science and Law. My colleagues were the

patient audience when I was talking about my flipped classroom. They gave me helpful advice on improving the design and implementation of the class. My heartfelt gratitude also goes to my students of 2020 for agreeing to participate in my dissertation project and sharing their experiences in the focus group discussions. They provided valuable data for this dissertation and helpful suggestions to improve the flipped class.

This acknowledgement would never be complete without my family members, my dear parents, husband and son, who are always behind me, unconditionally supporting whatever I decide to do. My thank-you also goes to our lovely furry Bobo. He came to us the year I started this PhD programme and has been a great helper and amazing companion in following a well-scheduled life.

Author's declaration:

I hereby declare that my dissertation

- Is my own work and has not been submitted in substantially the same form for the award of a higher degree elsewhere.
- Does not exceed the prescribed word limit of 45, 000 words, including the main text of the thesis, footnotes, data and text incorporated into diagrams, tables or figures, excluding the material preceding the main text of the dissertation (e.g. the title page, contents and abstract) and the material following the main text of the thesis (e.g. the appendices, and the list of references). The word count for this dissertation is 44,881 words.

Signature ……Yan Shen…….

Chapter 1: Introduction

1.1 Research Background

Teaching and learning have always been at the core of higher education (HE). However, the two have never been as simple as teachers teach, and students learn. That partly explains the ongoing scholarly interest in researching teaching and learning practices in HE despite the tremendous change in its systems over the years.

The last few decades have witnessed a global massification of HE. Based on the assumption that “most people will participate and on more than one occasion” (Tight, 2012, p. 3), universities moved from elite systems that involve only a small minority of the population to mass systems. The result is that universities worldwide are now having bigger class sizes with more diversified student groups. The mass systems add to the challenges of good teaching practices, which, according to Tight (2012), are defined by teachers’ having enough knowledge of their students’ learning needs and the teaching and learning context and being able to cater to different student needs. Along with the increased challenges in teaching is the growing attention to teaching quality and excellence. To demonstrate the effectiveness of innovation in teaching is now one of the standards of the HE literature (O’Flaherty & Phillips, 2015; Praslova, 2010; Sozer et al., 2019).

Another remarkable change in the contemporary HE is the integration of digital technology with the fabric of teaching and learning. Research interest in the related areas has mushroomed in recent years (Cheng et al., 2019). The

popularity of the blended/flipping pedagogy in HE is an example (Abeysekera & Dawson, 2015; Akçayıra & Akçayır, 2018). This conjunction of e-learning with conventional teaching is advocated by many of its practitioners as both a solution to the growing class sizes and a channel to enhance student engagement in learning. Though voices are appealing to in-depth investigations into the extent digital technology can enhance teaching and learning (Selwyn, 2016), it is undeniable that digital technology has become the norm in modern universities (Henderson et al., 2015; Macfadyen & Dawson, 2012) and irreversibly changed conventional teaching and learning practices in HE (Lin & Mubarok, 2021).

As such, this study evaluates the effectiveness of a flipped EFL classroom, which is a practice based on exemplars of the flipping pedagogy that is gaining growing recognition in higher education institutions (HEIs) worldwide (Cheng et al., 2019). The flipped EFL classroom under evaluation in this project is an outcome of China's educational reform in English teaching. It is a professional imperative to change in curriculum policy and is resource-driven by modern technology. Meanwhile, it is a by-product of crisis management in the Corona Virus Disease 2019 (COVID-19) pandemic.

This evaluative project is situated within the global HE communities of practitioners of the flipping pedagogy as it is now increasingly recognized and practised in HEIs worldwide. However, it should still be noted that this project was carried out in the HE context in China. Robertson and Dale (2015) note that the civilisationally based culture through which the education ensemble is

constructed and mediated should be distinguished because contextual factors can shape the research outcomes.

China's education is a highly centralized system run by the Ministry of Education (MoE) of the central government. Its educational policies run in a top-down direction and embody the state's will and interests. English is now the dominant international language, and the importance of teaching and learning English is self-evident to any non-English-speaking country (Sun et al., 2016). Since it was officially listed as China's first foreign language by the MoE in 1964, English has become a national prescribed school subject at all educational levels. It is also a subject to be tested in all high-stakes tests for almost all officially registered students to further school education in the country.

In China, the status of English is closely related to the changing socio-political context (Cheng & Wei, 2021). English education policies are made and enacted to serve the country's need for political, economic, and social development and safeguard national interests. English teaching in China has long focused on developing learners' language knowledge and skills. Such orientation has a strength in developing learners' generic language skills, especially reading skills, in a relatively short period. It used to serve well China's need to learn from the outside world for advanced technological and economic development upon its Opening-up Reform (Sun et al., 2016). However, with China's entry into the World Trade Organization (WTO) and its growing involvement in international exchange, the country needs a competent workforce to promote its position in global competition. University graduates with only generic language

skills in English fell short of this challenge. To strengthen international cooperation and keep up with the latest technological development, in 2001, the same year China entered into the WTO, the MoE issued an ambitious policy *Guideline for Improving the Quality of Undergraduate Teaching* (MoE, 2001) to promote English as media of instruction (EMI). Though faced with heated controversy, it ushered in the reform of English teaching in mainland China's HEIs. In 2010, the MoE issued *Outline of China's National Plan for Medium and Long-term Education Reform and Development (2010-2020)*, explicitly stressing the need for HEIs to *produce students and researchers equipped with international perspectives, familiar with international rules and capable of global competitions* (MoE, 2010). The *Outline* became the guideline for many educational policies and initiatives during the following ten years. One example is that HEIs across the country started converting their English courses for generic purposes (EGP) to English courses for academic purposes (EAP) or specific purposes (ESP).

1.2 Research Problems

Effective evaluation is conducive to quality assurance and enhancement in teaching practice. In terms of quality control, evaluation systems incorporating the measurements of technology application in teaching have not been fully established yet at the institutional level (Macfadyen & Dawson, 2012). Practitioners and researchers are still looking for frameworks or tools that can systematically and effectively evaluate classes blended with modern technology (Feng et al., 2018). E-learning is different from traditional classroom teaching, and particular techniques need to be adapted for evaluation. However, most

evaluations at the institutional level are still using frameworks that cater to traditional classroom teaching. The lack of evaluation standards in HE results in the absence of practice norms that guarantee effective technology implementation in teaching. The challenges of how best to teach with technology have been persisting ever since computer technology was introduced into the education sector (Strayer, 2012). However, implementation and evaluation in this regard have so far fallen mainly within the initiative of individual faculty at the course level. In addition, the current HE evaluation systems generally suffer from the following handicaps: evaluation outcomes are not communicated to students in a useful or timely manner; student feedback in course evaluation is seldom acted on (Alderman et al., 2012).

The flipping approach is not just re-ordering learning activities in and out of class but an expansion of traditional classrooms. In flipped classrooms, direct instruction is moved before class, often via asynchronous online instructional videos, and the class time is freed up to entice active learning via various learning activities, such as problem-solving, discussion, and peer collaboration. There is no fixed practice of a flipped model; learning activities in a flipped classroom can be organized flexibly to realize various teaching purposes or cater to specific learning needs. The core concept is to transmit subject content to prepare learners for active classroom participation before class and promote deep learning and higher-order cognitive skills via constructive learning processes in class.

In practice, the existing evaluation or research on flipping pedagogy measures more the “effects” of the pedagogy on student learning outcomes or satisfaction

than the “processes” that lead up to these effects (Betihavas et al., 2016), resulting in a knowledge gap in what is actually happening in the process of teaching and learning (Cheng et al., 2019; Kim et al., 2017). Consequently, it adds to the difficulty of understanding the inconsistent “effects” in the existing findings. For instance, in addition to the enhanced outcomes, research findings also indicate equal academic achievements between traditional and flipped classes (Al-Zahrani, 2015; Jensen et al., 2015) and mixed student satisfaction with flipped practices (Lombardini et al., 2018; McNally et al., 2017). Without a careful investigation into the processes as to how the flipped class is implemented or how students engage themselves in the flipped class, what brings about these inconsistent outcomes is not fully understood yet, and improvement is difficult to be made.

Though the decisive roles of learning activities in flipped classrooms have been widely acknowledged (Jensen et al., 2015; Lombardini et al., 2018), they are the most controversial part related to student satisfaction. In a flipped class, a succession of learning activities should be in place to prepare students for learning and engage them in learning so that students not only learn course content and acquire professional skills but, ideally, develop their higher-order thinking skills. However, quite a few flipped practice studies report divergent student attitudes towards learning activities (Al-Zahrani, 2015; Lombardini et al., 2018; McNally et al., 2017; Munir et al., 2018). As most flipped practice research measures outcomes of the flipping pedagogy, little is known about the extent to which learning activities enhance student engagement in the process

of learning, what leads to the differences in student satisfaction, and how improvements in design and implementation should be made.

Methodological limitations in current evaluative or academic research on flipping pedagogy have also resulted in a vague understanding of the pedagogical effects and limited insight into the enhancement of the pedagogy (Lee & Wallace, 2018). Much of the existing research on flipping pedagogy adopts quantitative positivist methods (Cheng et al., 2022; McNally et al., 2017), which are adequate for the judgemental purpose (Steyn et al., 2019). However, they are not robust enough to capture diverse learning experiences that lead to learning outcomes or produce in-depth insights into effective changes to enhance teaching and learning practices (Sozer et al., 2019). As a result, many researchers call for more robust research designs to evaluate the effectiveness of the flipped classroom model (Abeysekera & Dawson, 2015; Bishop & Verleger, 2013; Hung, 2015; Låg & Sæle, 2019; O'Flaherty & Phillips, 2015).

1.3 Rationales for the Study

The rationales for this study derive from the current status of flipping pedagogy as a global practice in HEIs, the widespread evaluative practices in HE as a means of quality assurance and enhancement, and my stance as a teacher-researcher.

1.3.1 **The Flipping Pedagogy**

Flipping pedagogy features direct computer-based individual instruction before class and interactive group-based classroom learning activities (Bishop & Verleger, 2013). The popularity of flipping pedagogy is high and projected to increase in the current HE sector (Akçayıra & Akçayır, 2018; McNally et al., 2017; Turan & Akdag-Cimen, 2020). Numerous studies have confirmed that flipped classrooms are more flexible and student-centric than traditional classrooms and, therefore, conducive to active and deep learning (Akçayıra & Akçayır, 2018; Betihavas et al., 2016; Bishop & Verleger, 2013; Hung, 2017). Another critical feature of a flipping approach is its applicability to various disciplines and student groups (Hao, 2016; Hung, 2017; Strelan et al., 2020). English teaching in China has been long criticized as *a kettle that never boils*. *Time-consuming* and *ineffectiveness* are the rhetoric frequently associated with it. Academics and practitioners in China have been dedicated to seeking effective and efficient approaches to EFL teaching and learning. As such, a flipping pedagogy is readily welcomed and received in China (Liu et al., 2019). Along with its application comes a rise of interest in researching the effectiveness of the pedagogy, which conforms with the quality systems of the current HE sector.

1.3.2 **Course Evaluation**

Course evaluation lies in the most basic level of quality assurance and enhancement in HE (Tight, 2012). Within the current HE sector, most of the evaluation attention is directed to quality appraisal. Course evaluation for

enhancement purposes in HEIs, therefore, falls much within the remit of individual academics (Saunders et al., 2011). Driven chiefly by academic interest, course evaluations are of more academic value than externally imposed evaluative practices (Bamber, 2011a). They have become one of the most commonly used approaches to measuring the effectiveness of teaching and learning practice and consulting students with their experience in learning (Freeman & Dobbins, 2013; Nicolaou & Atkinson, 2019). Moreover, many scholars and practitioners believe that judging quality is the least important aspect of course evaluations; the most important lies in their value in providing intelligence for improvement (Edström, 2008; Nicolaou & Atkinson, 2019). By obtaining sufficient information for valid decision-making from evaluating their courses, teachers are able to make evidence-based improvements in their teaching practices and exert more impact on student learning.

1.3.3 **Personal Motivation**

Admittedly, the primary driving force of my commitment to this project comes from my stance as a teacher-researcher. This positioning has long shaped and influenced my teaching practice and research interest. My main concern in this project is to seek evidence of a flipping approach's strengths and enhance teaching and learning practices in the course.

The worldwide online educational provision during COVID-19 witnessed the strengths of digital technology in education as a means of crisis management. After that, the use of technology appeared to turn from a norm in HE to a professional imperative in teaching. The teaching staff in my university was

encouraged to move their courses into the learning management system (LMS) even after school reopened and teaching was back to normal. So, in the semester after the COVID-19 lockdown in 2020, I converted one of my blended Integrated English Course (IEC) classes into a flipped class. In response to the university policy, I uploaded video-recorded instructional content into the LMS and redesigned learning activities for IEC. I did that also in an attempt to understand, through comparison, which approach, the flipped design or the blended one, worked better in IEC. Technology is, after all, a tool; the extent it enhances teaching and learning depends on how it is used. As technology was becoming a must in my work, the question then became how I could exploit its full potential in my class.

My knowledge of the HE context's complexity led to my pragmatic epistemological stance and my data collection decision in this project. I hold that learning outcomes can only be clearly understood and reasonably interpreted when examined in relation to the learning process and student experiences. I brought in student voices via focus group discussions to shed light on what contributed to student learning outcomes in IEC. To curb the potential bias from self-reported data, I took field notes from observation and collected log data from the LMS for triangulation.

1.4 Research Aims and Research Questions

Driven by my stance as a teacher-researcher and in response to the research problems mentioned above, this project aims to explore and understand the effectiveness of learning activities in enhancing student engagement in the

learning process and their learning outcomes in a flipped EFL classroom from an evaluative perspective. By doing so, it hopes to bring about evidence-based improvement to the design and implementation of learning activities in flipped EFL classrooms, offer pedagogical insights into teaching and learning enhancement to the wider communities of practitioners of a flipping pedagogy and contribute to the literature on flipping pedagogy and formative course evaluation.

Based on the research aims, this project seeks to answer the following overarching research questions (RQs):

RQ 1. To what extent are students in the flipped classroom (FC) engaged in the learning activities?

RQ 2. What are the overall effects of the learning activities on student engagement in the learning process and learning outcomes in FC?

RQ 3. What affects student engagement in learning activities in the flipped classroom?

RQ 4. What are the reflections and areas for improvement regarding flipped EFL classrooms and formative course evaluation?

An adapted Kirkpatrick Model (Kirkpatrick & Kirkpatrick, 2005; Kirkpatrick & Kirkpatrick, 2006) was adopted as the evaluative framework for this project, where it played multiple evaluative purposes. The adapted Model, as a whole, worked as an evaluation plan for this project. The four levels, the affective, cognitive, behavioural, and results levels, in the Model served as evaluation

foci. Indicators in each level were either data-driven or derived from the findings of the extant flipped practice literature. The Model managed to measure both the processes and the outcomes of student engagement in learning activities and contributed to answering the research questions of the project. A convergent parallel mixed-methods design was employed for data collection. Qualitative data that aimed to capture student engagement in the learning activities during the learning process were collected from focus groups and the researcher's field notes of participant observation. Quantitative data on student performance in the learning tasks throughout the semester and the subsequent learning outcomes were collected from LMS and iWrite. Data were collected from both FC and BC to understand the features typical of FC. The qualitative and quantitative datasets are then merged together and discussed, by referring to Kahu's conceptual framework of student engagement (Kahu, 2013; Kahu & Nelson, 2018), to uncover the factors influencing student engagement and the areas for improvement in flipped EFL classrooms and formative course evaluation. In a word, the mixed-methods design aims to produce a holistic understanding of both the learning process and outcomes in the flipped IEC.

1.5 Significance of the Study

In practice, this project embodies the unique strengths of evaluative studies. What sets this evaluative project apart from other scholarly research is that the participants of this project can be the direct beneficiaries due to the process use of evaluation (Patton, 1998). By participating in the evaluation, students maximized their voice and increased their course ownership (Wilson et al., 2022). As Edström (2008) argues, the empowerment of students in course

evaluation makes them more active learners in the educational community than consumers in the educational market. To the teacher-researcher, the increased knowledge of the project and its participants under evaluation led to more precise goals for improvement. Moreover, as the popularity of flipping pedagogy and the emphasis on quality assurance and enhancement in HEIs are international phenomena, this project can produce insights for the wider communities of practitioners of flipping pedagogy and course evaluation. In addition, administrators in HEIs might find this research useful in making policy, establishing evaluation systems, and planning faculty training programmes.

Methodologically, this project demonstrates how qualitative and quantitative approaches work well together in the Kirkpatrick Model to enhance the credibility of course evaluation findings. As an evaluative study in a field where quantitative methods prevail, the project brings qualitative approaches to capture student learning experiences from their voices to situate the quantitative findings. The mixed-methods design demonstrates how log data from LMS can be matched with student voices to produce rich and precise empirical evidence of student engagement in learning, which is more valuable than qualitative or quantitative data alone to inform pedagogical decisions and improve research practices. When doing the aforementioned, this project verifies the applicability of the Kirkpatrick Model in evaluating the effectiveness of technology-enhanced learning. Most importantly, by focusing on learning activities, the project captures a more detailed collection of events in learning experiences, hence, offering better grounds for data analysis and prediction.

Theoretically, by investigating student engagement in the learning process and the subsequent outcomes, the project brings a holistic understanding of student experiences in the flipped classroom. By focusing on learning activities, it adds to the knowledge of what is happening in the learning process of the flipped classes, hence, filling the knowledge gap in the current flipped practice literature. The learning activities open up a window to look into the complex interplay between the individual psychosocial factors, the institutional structural factors, and the wider socio-cultural factors that affect student engagement in learning. This furthers our understanding of the mechanisms contributing to student engagement in flipped learning, which the existing literature has not clearly articulated (Zhoc et al., 2018). Meanwhile, the project contributes to an improved understanding of how learning activities should be designed, organized, and implemented in flipped classrooms to enhance learning, thus helping pave the way for flipping innovation.

1.6 Structure of the Thesis

This dissertation consists of seven Chapters.

Chapter One begins with a brief introduction of the research background for this project. Next, it points out the research problems in the fields of flipping pedagogy and course evaluation and explains the rationale for this study and the author's role in this research. It then puts forward the research aims and questions, briefly introducing the methodology employed to answer these questions. The chapter concludes with a discussion of the significance of the project.

Chapter Two reviews the literature on course evaluation, flipping pedagogy, and student engagement to lay the ground for the research focus, research design and conceptual framework for this project.

Chapter Three elaborates on the methodology of this research. It begins with the rationales for the mixed-methods design of this evaluative study and the sampling strategy and the participants and then goes on with the broader issues of a pragmatic epistemological stance that underpins this design. The evaluative framework for this study is presented next. After briefly describing the research context, the chapter then elaborates on the convergent parallel mixed methods design with a detailed explanation of the data collection and analysis methods. The chapter concludes with discussions of the validity and reliability of the study and the ethical issues involved.

Chapter Four reports the study's qualitative findings, which focus on students' affective, cognitive and behavioural engagement in the learning activities and answer RQ1 and its sub-questions.

Chapter Five reports the study's quantitative findings in corroboration with the qualitative results. This chapter answers RQ2 and its sub-questions: student performance in learning tasks and the overall outcomes.

Chapter Six discusses the factors affecting student engagement and learning outcomes. Critical reflections on the flipped classrooms and formative course evaluation are made based on the merged qualitative and quantitative findings. RQ3 and RQ4 are answered in this chapter.

Chapter Seven concludes the research findings in full based on the foregoing analyses and discussions and summarizes the contributions and limitations of this study.

Chapter 2: Literature Review

2.1 Introduction

This chapter aims to lay the ground for this project's research focus, research design, and theoretical framework by reviewing literature in the related fields. Peer-reviewed studies on course evaluation, flipping pedagogy and student engagement are the main influences in this project's literature. Criteria of relevance (Maxwell, 2006), importance (Hwang & Tsai, 2011) and duration (Denner et al., 2016) were applied in the initial search for the related literature. The initial search started with the terms "flipping pedagogy", "flipped classroom", "course evaluation", or "student engagement" in titles of SSCI-indexed journal articles within the latest five years in the category of "higher education" by using Lancaster's library search engine OneSearch. The abstracts of the resulting papers were then read to identify literature having important implications for research design, conduct and interpretations of flipped teaching and learning. When reading the full texts of the resulting articles, papers highly cited by the authors were further included into the literature list. By doing so, a systematic body of literature for this project was built up, covering domains of learning theories, foreign language teaching and learning, technology-enhanced learning and curriculum/course design, in addition to flipping pedagogy, course evaluation and student engagement. Based on the criteria of relevance and importance in literature review and given the research focus of this project to investigate the effectiveness of learning activities in a flipped classroom, literature on flipping pedagogy, course evaluation and student engagement is reviewed in this chapter. Literature that

is not directly related to the scope of investigation of this project, such as that on curriculum/course design, is not included in the review.

This project evaluates a flipped EFL classroom; this review chapter starts with literature on course evaluation. It first elaborates on the purposes and foci of current course evaluation practice and then discusses the strengths and weaknesses of the evaluative methods in use to justify using a mix-methods design in this project. The following section reviews the literature on flipping pedagogy, the focus of this project. It centres on the theoretical underpinning, pedagogical strengths, the factors affecting pedagogical effectiveness, and the applicability of the flipping approach in EFL classrooms. When doing so, it identifies the research gap in the existing literature on flipped practices and rationalises the in-depth examination of the effectiveness of learning activities in this project. The chapter then zooms in on the literature on student engagement, discussing the existing conceptualization issues and, hence, the challenges in its measurement, justifying Kahu's conceptual framework to examine student engagement in-depth in this project. This section further illuminates the use of student engagement and learning outcomes as indicators of pedagogical effectiveness of the flipped EFL classroom under evaluation in this project and further clarifies how these two indicators are to be measured effectively.

2.2 Course Evaluation

One essential feature of evaluative studies is that they generate new knowledge or reveal the hidden knowledge that can be used by all participants

(Saunders, 2011). Evaluation has two classic purposes: quality assurance and quality enhancement, also widely referred to as judgemental or developmental purposes (Edström, 2008). Quality assurance or judgemental purpose aims to create a base for fair quality appraisal, while quality enhancement or developmental purpose investigates aspects that can be improved. Course evaluation falls into the domain of self-evaluation practice that aims to achieve diagnostic or improvement purposes through practitioners' reflective practices (Bamber, 2011b; Saunders, 2011; Saunders, 2012). Self-evaluative practices are often regarded as a form of informal evaluation nested in academic cultures, therefore, of more academic values than externally imposed evaluative practices (Bamber, 2011a). They are self-driven and characterized by the significant discretion of individual practitioners. At the same time, they are bounded by the cultural and social practices of the evaluative context (Saunders, 2011).

2.2.1 **Formative Course Evaluation for Improvement Purposes**

Course evaluation inherits the classic dichotomous classification as summative and formative. Both play the role of informing teaching and learning practices (Scriven, 1996). Summative course evaluation generally serves a judgmental purpose. It is widely applied to measure success or effectiveness that helps identify good teaching practice, examine the curriculum quality, measure student satisfaction and progress, compare modules, inform appointment, promotion, tenure decisions and university funding and ranking (Alderman et al., 2012; Braun & Leidner, 2009; Holland, 2019; Nguyen & Foster, 2018; Steyn et al., 2019). In contrast, formative course evaluation is believed to serve a

developmental purpose. It helps improve teaching practices and student learning experiences by investigating the strengths and weaknesses of a course and lending insights into student learning and progress (Nguyen & Foster, 2018), student motivation, and their changing needs (Steyn et al., 2019) and gains. Hence, it allows for informed decisions about course design, amendments and development (Carbone et al., 2014; Fisher & Miller, 2008; Freeman & Dobbins, 2013; Sozer et al., 2019; Steyn et al., 2019).

While both purposes share the same long-term goal, to improve educational quality, summative evaluations are often criticized as disconnected one-offs and having limited contribution to course development. Formative evaluations, on the other hand, via timely feedback, offer the greatest scope for improvement (Fisher & Miller, 2008), benefiting not only instructors and students in the course under evaluation but the future cohorts (Holland, 2019). Though some scholars argue there is a tension between judgemental and developmental purposes, and the same data seldom serve both purposes well (Patton, 2012), others hold that if the developmental purpose is adequately addressed, evidence for quality appraisal or assurance is decided automatically, which is a similar idea to “audit through self-audit” (Edström, 2008, p. 99). Researchers argue that course evaluations should provide intelligence on making improvements rather than be limited to testing the effectiveness (Edström, 2008; Nicolaou & Atkinson, 2019).

2.2.2 Course Evaluation Focusing on Learners

Shifts are found in the focus of course evaluation practice. Conventionally, course evaluations are basically a practice of rating teachers and teaching, which dominantly focus on student satisfaction with teaching quality, course resources, learning support, and their experiences (Carbone et al., 2014; Edström, 2008; Nguyen & Foster, 2018; Sozer et al., 2019; Steyn et al., 2019). Researchers find an alignment between teachers' effort and increased student satisfaction (Carbone et al., 2014). However, studies show that student satisfaction does not necessarily correlate positively with teaching quality and student learning outcomes.

The Bologna process shifts the focus of course evaluation to learning outcomes (Nguyen & Foster, 2018). Many educators endorse the effectiveness of competence-based evaluations in measuring learning outcomes (Braun & Leidner, 2009). However, concerns exist about how to accurately and reliably assess student competencies. Other factors, like student motivation and learning experiences, are also found to be positively related to learning outcomes (Carbone et al., 2014).

Increasing evidence in educational evaluation supports that evaluations should be learner-focused so that teaching and learning activities can be improved to better support desired learning outcomes (Praslova, 2010). Therefore, the evaluands are what students do and perceive during their learning process. Aspects covered so far in this regard in the literature include student perceptions of learning demand and learning per se (Edström, 2008), student

learning styles, expectations and concerns (Fisher & Miller, 2008; Nielsen & Kreiner, 2017), student approaches to learning, their extent of engagement in learning and interaction with peers (Park, 2014; Sozer et al., 2019; Steyn et al., 2019) and student perceived gains (Nguyen & Foster, 2018). Edström (2008) finds that students' perception of learning and the learning context influence how they go about their learning. Carbone et al. (2014) contend that students' perceptions of their learning environment are a stronger predictor of learning outcomes than their previous academic achievement. Students' educational background is a factor that affects students' perceptions, concerns and preferences in teaching and learning (Sozer et al., 2019). In addition to standardized tests to measure student gains, Kuh et al. (1997) testify that active learning and cooperation among students are the two best process indicators of student gains in general education, personal-social development and intellectual skills.

2.2.3 **Strengths and Weaknesses of Evaluation Methods**

Conventionally, quantitative methods are used as evaluative instruments in most course evaluations. The quantitative approach appears more objective and straightforward in administration and analysis (Erikson et al., 2016; Holland, 2019), with evaluators staying distant from the evaluands and analysis based on statistics. However, quantitative evaluations have often been criticized for their limited capacity to contribute to course improvement (Richardson, 2005) and questionable reliability and validity (Alderman et al., 2012; Holland, 2019).

Questionnaires, the widely adopted instrument in quantitative course evaluations, receive the most challenges. The existing literature suggests these surveys are mostly teaching-centric (Edström, 2008), which, however, does not indicate so much good teaching (Kember et al., 2008) as student satisfaction (Braun & Leidner, 2009; Freeman & Dobbins, 2013; Nguyen & Foster, 2018). Nevertheless, student satisfaction is not constantly correlated with student gains (Kuh et al., 1997) or teaching quality (Richardson, 2005). Therefore, the results of the questionnaire surveys do not necessarily point to any course of action for improvement. Besides, researchers argue that the predefined categories in these questionnaires lack flexibility and over-generalize disciplinary characteristics (Freeman & Dobbins, 2013), thus making it impossible to fully represent the diverse student experience (Steyn et al., 2019). Consequently, they not only limit their contribution to improving learning and teaching but also undermine the validity of the questionnaires.

Qualitative course evaluations have been cited as a viable alternative to quantitative evaluations. Supporters of the qualitative approach believe qualitative methods, free from the confinement of predefined paradigms, can capture richer, deeper and more context-specific pictures of student learning experiences. Hence, more appropriate remedial actions can be taken to improve learning and teaching. Steyn et al. (2019) found that more comprehensive and in-depth dimensions concerning course content, learning support, learning environment and course administration were covered in their qualitative course evaluation than in the previous quantitative evaluations used in their institution. They conclude that qualitative evaluations enable extensive

examinations of student experience and produce more relevant information to improve teaching and learning. In addition, supporters argue that a culture of shared responsibility in qualitative course evaluations engages students as active contributors and co-creators of learning, empowers them to contribute constructively to the teaching and learning process (Freeman & Dobbins, 2013) and motivates both students and teachers (Holland, 2019).

On the other hand, qualitative course evaluations are primarily commented as time-consuming and resource-intensive. Due to the complexity of administration and analysis, they are difficult to be applied to a large-sized student population. The trustworthiness of qualitative course evaluations is also in question (Mabry, 2003). The quality of the unstructured self-reported student feedback is hard to control; it relies on the extent of understanding and engagement of the participants (Freeman & Dobbins, 2013) as well as the capabilities of evaluators and may suffer from student biases (Braun & Leidner, 2009). Research suggests that student feedback can be either too diverse to handle or too specific to be useful; accurately reporting the participants' diverse perspectives is even more challenging (Steyn et al., 2019). Ethical issues are also a major concern. So far, far less research has been conducted into the efficacy of qualitative course evaluations compared to quantitative ones (Steyn et al., 2019).

The mixed-methods approach established its early presence in evaluative research (Cohen et al., 2011) and has been gaining increasing attention in the evaluation community (Saunders et al., 2011). This approach is recognized as more challenging for data collection, analysis and integration and more

demanding on research time and resources (Creswell & Plano Clark, 2017). Nevertheless, Mertens and Hesse-Biber (2013) contend that mixed methods make up for the weaknesses of the conventional quantitative methods that fail to address the social aspects in which the evaluators are interested. Literature shows that the mixed-methods design provides a fuller understanding of the evaluation problems and achieves the synergistic promise of enhancing the credibility of evaluation findings (Hall, 2013). Better still, researchers hold that mixed-methods innovations move the field of evaluation toward a more inclusive and socially just process than any mono-method approach (Mertens & Hesse-Biber, 2013).

2.3 The Flipping Pedagogy

Many researchers have categorized the flipping pedagogy as a specific type of blended learning design (Chen et al., 2014; Lai et al., 2021). The blended class features face-to-face classroom instruction and a range of resources and activities supported by online technology (Cheng et al., 2019; O'Flaherty & Phillips, 2015). The flipped and the blended designs share the qualities of active learning and technology support (Chang & Hung, 2019; McNally et al., 2017). Nonetheless, researchers hold that two core features distinguish the flipping approach: knowledge transmission before class and interactive learning in class (Abeysekera & Dawson, 2015; Bishop & Verleger, 2013; Cheng et al., 2019; Låg & Sæle, 2019; Lai et al., 2021; Strelan et al., 2020; van Alten et al., 2019).

2.3.1 **Theoretical Underpinning and Pedagogical Strengths**

Numerous studies have confirmed that, by flipping the traditional lecturing out of the classroom and exposing students to course content and key concepts before class, teachers can utilize student-centred teaching more fully in class (Akçayıra & Akçayır, 2018; Betihavas et al., 2016; Chuang et al., 2018; Elmaadaway, 2018; Hung, 2017; Munir et al., 2018). A student-centred approach values and supports diverse learning styles where students are active and responsible learners (Betihavas et al., 2016), and teachers are facilitators of deep learning (Hung, 2017). In their review of flipping classroom approaches, Bishop and Verleger (2013) put forward that the flipping pedagogy is based on the theoretical framework of student-centred learning theory, which embodies learning theories of active learning, peer-assisted learning and collaborative learning. At the heart of the student-centred learning theory is active learning. Abundant research has supported that the flipping pedagogy involves students in active learning (Hew et al., 2021; Låg & Sæle, 2019), contributing to students' increased emotional, cognitive and behavioural engagement (Elmaadaway, 2018; Lai et al., 2021) and higher order thinking capacity (Chuang et al., 2018), and therefore brings positive learning outcomes. Akçayıra and Akçayır (2018) did a large-scale systematic review of the advantages and challenges of flipped classrooms. They conclude that flipped classrooms bring opportunities for peer-assisted learning and collaborative learning, leading to improved teamwork abilities, better social and communicative skills, and shared understanding, in addition to better classroom engagement and deeper learning. Munir et al. (2018) add that students tend to

take more initiative, put in more effort, and handle more complicated learning tasks when working with their peers.

2.3.2 Factors Affecting the Effects of Flipped Classrooms

Research shows a variety of factors affect the effectiveness of a flipped classroom: the extent of flipping, class design, course languages, learners' motivation, beliefs, ages, genders, class time and class size (Chuang et al., 2018; Lombardini et al., 2018; McNally et al., 2017; Strelan et al., 2020; van Alten et al., 2019). Strelan et al. (2020) conclude from their recent meta-analysis of flipping effects on student performance that active learning and problem-solving are the two primary contributing factors of student enhanced performance in flipped classrooms. Meanwhile, research results also indicate equal learning outcomes between traditional and flipped classrooms when both are equipped with active learning activities (Adams et al., 2016; Jensen et al., 2015). These findings raise the question of what it is that brings about the positive learning outcomes, flipped practice or active learning? Some scholars argue that even though the flipping approach can at least do as much to students' academic achievement as does the traditional pedagogy, it is far from valid to advocate the flipping approach considering the vast amount of time and effort and the substantial changes involved (McNally et al., 2017). These arguments point to the need for evaluative studies to verify the value and worth of the flipped practices, which are in scarcity in the existing flipped practice literature (Betihavas et al., 2016; Låg & Sæle, 2019; O'Flaherty & Phillips, 2015; Stöhr et al., 2020).

Researchers support that learning activities in flipped classrooms serve as a vehicle for student-centred active learning (Al-Zahrani, 2015; Bishop & Verleger, 2013; Chang & Lin, 2019). It is commonly held that flexible pre-class learning activities prepare students for better classroom learning (Elmaadaway, 2018) by boosting their active participation in classroom activities and enhancing their interactions with peers and teachers (Hung, 2017). Chuang et al. (2018) suggest that pre-class exposure to lecture content is crucial in guaranteeing students' success in classroom participation. In class, multiple learning activities like discussion, feedback, problem-solving, and group work involve students in active learning (Elmaadaway, 2018).

While a wide range of studies acknowledges that learning activities in flipped classrooms are essential components of enhancing student engagement, they are far less researched than other components in the flipping pedagogy (Stöhr et al., 2020). Meanwhile, they are the most controversial part related to student satisfaction (Akçayıra & Akçayır, 2018; Chuang et al., 2018). Students are mostly reported as not satisfied with the class structure that orientates them to their learning tasks (Al-Zahrani, 2015; Elmaadaway, 2018). Studies endorse that increased workload, self-regulation, and demanding activity tasks can reduce student satisfaction (McNally et al., 2017). Nevertheless, some research has noted changes in student attitudes from apparent resistance at the beginning of the pedagogical change to acceptance in the final stage of the semester (Betihavas et al., 2016; Munir et al., 2018), suggesting that students' initial dissatisfaction may result from changes in the learning habit. As a result, scholars urge that learning activities, especially the collaborative ones which

involve a variety of factors, such as task complexity, personal expertise, individual contribution and personality, shall be carefully designed and managed in flipped classrooms (Betihavas et al., 2016; Chuang et al., 2018). More research dedicated to learning activities in flipped classrooms is needed to shed light on how to improve the design and implementation of these activities.

2.3.3 **Flipping Pedagogy in EFL**

The flipping pedagogy, which features active and deep learning through multiple learning activities (Jensen et al., 2015; Lombardini et al., 2018), is receiving wider support from language educators (Chen Hsieh et al., 2017; Turan & Akdag-Cimen, 2020). Effective foreign language learning takes place by exposing sufficiently to quality language input, applying rules in practice, making sense of the activities, interacting actively with others (Spolsky, 2000) and having more learner autonomy, choice and responsibility (Ellis, 2005). Hung (2017) elucidates the applicability of flipping pedagogy in the EFL context through the theoretical lens of second language acquisition (SLA). She argues that the flipping pedagogy is conducive to lowering the affective filters of second language (L2) learners, maximizing their interactive opportunities in the target language during class time, and developing learners' L2 fluency through communicative practices from careful instructional design. Turan and Akdag-Cimen (2020), in their systematic literature review of flipped practices in English language teaching, conclude that the majority of the reviewed studies support positive language learning outcomes in flipped classrooms, confirming the applicability of the flipping approach in language classrooms in practice. Their

review further suggests that language output abilities have been the most commonly studied language skills. It also reveals that increased workload, a problem common to the flipping pedagogy, and learning anxiety, a challenge typical of foreign language classes, are the most often reported in the existing literature on flipped EFL classrooms. The extant literature further indicates that in a sea of studies on flipped classrooms, relatively fewer are conducted in the EFL context (Hung, 2017; Turan & Akdag-Cimen, 2020). In those limited studies, most focus on learning outcomes rather than the process. Therefore, more empirical studies are needed to add to our knowledge about what works for English language learners and under what conditions of flipped learning environments.

2.4 Student Engagement

Student engagement is widely recognized as an important influence on learning and a critical pathway to success (Kahu, 2013; Kuh, 2009; Reeve, 2013; Wilson et al., 2022; Xerri et al., 2017). The concept of student engagement in HE is generally defined in two ways: broadly, student engagement conflates both the time and energy students invest in educationally purposeful activities and the effort institutions devote to using effective educational practices (Kahu, 2013; Wilson et al., 2022; Xerri et al., 2017; Yu et al., 2018; Zhoc et al., 2018). More specifically defined and widely adopted is the psychological perspective that focuses on students' affective commitment to, cognitive investment, and behavioural participation in learning (Kahu, 2013; Kahu & Nelson, 2018; Yu et al., 2018; Zhoc et al., 2018). The extant studies have shown a positive relationship between student engagement and improved learning outcomes,

including academic performance, cognitive and psychological development and generic abilities for personal growth (Kahu & Nelson, 2018; Kuh, 2009; Trowler, 2015; Zhoc et al., 2018). As such, the HE sector always uses student engagement as a proxy for quality (Kahu, 2013) and has a particular focus on enhancing student engagement to maximize the effectiveness of teaching and learning practices. Alongside the emphasis on student engagement comes the increasing effort to evaluate and improve the quantity and quality of student engagement in the HE sector (Wilson et al., 2022).

2.4.1 **Challenges in Measuring Student Engagement**

Student engagement has so far been widely theorized and researched. Researchers consistently hold that student engagement is a multi-dimensional construct (Kahu & Nelson, 2018; Zhoc et al., 2018). Studies on student engagement always approach the issue under investigation with multi-dimensional frameworks. The following dimensions are frequently researched in the existing literature.

The behavioural dimension is the most widely accepted view of student engagement in HE literature (Kahu, 2013). It emphasizes, in particular, the relationships between teaching practices and student behaviour that contribute to high-quality learning outcomes (Kuh, 2009; Lai, 2021; Zhoc et al., 2018). The cognitive dimension of engagement focuses on students' psychological investment in learning, understanding and mastering knowledge to go beyond the minimal requirement of study (Krause & Coates, 2008; Zhoc et al., 2018). Similar to cognitive engagement is the dimension of academic engagement,

which refers to observable behaviour directly related to the learning process (Zhoc et al., 2018). Unlike the cognitive dimension, the academic dimension of engagement focuses on behaviours essential to achieving the minimum threshold level of learning. Research into the affective dimension of engagement has examined aspects like students' sense of belonging, identification with school, sense of relatedness, self-efficacy and well-being, which are influential to students' motivation, participatory behaviour and learning outcomes (Yu et al., 2018; Zhoc et al., 2018). Research into the dimension of social engagement mainly focuses on student interactions with faculty members and peer students, which abundant research evidence has supported as crucial to improving student learning and development (Chen et al., 2021; Kuh et al., 1997; Zhoc et al., 2018).

It should be noted that each of these dimensions, though having its focus, is by no means isolated but somewhat closely interrelated. Existing studies into student engagement have used frameworks with different components of these dimensions. The inconsistency in the inclusion of dimensions suggests that the current construct of student engagement is far from clearly specified (Wilson et al., 2022). Kahu (2013) points out that the construct of engagement overlaps not only with those of motivation and learning approaches but between the different dimensions within the construct itself. The problems of conceptualization have led to inconsistencies in measurement in current research, so measures based on more precise operational definitions are in need.

Researchers have consistently recognized student engagement as a critical mediating mechanism that explains how learning contexts influence learning outcomes (Kahu, 2013; Kahu & Nelson, 2018). Studies have found that learner characteristics, such as gender, learning experiences, family background, cultural values (Zhoc et al., 2018), learning attitude, and self-efficacy (Kahu & Nelson, 2018), impact student engagement. Studies further indicate that these individual characteristics interplay with contextual factors, such as teaching practices (Xerri et al., 2017), institutional actions, national policies (Trowler, 2015) and social values (Kahu, 2013), all mediating student engagement in learning. However, as the concept of student engagement is still fuzzy (Kuh, 2009; Wilson et al., 2022), and learning contexts are always messy, student engagement is still a black box too complex to measure or map all of its properties. Despite the proliferation of studies on student engagement, we still do not fully understand the complex ways individual and contextual factors interact to influence engagement. Consequently, more research is needed to explore how to appropriately evaluate student engagement.

2.4.2 **Kahu's Conceptual Framework of Student Engagement**

Kahu's work and conceptual framework is one of the most widely acknowledged and cited in a plethora of literature on student engagement. Kahu (2013) attributes the existing conceptualization problems of engagement to a lack of distinction between the state of engagement, its contributors and its outcomes. By drawing on four distinct perspectives on engagement in the existing literature - the behavioural, psychological, socio-cultural and holistic perspectives - Kahu (2013) proposes a six-element conceptual framework on

student engagement that incorporates the socio-cultural context, the structural and the psychosocial influences, student engagement, and the proximal and the distal consequences. This framework is known for its integrativeness (Wilson et al., 2022; Xerri et al., 2017). The framework has student engagement at its centre and depicts it from the psychological perspective of engagement. Recent studies support that the affective, cognitive, and behavioural dimensions can adequately capture the psychological state of student engagement (Kahu, 2013; Lai, 2021; Reeve, 2013; Yu et al., 2018). This individual psychological experience is embedded within the socio-cultural context, influenced directly by students' psychosocial characteristics and mediated by institutional structural features, leading up to proximal and distal consequences of academic success and personal growth.

The key strength of this framework is that it acknowledges the unique nature of individual experiences and highlights that student engagement is more situational than static. By delineating the state of being engaged from its contributors and outcomes and depicting the complex array of factors influencing student engagement within the wider socio-cultural context, the framework facilitates investigation into the factors that hinder or enhance student engagement. The concept of "educational interface" (Kahu & Nelson, 2018, p. 1) further illuminates how and where the contributing factors interact and impact the underlying psychological mechanism that influences learning outcomes. Kahu contends that the framework is not only valuable for guiding in-depth research on engagement but useful for targeting interventions to enhance student engagement and informing the design and implementation of curricular

initiatives (Kahu, 2013; Kahu & Nelson, 2018). Wilson et al. (2022) support Kahu's claim with their review study and further conclude that Kahu's framework is powerful in assessing the contributors to student engagement. However, after evaluating four frameworks for student engagement in practice, they point out that Kahu's framework is relatively limited in measuring the engagement process (Wilson et al., 2022), indicating a need for additional indicators to investigate the psychological mechanism of student engagement effectively. Additionally, researchers maintain that to achieve the full potential of Kahu's framework, rich and in-depth qualitative data and analysis are required (Kahu, 2013; Wilson et al., 2022). Despite its wide recognition, the framework, so far, still lacks for empirical evidence to verify its effectiveness.

2.5 Summary

Course evaluation has been a norm in the current HE sector. It aims to inform evidence-based improvement for teaching and learning practices in a course. The formative course evaluation is particularly strong for this developmental purpose due to its use of qualitative methods to collect context-specific data. However, course evaluation with a qualitative approach is less used and researched in practice, indicating a need for more studies in this field.

The flipping pedagogy has been winning global popularity in the current HE sector because it is believed to be student-centred and, therefore, to enhance student engagement in learning. The flipping pedagogy has been proven conducive to EFL classrooms for its emphasis on active and interactive learning. Among the various factors that research shows to impact the

effectiveness of a flipped classroom, the factor of learning activities is admittedly crucial but has not received adequate research attention so far. Hence, this evaluative project focuses on learning activities in the flipped EFL class.

Student engagement, an essential indicator of learning and a proxy for quality, has become a focus in course evaluation. However, because of its overarching and multifaceted nature, it is challenging to measure all of its properties. Kahu's integrative conceptual framework for student engagement is widely recognized as valuable for guiding in-depth research on student engagement and interventions that aim to enhance student engagement. Kahu's framework is powerful in assessing the contributors of engagement. Therefore, it was employed to frame the discussion of the factors influencing student engagement in this project. Meanwhile, the weaknesses of Kahu's framework in measuring the engagement process lay the ground for bringing in an additional evaluative framework to measure the state of student engagement, which is to be elaborated on in the Methodology Chapter. Last but not least, it should also be noted that all the related literature reviewed in this chapter points to the necessity of bringing in qualitative data to probe into student learning experiences, which led up to the mixed-methods design of this project.

Chapter 3: Methodology

3.1 A Mixed-Methods Approach

This mixed-methods case study aims to evaluate the effects of the learning activities on student engagement and the subsequent academic outcomes in a flipped EFL class. As HEIs have been placing growing importance on the quality agenda, evaluation of educational outcomes is playing an increasingly important role in HE and has become a focus of discussion in recent literature (Saunders et al., 2011; Tight, 2012) and everyday practice.

Course evaluation is one of the most commonly used approaches in HE to measure the effectiveness of teaching and learning practices and consult students with their experiences in learning (Freeman & Dobbins, 2013; Nicolaou & Atkinson, 2019). It is an ongoing and iterative process that improves a course's efficiency and effectiveness by obtaining the right amount of information to ensure valid decision-making that leads to high-quality teaching (Nicolaou & Atkinson, 2019). When performed credibly, it plays a pivotal role in lending insight into the extent of student learning and progress. Many scholars and practitioners hold that the value of course evaluations lies more in providing intelligence for improvement than judging quality (Edström, 2008; Nicolaou & Atkinson, 2019).

The quantitative positivist approach is prevalent in extant research on both flipping pedagogy and course evaluation in HE. Robust quantitative methods produce succinct outcomes of generalized results or predictable trends. However, these findings can be too general for direct application to local

situations or contexts (Johnson & Onwuegbuzie, 2004). Therefore, in this project, the qualitative approach to exploring student engagement in the learning activities was brought in to situate, explain and complement quantitative data.

The underpinning rationale for a mixed-methods design in this project is multi-fold: first of all, a mixed-methods design can yield more valid research results due to the triangulation of methods (Saunders et al., 2011) and the expanded understanding obtained from different data sources (Johnson & Onwuegbuzie, 2004). Research on teaching and learning practice is highly context-specific, and the contextual factors in the educational environment vary tremendously. Qualitative methods enable in-depth and responsive investigations, which can add insight to the findings from the current dominant quantitative study on flipped classrooms and enhance the credibility of the findings of this project. Second, published evaluative studies show that multiple evaluative methods can lead to more complete and constructive responses from the participants because of increased participant satisfaction and voice (Nicolaou & Atkinson, 2019). Students may be more willing to participate in the research process due to different preferences for what constitutes a better course evaluation. When stakeholders buy into the effectiveness and validity of research methods, evaluation use can, in turn, be enhanced (Patton, 2012). Last and most importantly, my pragmatic stance gave incentives to answer the evaluative questions of this project with qualitative and quantitative data together. Patton (2012) warns that the strength of an evaluative investigation comes from

appropriate methods to address the research questions and adequate data collection and analysis techniques.

3.2 Sampling and Participants

The self-evaluation purpose of this project determined that I researched my own classes and recruited the students I was teaching as the participants of this study. The recruitment took place in the autumn semester of 2020 after the Ethical Approval (Appendix One) for this project was issued. In addition to the face-to-face oral introduction and invitation to this project, a Letter of Invitation (Appendix Two) was posted in LMS, explaining briefly to students the research purpose, participants' role in the project and the research methods, along with a Participant Information Sheet (Appendix Three) and a Consent Form (Appendix Four). Students were not recruited as participants unless they signed the Consent Form and sent it back.

The participants of this project were first-year students from a FC (n = 25) and one of the blended classrooms (BCs) (n = 28). I was the lecturer for both classes. BC was used for comparison to understand features typical of FC. Because of the evaluative nature of this project, all students were invited. Participants' demographic information is presented in Table 3.1. Though convenience sampling was used, the sampling strengths are worth mentioning. The first-year students at the university where this project was carried out are randomly assigned to each class, which minimizes selection bias. Additionally, the initial equivalence of the participants in the two classes added to the project's validity (Creswell, 2014). Students in the two classes were of the same

age and had learned English for similar years on average. Students in these two classes were all English majors. All of them had the same courses in that autumn semester; no students were from the other departments or majoring in additional subjects. FC and BC were both using LMS, having the same class schedule, sharing all the learning resources and completing the same learning activities. Except that students in BC displayed statistically significantly higher proficiency in writing in the university-level placement test, the test results indicated no significant difference in all the other language skills and overall English proficiency between the two classes. Hence, factors that led to participants' differences in academic performance between the two classes were minimized, and the effects of learning activities were viable to be attributed to the difference in class design.

	FC (n = 25)	**BC** (n = 28)
Gender		
male	5	8
female	20	20
Age		
M (SD**)**	19.52 (0.64)	19.21 (0.41)
Years learning English		
M (SD)	11.32 (2.14)	10.68 (1.59)
Language proficiency		
listening	14.76 (2.31)	14.42 (2.95)
cloze	17.04 (3.01)	17.07 (4.09)
reading	12.60 (3.03)	12.61 (3.19)
grammar & vocabulary	12.32 (2.46)	12.86 (1.46)
writing*	18.92 (3.70)	21.13 (2.61)
total	75.64 (8.28)	78.09 (9.40)

*Mann-Whitney Wallis test showed a statistically significant difference in writing proficiency between the two classes with a moderate effect size (*U = 219.5, Z = -2.34, p = .020, r = .32)* in the placement test for first-year students.

Table 3. 1 Student demographics in the two classes

It is worth noting that although the participants of this project were English majors, they represented, at most, students of average English competence in

SHUPL due to its distinct disciplinary orientation towards law and politics. The English Department had a lower entry score than the key disciplines did by about 30 points in 2020. Neither was there any additional entry requirement for English-major students in SHUPL. English was not the ideal major for some students because they were transferred to this Department by SHUPL for not being qualified for their ideal majors.

3.3 Epistemology

A pragmatic epistemology underpins this project. Pragmatism is a philosophical method for doing; pragmatists aim to uncover practical knowledge and contextual truth (Biesenthal, 2014). The pragmatist maxim holds that practical knowledge that works in a particular situation is evaluated by its problem-solving capacity and practical consequences in everyday life rather than its universal applicability. Problem-solving incorporates the successful application of concepts, beliefs or theories in a particular situation and is measured by a practical consequence: a satisfactory outcome concerning the problem to be addressed. Truth is defined by its practical use in ongoing experiences and is the result of ongoing inquiries by applying practical knowledge in different situations. Inquiry in service does not establish universal or absolute truth (Hall, 2013). Instead, truth is provisional and instrumental (Biesenthal, 2014), therefore, always subject to fallibility through further human inquiries, leading to larger truths to inform future actions. The pragmatic philosophical stance offers a practical method of inquiry based on iterative actions, oriented to both processes and outcomes and aiming at eliminating doubt (Johnson & Onwuegbuzie, 2004). The process of inquiry taking the pragmatic stance is

inherently rigorous, as theories or the problem-solving ability of a particular truth is constantly tested, then verified or falsified by the practical community of inquiry (Biesenthal, 2014).

The pragmatic philosophy has long been applied in evaluative studies as a problem-solving method that aims to create practical knowledge through scientific inquiry. Corresponding to the pragmatic philosophy, evaluation focuses on doing; it is a field of inquiry that generates questions, seeks answers, examines actions and impact and promotes change (Rosenstein, 2014). The synergy between theory and practice in pragmatic inquiry accomplishes contextual sensitivity and tangible processes for credible evidence to be achieved in evaluative research (Hall, 2013). Pragmatism expands the narrow focus on methodology rigour in traditional evaluation to include a broader view of credibility and validity, addressing both the processes and the outcomes of evaluation (Hall, 2013). Consequential validity is prioritized in pragmatism, which is primarily concerned with warranting evidence and assessing the implications of evaluation findings in concrete situations. This continual, contextually responsive approach enhances the credibility of evaluations. Besides, Deweyan pragmatists advocate that reflection, an added dimension of credibility, be employed to understand how problems can take on new meanings in the larger context by continuously reflecting on evaluation practices and their consequences. In addition, evaluation taking a pragmatic stance aims to promote democracy (Hall, 2013). This means evaluators are purposeful in sharing information, taking collaborative action to meet human needs, and reflecting on how their evaluative endeavours address power

dynamics and the interest of the less empowered. These democratic aims of pragmatism oblige evaluation to be inclusive and responsive.

Dewey's works contributed, to a great extent, to the long-established position of pragmatism in education and related research (Johnson & Onwuegbuzie, 2004). Dewey is distinguished from other classic pragmatists by his educational philosophy and his ethics of democracy. For Dewey, democracy is learned by living it. Education has its social function of achieving democracy. Schooling for democracy is central to social reconstruction dedicated to the continual betterment of humanity. Schooling should cultivate in young people the inclination and ability to engage in collaborative social problem-solving using the logic of scientific inquiry. Pedagogically, schools should be organized as a "miniature community and an embryonic society" permeated with "the spirit of social cooperation and community", where students can develop in the direction of "social capacity and service" by working on small-scale inquiries and community-building activities side by side cooperatively (Harkavy & Puckett, 2014). In the same vein, Dewey, holding firmly that the benefits and privileges of democracy should be extended to every member, advocates democratic dialogue and collaborative decision-making when conducting scientific research.

Pragmatism is "the philosophical partner for mixed methods research" (Johnson & Onwuegbuzie, 2004). The pragmatic stance places high regard for warranted evidence by recognizing both the existence of natural reality and the influence of the emergent social and psychological world (Johnson & Onwuegbuzie, 2004). By using the "both-and" logic and finding a synthetic or balanced

approach to the poles of dualisms, pragmatism justifies the use of multiple approaches that are compatible with different sets of philosophical assumptions. Hence, compared with the mixed-methods design based on other prevailing research philosophies, such as critical realism and post-positivism, the mixed-methods research relying on the pragmatic stance is more flexible and creative, more likely to avoid overlapping weaknesses intrinsic in each research method and more able to address existing complexity (Johnson et al., 2016).

My pragmatic position led to the mixed-methods design of this evaluative project, which aims to have a holistic understanding of what works effectively for learning activities in a flipped EFL classroom and maximize the utility of evaluation findings. In this evaluative project, the pragmatic epistemology sheds light on how evaluation methods could be mixed fruitfully by adopting a practical, context-based and problem-solving method of inquiry, thus allowing the best opportunities to answer important evaluation questions. Meanwhile, it allows due importance to be attached to student voices, enabling a deeper understanding of the evaluands and a more credible evaluation process, which, in turn, contributes to warranted decisions for improvement (Patton, 2012; Rosenstein, 2014).

3.4 Evaluation Framework

An adapted Kirkpatrick Model (Kirkpatrick & Kirkpatrick, 2006) was employed as the evaluation framework for this project. The Kirkpatrick Model involves measuring both the learning process and the outcomes, which aligns with my

pragmatic stance that student experiences and their learning outcomes can by no means be understood separately. Better still, the Model makes up for the weakness of Kahu's engagement framework by offering flexible and context-specific indicators denoting the engagement process. Both strengths are the incentives for its adoption as the evaluation framework for this project. The Kirkpatrick Model was multifunctional in this project. It served to create evaluation plans, clarify criteria for learning effects, define their indicators accordingly, inform the formulation of RQs and guide the qualitative data coding for this project.

3.4.1 The Classic Kirkpatrick Model

The Kirkpatrick Model is an evaluation framework for assessing the effectiveness of organizational training programmes. Known for its systematicity and applicability, the Model has been widely applied in different contexts for different levels of evaluation, including HE contexts (Kirkpatrick & Kirkpatrick, 2006; Paull et al., 2016; Praslova, 2010; Taras et al., 2013). Over six decades since it was first introduced (Kirkpatrick, 1996), the original Model has been amended, modified and developed by either the author of the Model himself or other researchers (Holton, 1996; Kirkpatrick & Kirkpatrick, 2006). Nonetheless, the original four levels of criteria have remained widely used and kept finding new applications in additional contexts (Praslova, 2010). Table 3.2 gives a brief overview of the classic Kirkpatrick four-level Model.

Level	Description
1. reaction	measures participants' feelings about and perceptions of the training programme.
2. learning	measures knowledge learned, skills developed, or attitudes changed due to training.
3. behaviour	measures changes in behaviour due to knowledge, skill or attitude transfer.
4. results	measures the desired final results that occur because the participants attend the training.

Table 3. 2 Overview of the Kirkpatrick four-level evaluation model

These four levels of criteria offer a straightforward guideline for evaluation practice. Effective indicators available in a particular training context can be mapped onto each corresponding criterion. It realizes the alignment between evaluative criteria and their specific indicators, ensures effective data collection for measurement and satisfies the differentiated need in each programme. The extant literature on the Model notes that causal links do not necessarily exist between two adjacent levels (Holton, 1996; Kirkpatrick & Kirkpatrick, 2006; Praslova, 2010). However, Kirkpatrick himself warns that each level is essential and impacts the next level, so skipping straight to the later levels without examining the previous ones in sequence can lead to wrong conclusions of the evaluation results (Kirkpatrick, 1996; Kirkpatrick & Kirkpatrick, 2006). Except that the levels need examining in sequence, the Model is flexible in that evaluators can decide the number of levels they are to measure and choose the most effective indicators available to them for each criterion following their particular evaluation purposes and contexts. The extant literature shows that the first two levels are assessed more often in practice (Praslova, 2010). It may be partly due to their relative ease of administration, as Kirkpatrick testifies that

level 3 and level 4 are more complicated and challenging to measure (Kirkpatrick & Kirkpatrick, 2006).

3.4.2 The Adapted Kirkpatrick Model

The Kirkpatrick Model in organizational settings has been proven to have clear parallels in the HE context (Praslova, 2010). The *reaction* level primarily measures student affective experiences, and the *learning* level measures learning outcomes, which are the conventional foci of evaluation practice in HEIs (Nguyen & Foster, 2018). Better still, the four levels in the Model go beyond measuring immediate effects of student class work (level 1 & 2) to include outcomes of transfer (level 3) and impact of learning (level 4) in evaluation. Such multi-level evaluation provides rich and well-grounded evidence that enhances HEIs' educational endeavours (Paull et al., 2016; Praslova, 2010).

The Kirkpatrick Model was adopted as the evaluation framework for this project in that it aligned with the project purpose to measure both the process and the results of the flipping intervention. The first three evaluation levels correspond with the key psychological perspectives of student engagement. The *reaction*, *learning* and *behaviour* criteria in the Kirkpatrick Model match the *affective*, *cognitive* and *behavioural* dimensions of the psychological state of student engagement, which is the central element of Kahu's integrative engagement framework (Kahu, 2013). The *results* criteria, the fourth level in the Model, measure student enhanced performance in learning, the primary indicators of learning effectiveness. This level corresponds with the consequence elements

in Kahu's engagement framework. Thus, both the engagement process and the pedagogical outcomes of flipped classrooms, that is, student experiences, subject content learned, language and learning skills developed, and cooperative and collaborative behaviour established (Akçayıra & Akçayır, 2018) managed to be measured within the Kirkpatrick Model.

The Kirkpatrick four-level framework was adapted to this project's particular setting and circumstances. The adapted Model in this project contained the same four levels as the classic Model but with modified indicators and instruments catering to this evaluative project. The following were adjustments made in this project.

The data collection methods were the foremost innovation of the adapted Model. The original Model dominantly collects quantitative data to measure outcomes, which are, however, too general to serve this project's formative course evaluation purposes. Therefore, qualitative data were added to this project to help understand students' situated practice in the flipping context and generate more credible evidence for improvement. Specifically, data for the first three levels, the *reaction*, *learning* and *behaviour* levels, were collected via the qualitative approach. Wilson et al. (2022) argue, in their evaluation of frameworks for student engagement, that rich qualitative data help capture myriad forms of student engagement. That leads to an in-depth understanding of how different student groups are affected and helps realize the full potential of evaluation.

Whilst Kirkpatrick (Kirkpatrick & Kirkpatrick, 2006) stresses that participants' positive reaction is vital to the success of a programme, and in practice, most evaluation programmes focus on how well training is received, the *reaction* level in this project collected both positive and negative reactions from the participants. This is done for the improvement purpose of the project. For a course to be improved, drawbacks in the course design and implementation must be detected timely so that amendments and modifications can be made early (Nguyen & Foster, 2018; Steyn et al., 2019). Therefore, negative feedback from the participants is as essential as their favourable comments for this project.

The indicators in the *learning* level and *behaviour* level were sorted out from the extant flipped practice literature. The primary indicators in these two levels included students' knowledge learned, skills developed, attitudes changed, and behaviour established due to their participation in the learning activities. These indicators are unanimously held as essential indicators of effective EFL flipping teaching by researchers of a flipping pedagogy. The *behaviour* level focused on executing the knowledge and skills learned at the learning level. This level is important because the final results of a project depend primarily upon the extent to which the knowledge and skills learned are executed. In the flipped classroom, learning behaviours of self-regulation, interaction and collaboration are both contributors to and indicators of effective learning. However, the *behaviour* level is the most difficult to measure among the four criteria (Kirkpatrick & Kirkpatrick, 2006), as factors that impact the transfer of knowledge and skills learned in classrooms to on-the-job behaviour are many

and various. Some factors, such as time and climate for the expected behaviour, the opportunities available for behaviour changes to be measured and human inertia, are challenging to control, but all impact the results (Kirkpatrick, 1996; Kirkpatrick & Kirkpatrick, 2006). So, as Kirkpatrick and Kirkpatrick (2006, p. 59) suggest, when evaluating the *behaviour* level, the dictum that “something beats nothing” was applied.

The *results* level of this project was measured by the participants’ enhanced engagement in the learning activities and their improved language proficiency in reading and writing. The nature of this project as a course evaluation decided that participants’ performance in the learning activities and the final examination were the most relevant results that could be measured within the time limit of this project. When this evaluative project was carried out, no standardized tests were administered to measure participants’ listening and speaking proficiency at the end of the semester. So, data on these two skills were not available. Admittedly, the highly complex HE context makes it difficult to determine how much participants’ improved language proficiency should be attributed to the flipping design compared with other factors. Kirkpatrick’s suggestion here is to make the best of the evidence if there is no direct proof (Kirkpatrick, 1996; Kirkpatrick & Kirkpatrick, 2006).

It also should be noted that the four Kirkpatrick Model levels are reported as lacking causal linkage (Holton, 1996; Praslova, 2010). That said, such weakness did not have much impact on the validity of the findings of this project. The primary research purpose of this evaluative project was to make evidence-based improvements by exploring and understanding students’

enhanced engagement in learning activities in the flipped EFL class. To confirm causal relations was not the purpose of this project. Bazeley (2013) approves that understanding human experience is a matter more of chronologies than of causes and effects. The four levels in this project represent “a sequence of ways to evaluate the programme” (Kirkpatrick & Kirkpatrick, 2006, p. 21) so that the evaluation results can be interpreted with more precision. The data collected via qualitative and quantitative approaches suffice to generate information to understand the practical consequences (Hall, 2013) and provide evidence for improvement in the flipped EFL class. A brief overview of the adapted Kirkpatrick Model for this project, with sample indicators listed for each level, is presented in Table 3.3.

level	**description**	**sample indicators in IEC**
1. reaction	Student affective reactions to learning activities	student feelings about, attitudes towards and perceptions of the learning activities
2. learning	Students’ knowledge learned, skills developed, or attitudes changed due to their participation in learning activities	language knowledge learned; language skills developed; higher order thinking skills developed; effective learning skills developed; attitudes changed towards learning
3. behaviour	Student behavioural changes due to participation in learning activities	self-regulation, interaction, collaboration in learning
4. results	The desired final results that occur because students take part in learning	students’ enhanced engagement in the learning activities; students’ improved reading and writing proficiency in English

Table 3. 3 Overview of the adapted Kirkpatrick Model for this project

3.5 Research Questions

Aside from offering evaluation criteria, the Kirkpatrick Model informs the formulation of the RQs of this project. The first two main RQs and their sub-RQs were formulated under the four levels of the Model. Table 3.4 summarizes the main RQs and their sub-RQs for this project.

RQs & sub-RQs
RQ1. To what extent are students in FC engaged in the learning activities?
RQ 1.1 To what extent do students in FC react affectively to the learning activities?
RQ 1.2 To what extent do students in FC perceive they learned from the learning activities?
RQ 1.3 To what extent are students in FC engaged in the learning activities behaviourally?
RQ 2. What are the overall effects of the learning activities on student engagement in the learning process and learning outcomes in FC?
RQ 2.1 To what extent is student engagement in the learning activities enhanced in FC?
RQ 2.2 To what extent are students' learning outcomes enhanced in FC?
RQ 3. What affects student engagement in the learning activities in FC?
RQ 4. What are the reflections and areas for improvement regarding flipped EFL classrooms and formative course evaluation?
RQ 4.1 What are the critical reflections on the design and implementation of flipped EFL classrooms?
RQ 4.2 What are the critical reflections on formative course evaluation of flipped classrooms?

Table 3. 4 Summary of RQs and sub-RQs

3.6 Research Context

When discussing self-evaluative studies, Bamber (2011a, p. 196) emphasizes the importance of context as it "provides important data and evidence which, if ignored, can skew the evaluative outcome". Understanding the project in the context where it is carried out produces a more reliable understanding and

interpretation of its findings and increases its case-to-case transferability. Pragmatists assert reality is nature and experience in transaction, where “experiencing organisms interacting in environments” (Johnson et al., 2016, p. 268). In the HE context, these environments may involve the “Big Culture” of the society as a whole, resources, power relations and technological development (Trowler, 2016). The environments directly related to this project are the university where this project is carried out and the LMS that supports the course under investigation.

3.6.1 **Setting**

This single-sited case study was carried out at Shanghai University of Political Science and Law (SHUPL), where I am based as an EFL lecturer. SHUPL is a profession-oriented municipal university preparing professionals in laws and politics. In 2016, with the nationwide English education reform (briefly introduced in section 1.1), SHUPL decided its English curriculum should be law-oriented and started to cut down credits for EGP courses to make space for EAP/ESP courses. A direct result of this move was reduced class hours for EGP courses. However, under the prevalent assumption that high schools were producing students with higher English proficiency, the curriculum objectives remained unchanged or became even higher.

Alongside the nationwide credit cut-down of EGP courses, the flipping pedagogy was gaining increasing scholarly recognition in HEIs worldwide simultaneously (Akçayıra & Akçayır, 2018; Cheng et al., 2019). Under the attractive idea of recovering class time by flipping classes, teachers for EGP

courses in SHUPL took the initiative to move part of the subject content and learning activities outside classes. Online social media, like WeChat, were used to deliver course materials and carry out learning activities such as group discussions out of class. However, back then, teachers did not know much about how students dealt with learning activities outside class, as public applications did not provide the function of tracking student performance. It was not until the nationwide online education provision in the spring semester of 2020 that an online LMS was made available to all the teaching staff in SHUPL due to the COVID pandemic lockdown. As one of its crisis management measures, SHUPL encouraged all its teaching staff to move their courses into the online LMS even after school reopened and classes were delivered face-to-face.

3.6.2 **Online LMS**

SHUPL joined Chaoxing, an LMS for HEIs nationwide, to support its online delivery of classes during the COVID lockdown when face-to-face classroom teaching was impossible. Unlike the other public online teaching platforms available to individual teachers, Chaoxing is also accessible to educational institutions. It allows its registered institutional members to display their model courses, share their massive open online courses (MOOCs) or small private online courses (SPOCs) and access its digital library resources. Teachers and students of the registered member institutions can log into the LMS via their institutional ID and access these shared online resources. Since the nationwide online education provision, almost all the HEIs in mainland China have been registered members of Chaoxing. All learning and teaching practices via

Chaoxing can be tracked in real-time. For administrative and quality control purposes, SHUPL has been encouraging its teaching staff to deliver online classes via Chaoxing in preference to other media.

Teachers can manage almost all teaching activities via Chaoxing. Both recorded lectures and live online classes can be delivered via it. Teaching activities (as shown in Figure 3.1) can be delivered both in-class face-to-face and online out of class synchronously or asynchronously. Chaoxing tracks the duration and frequency of student participation, automatically rates student assignments, and produces activity reports. Thus teachers can obtain immediate feedback on student performance, which facilitates formative evaluation of student performance. For students, Chaoxing serves as a “one-stop shopping” area to access all the materials and learning activities for every course they enrol in.

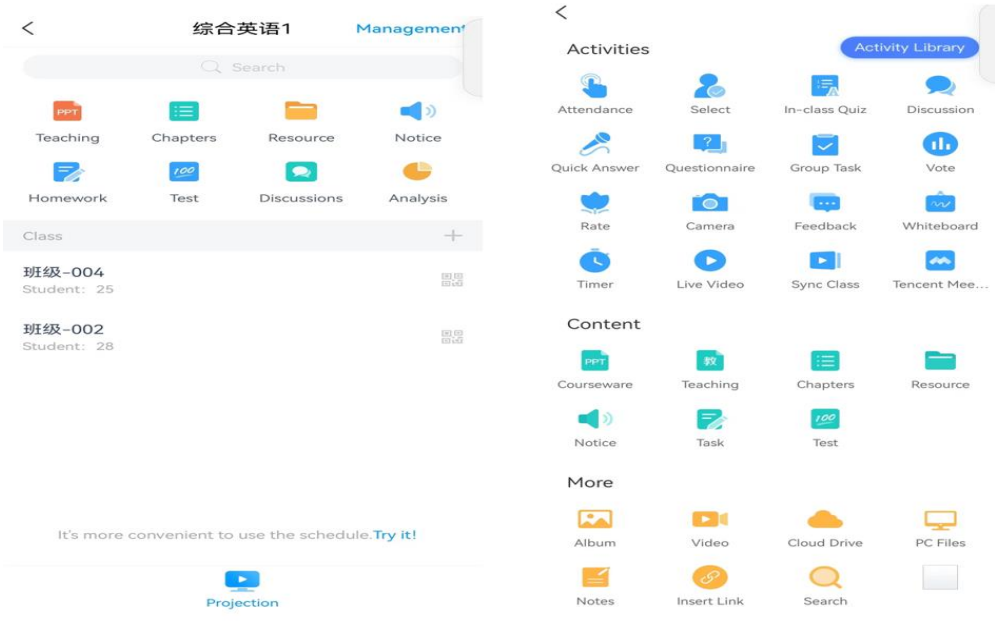

Figure 3. 1 A snapshot of Chaoxing LMS App interface

3.6.3 **IEC and its Learning Activities**

The learning activities evaluated in this project are in an Integrated English course (IEC). IEC is an EGP course. It aims to enhance the first- and second-year students' generic English language skills and prepare them for advanced English courses. It is a degree course, therefore compulsory, for English-major students but meanwhile selective to all non-English majors in SHUPL. The IEC is offered twice a week with a duration of 90 minutes each period and a maximum quota of 30 students in each class. There are altogether 30 periods in a 15-week semester. Figure 3.2 presents a brief overview of the course flow over the semester of this project [1]. Each semester IEC enrols eight classes of students, with four classes of freshmen and four sophomores. During the autumn semester of 2020, when this evaluative project was carried out, the eight IEC classes were taught with different class designs: one in the flipped design, three in the blended design, and the other four via traditional direct instruction. The instructor of each class decided how the class was delivered.

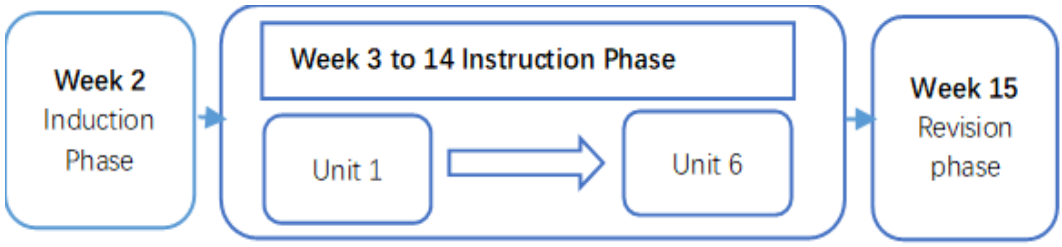

Figure 3. 2 Overview of the course flow

[1] Freshmen students started their first autumn semester a week later than the other students due to the COVID-19 pandemic, so they had a 14-week semester instead of the usual 15 weeks.

Learning activities are both the learning process and the outcomes that can be evaluated in terms of whether the correct or appropriate propositional content has been conveyed (Müller-Hartmann & Ditfurth, 2012). Language learning activities involve receptive or productive, oral or written skills, and various cognitive processes (Ellis, 2003). Accordingly, the learning activities in IEC are used as a pedagogical tool to engage students in learning and entail language learning objectives to develop students' English language and cognitive skills via language use. As such, they can be measured in terms of student engagement and learning outcomes.

The learning activities in the flipped IEC are organized based on the revised taxonomy of cognitive processes proposed by Anderson et al. (2001). Bergmann (2017) claims the diamond model (as shown in Figure 3.4) depicts better how learning activities are organized in flipped classrooms than the pyramid model in Figure 3.3. More importantly, the diamond model serves the purpose of EFL classes: to develop learners' language skills through abundant practice.

As the pyramid model shows in Figure 3.3, in a traditional EFL classroom, most of the class time is devoted to lecturing, where teachers impart language knowledge to students to remember. Under this traditional model, most EFL students in China, having learned English for more than ten years and passed thousands of paper-and-pencil tests, come to university with adequate knowledge of the language yet feel it is challenging to express themselves

clearly. It is because, without sufficient practice in authentic language contexts, language knowledge does not turn into proficiencies effectively (Spolsky, 2000).

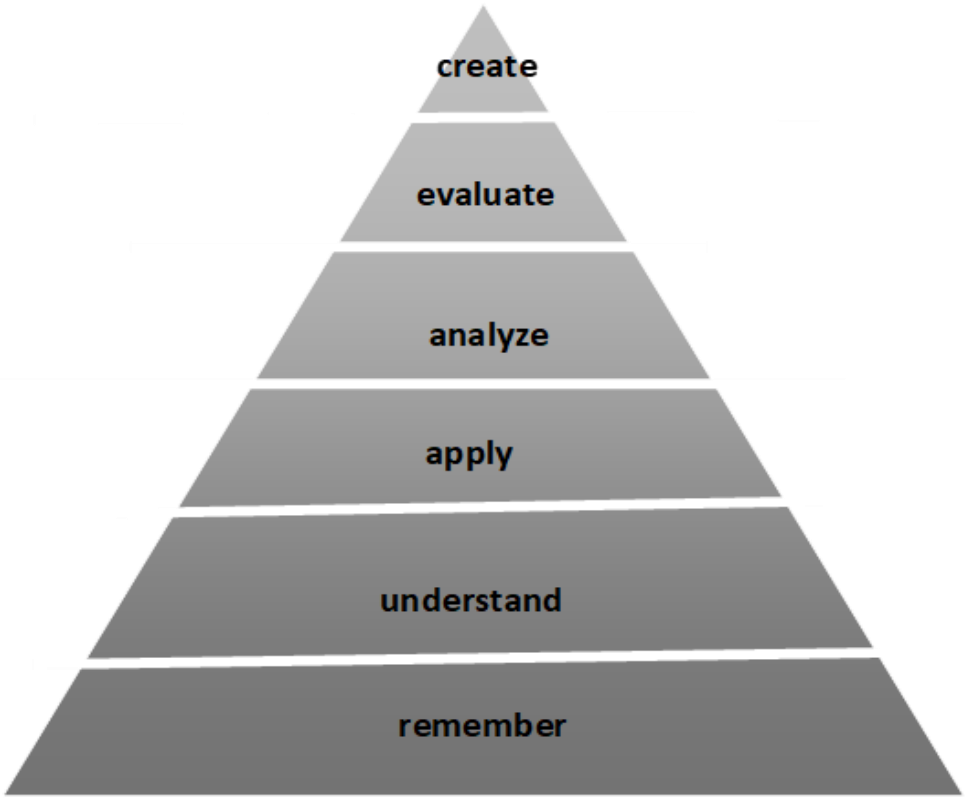

Figure 3. 3 The pyramid model of the cognitive process taxonomy

In contrast, in the flipped EFL class, when language knowledge is lectured via recorded videos before class, most of the class time is freed up for individualized or interactive tasks for students to practise and apply what they have learned, as depicted by the diamond model in Figure 3.4. Better still, more class time can be arranged for learning activities on the top levels of the diamond model to develop students' higher-order thinking capacity. As such, learning goes deeper in a flipped class (Bergmann & Sams, 2014).

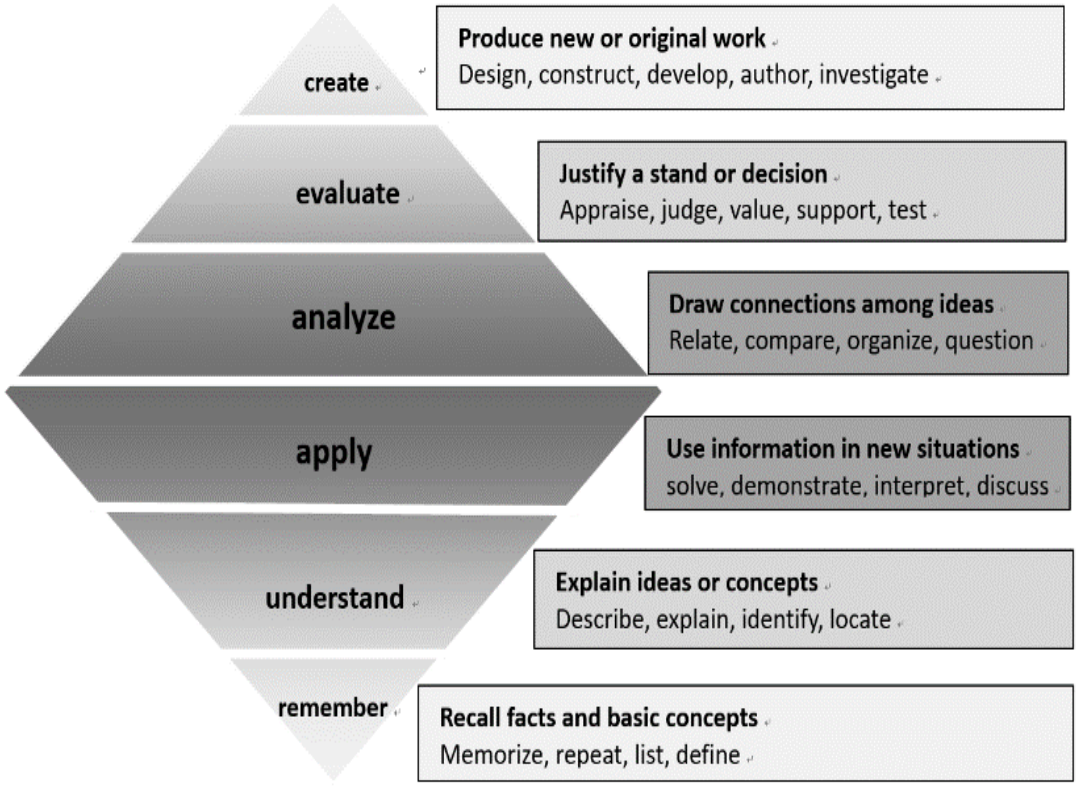

Figure 3. 4 The diamond model of the cognitive process taxonomy and examples of classroom activities

The class design of FC under evaluation in this project is presented in the left column of Figure 3.5. The pre-class learning activities in FC aimed to prepare students for their participation in face-to-face classroom learning. They began with students watching **recorded instructional videos** for that class period. The videos presented students with text-related language knowledge of vocabulary, grammar and writing that they would need to complete learning tasks in class. During a video, students could stop where they were confused and ask questions by writing real-time comments.

A **quiz** followed to check the first two levels of the cognitive process taxonomy, students' memory and understanding of the video content. Quizzes usually consisted of objective items of multiple choice, blank filling or true or false

questions so that they could be rated upon submission, and both the instructor and the students could have immediate feedback. Students were encouraged to share and discuss with classmates their questions from the videos and quizzes in the **online Q & A forum**[2]. It was designed for problem-shooting to further prepare students for participation in classroom learning activities. More importantly, the forum aimed to foster a community of learners where learning could be enhanced through peer tutoring and collaboration.

With the lower levels of the cognitive process addressed before class, classroom time in FC was devoted to developing the higher levels of the cognitive process taxonomy. In-class activities of FC began with **teacher feedback** on the outstanding questions based on students' shared problems from the video, quiz and Q & A forum discussion, then moved on to student **individual/paired work**. This part was mainly hands-on learning activities designed to develop the third and fourth levels of the cognitive processes in the taxonomy. Students were to apply what they learned from the video for translation and paraphrase tasks, and answer questions related to text content. Individual support was available to students in need of help to complete these tasks. For each class period, extended topics deriving from the text were assigned in the online forum. After completing their individual/pair work,

[2] All the forum discussions in IEC were optional, so they are not included in Figure 3.5. Forums in IEC served multiple purposes. In addition to problem shooting, brain-storming, teacher feedback, and peer tutoring, forum discussions were also a time-coordinator in FC, so they were designed as optional to satisfy different student needs.

students were encouraged to feed their ideas on the extended topics into the online forum and respond to peers' postings. This **online forum group discussion** served as brainstorming for the project tasks, as well as a time coordinator to guarantee that all the students were able to join in the following face-to-face group discussion in time.

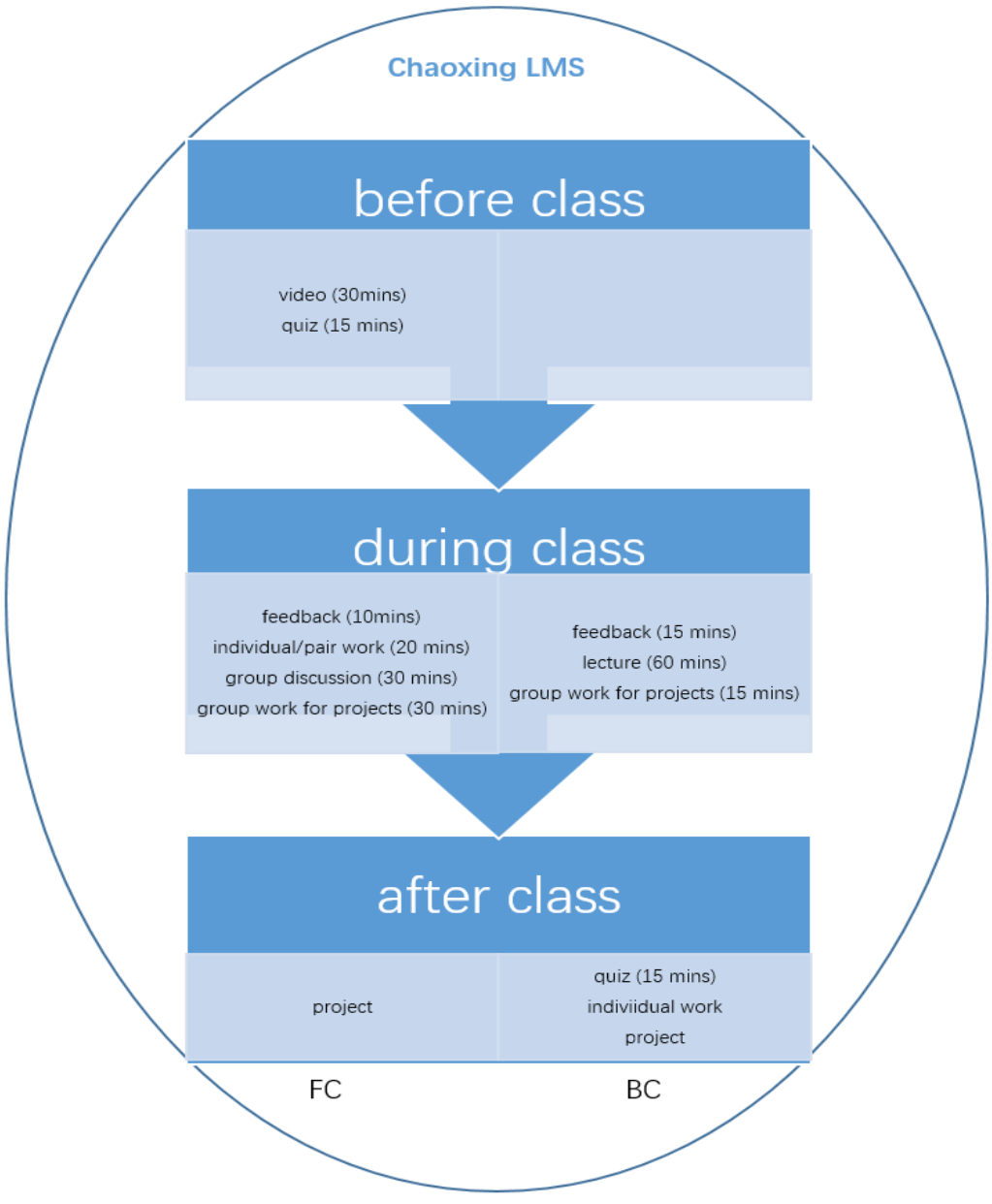

Figure 3. 5 Class design of the flipped IEC & the blended IEC

The last two classroom activities in FC aimed to develop students' higher-order thinking capacity. In the **face-to-face classroom discussion**, students presented and discussed their finished work in class. They were invited to complement, clarify or justify their ideas while the instructor gave on-the-spot comments. Classroom activities in FC ended up with **group work**. At the beginning of each unit, students were assigned a group work project deriving from the theme of its texts. The project tasks aimed to develop students' higher-order thinking skills at the top levels of the diamond model. In FC, students completed most of a project in class, collaborating with group members face-to-face. Teacher support was available. Due to the time limit in class, some projects were extended after class, where students collected materials, completed their share of the task and prepared for in-class collaboration in the next class period. Each project was jointly assessed by the instructor, the other groups and the participants themselves. Student work that got the highest mark in each project was demonstrated to all the students via the online forum.

The course design in BC is presented in the right column of Figure 3.5. BC began its class with teacher feedback on students' shared problems in the quiz, individual work and forum discussions for the previous class period. As students in the BC finished their quizzes and individual work after class, more outstanding questions needed addressing in the classroom feedback session. Therefore, more time was needed for feedback in BC than in FC. Most class time in BC was spent on interactive lectures delivered with Power Points (PPTs), covering both the video content and the topics in group discussion in FC. The last 15 minutes in BC were assigned to group projects. Due to the time

limit, groups could barely work out project plans, divide labour, check progress or make coordination in class. All groups in BC did their projects out of class. Except for class design differences, FC and BC shared all the learning resources, completed the same tasks in the LMS, and had the same class schedule over the semester.

The traditional IEC featured direct instruction with PPTs in class and assignments after class. Each student was also to do a presentation in groups with peers and write essays during the semester. The instructors of the traditional class seldom used LMS. As the traditional classroom has little relevance to this project, it is not presented here.

3.7 Research Methods

This single-site case study adopted a convergent parallel mixed-methods design with nested samples to explore and explain the extent to which learning activities in flipped EFL classrooms enhance student engagement in the learning process and the subsequent learning outcomes.

The research purpose of this project, to evaluate the effects of learning activities on both the learning process and the outcomes, was an incentive for this convergent parallel design (QUAL + QUAN). This design attaches equal importance to the qualitative and quantitative data (Creswell & Plano Clark, 2017). A diagram depicting the design procedures for this project is presented in Figure 3.6. Both datasets were collected simultaneously. The two datasets were first analysed separately, then merged, related and compared side by side. The qualitative findings answered RQ1 and its sub-RQs; the quantitative

findings answered RQ2 and its sub-RQs. Interpretation and inferences were made to answer RQ3 and RQ4. A summary of the RQs and sub-RQs, with their data sources and analysis methods, is presented in Table 3.5.

3.7.1 **The Qualitative Phase**

The qualitative phase sought to uncover students' affective reactions, perceived cognitive gains and behavioural engagement in the learning activities in IEC through the participants' voices. Data for this phase were collected from focus group discussions and participant observation and analysed thematically. The following sections expound on the data collection and analysis processes.

3.7.1.1 Focus Group Discussion

Focus group discussion was administered in this project primarily for its strength in uncovering convergent participant voices (Cousin, 2008). It is believed that interaction with peers of similar status or background in a permissive and non-threatening environment enables participants to give more prominence to their perspectives on the issues discussed, thus highlighting issues of importance. In addition to the data collection purpose, the interactive discussions among participants were used to maximize the process use in evaluation, which refers to the effects brought about by an evaluation while it is being undertaken (Patton, 1998). Divergence and consensus arising from interactions between participants illuminate insight into the issues being discussed (Cousin, 2008; Hennink & Leavy, 2014). Students, therefore, may

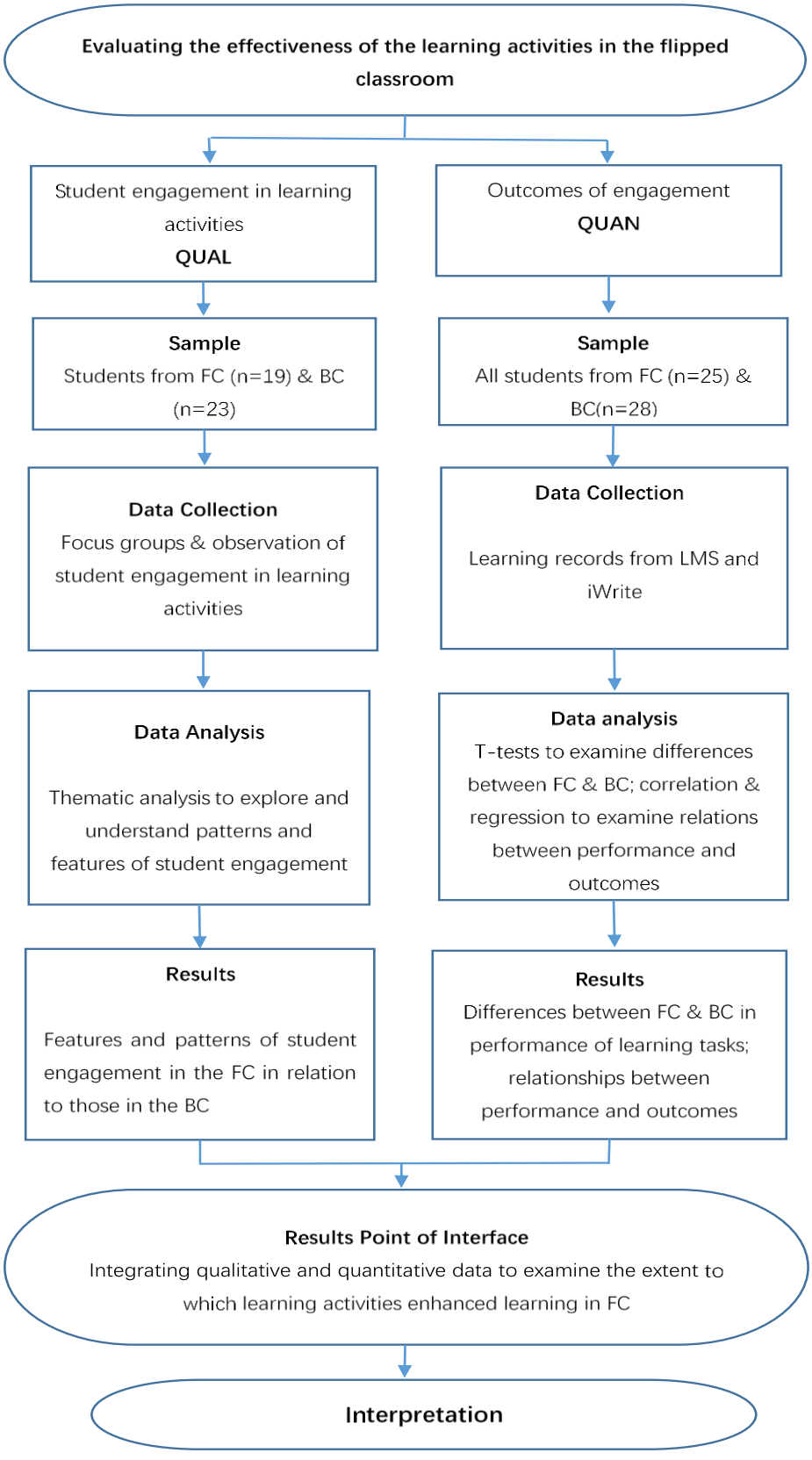

Figure 3. 6 A convergent design of the mixed-methods study of learning activities in FC

RQs & sub-RQs	Data Sources	Data Analysis
RQ1. To what extent are students in FC engaged in the learning activities?	**QUAL**	
RQ1.1 To what extent do students in FC react affectively to the learning activities?	Focus group interviews; Participant observation	Thematic analysis
RQ1.2 To what extent do students in FC perceive they learn from the learning activities?		
RQ 1.3 To what extent are students in FC engaged in learning activities behaviourally?		
RQ 2. What are the overall effects of the learning activities on student engagement in the learning process and learning outcomes in FC?	**QUAN**	
RQ 2.1 To what extent is student engagement in learning activities enhanced in FC?	Learning records of quizzes, assignments, group work and essays from LMS & iWrite; Student scores of final examination	Descriptive analysis; T-tests (or their equivalents for non-parameter data); correlation tests and regression
RQ 2.2 To what extent are students' learning outcomes enhanced in FC?		
RQ 3. What affects student engagement in learning activities in FC?	Interpreted and inferred from **QUAL+QUAN**	
RQ 4. What are the reflections and areas for improvement regarding flipped EFL classrooms and formative course evaluation?	Interpreted and inferred from **QUAL+QUAN**	
RQ 4.1 What are critical reflections on the design and implementation of flipped EFL classrooms?		
RQ 4.2 What are the critical reflections on formative course evaluation of flipped classrooms?		

Table 3. 5 Summary of RQs & sub-RQs, data sources and data analysis

have a better understanding of their performance in the course and clearer goals to work for. Focus groups were also adopted out of ethics and feasibility consideration. Collecting data via dynamic interaction between group members, focus groups are less affected by the interviewer than traditional interviews (Hennink & Leavy, 2014). That reduces the impact of the potential power differentials between the teacher and the students and improves data integrity.

Besides, focus group discussion is more time-efficient than one-on-one interviews; it produces a large amount of data in a relatively short time (Bazeley, 2013; Cohen et al., 2011). Without focus groups, it would have been impossible to hear voices from over 40 participants within one week and twice a semester.

For formative evaluation purposes, focus groups were administered at two time points. Literature on course evaluation supports that multiple time points evaluations bring teachers opportunities to make timely adjustments to improve teaching (Fisher & Miller, 2008; Holland, 2019), motivate student engagement in the course (Sozer et al., 2019) and increase the reliability and validity of measurements (Nguyen & Foster, 2018). The Pre-Focus Group (Pre-FG) was administered during the fifth week of the autumn semester of 2020 after the first two textbook units were completed. The timing was based on the assumption that students had become familiar with the course arrangement by then. The Pre-FG attempted to obtain a baseline understanding of student perception of and engagement in the learning activities and inform timely amendments to drawbacks in the course design. The Post-Focus Group (Post-FG) was carried out during the examination week at the end of the semester. It aimed to explore students' perceived gains from the learning activities and capture changes over the semester. Since participants had zero experience in focus group discussion, a video record of a focus group discussion was presented to them before they signed up so that they briefly understood their role in the discussion.

Different grouping strategies were used for Pre-FG and Post-FG. To discover commonalities and differences across groups through comparison, I grouped the participants with similar accumulated points in LMS in the same group in Pre-Focus. Not all the participants took part in the focus group discussion. Some students signed up but did not turn up as scheduled. The participation rate was high in Pre-FG. Nineteen participants from FC took part in the Pre-FG in four groups, and 23 from BC in five groups. There were 4 to 6 participants in each group. The coding results of the Pre-FG indicated not many differences in student engagement between groups of different academic performance.

Given this reason and the time limit, students were allowed to set up groups by themselves at their convenience in Post-FG. The participants in the Post-FG were those taking part in the Pre-FG but were fewer in number. COVID recurred towards the end of the autumn semester. For health and safety purposes, SHUPL cut down the two-week examination weeks to one and urged all the students to leave the campus as soon as they finished their final examinations. As a result, some participants could not attend the focus groups as scheduled. By the end of the examination week, 15 participants from FC took part in the Post-FG in four groups and nine from BC in two groups. Table 3.6 presents the participant size in the focus group discussions for this project.

Focus groups	Participant number in FC (groups)	Participant number in BC (groups)
Pre-FG	19 (4)	23 (5)
Post-FG	15 (4)	9 (2)

Table 3. 6 Participant numbers in the focus group interview

The Post-FG was closed by abiding by the ethical codes of protecting the participants from health risk (BERA, 2011) and after I ensured that data were saturated. Methodology books suggest that in the practice of qualitative study, data saturation is reached when no new information is being added to the coding categories and no new category is emerging (Bazeley, 2021). I went through BC's Post-FG transcripts right after the discussions and found that the participants in the two groups produced highly repeated views. I then mapped the transcripts onto the coding scheme for this project, and no alien information was found (which confirmed that the different grouping strategies used for Pre- and Post-FGs did not lead to group differences). Literature on qualitative study suggests that 95% of the descriptive categories in an interview are contributed by one-fifth of the cases in a homogeneous group (Bazeley, 2021; Collins & Onwuegbuzie, 2013; Guest et al., 2006) and that focus group discussions produce richer data than a one-on-one interview (Cohen et al., 2011; Hennink & Leavy, 2014). That led to my speculation that one-third of the participants from BC would suffice to produce views and ideas representing the whole class in focus groups. It should be mentioned that the participants turned out more active in their self-made groups and produced richer data. Apparently, familiarity between group members promoted interaction in the Post-FG discussion. However, the coding results show that these differences in responses did not result in group differences.

All the focus group discussions were held in the meeting room across from my office. Disruptions were prevented as far as possible. Every participant wrote their pseudonyms on a folded paper and displayed it in front of them. Prompt

sheets (Appendix Five) developed from the evaluation framework were used in both Pre-FG and Post-FG to ensure that important areas were covered in the discussions (Cousin, 2008). Considering the participants lacked focus group experience, I prepared 4-5 prompt questions for each discussion in case high moderation was needed (Cousin, 2008). The questions were used flexibly during the discussions. The discussions were held in Chinese, the participants' mother tongue, to minimize misunderstanding and maximize chances for the participants to express themselves freely and clearly. I acted as both the moderator and the observer, encouraging the participants to talk, ensuring all voices were equally expressed and taking notes of the crucial issues being discussed. Group discussions in the Pre-FG lasted one to two hours. Most group discussions in the Post-FG lasted about an hour.

At the end of each group discussion, I debriefed the group on the critical areas covered in the discussion and allowed them time to correct or expand on their contributions. I used both my laptop and iFly, a real-time transcription device, to record the group discussions. All the recordings and transcripts in iFly were destroyed and deleted right after I stored and encrypted them on my laptop. For member checking, I presented the transcript to each group member who contributed to it. Each transcript was labelled as "*Class-Focus Group Type-Group Number*". For example, the transcript from FC's first focus group in the Pre-FG discussion was labelled as *FC-Pre-FG-1*.

3.7.1.2 Participant Observation

Participant observation was employed to corroborate data collected from focus group discussions. The unique strength of observation lies in its affordance of live data from naturally occurring social situations (Cohen et al., 2011). In practice, what people do may differ from what they say they do. Observation, in this sense, provides a reality check and assists in understanding human actions. It is also used to uncover unexpressed student voices. Observational data are sensitive to contexts (Cohen et al., 2011). They enable researchers to be aware of the contexts of programmes and see things that might otherwise be unconsciously missed out, such as those students might not talk about freely in an interview. When used in combination with other research methods, observation becomes an essential tool for understanding the “backstage culture” (Cohen et al., 2011, p. 465) and “causation at work” (p. 471), leading to sound explanations and interpretations of the event. As a popular method of data collection in education, observation offers an unobstructed view of student learning practice in a realistic context, which informs “formative and developmental feedback to guide teacher improvement efforts” and illuminates linking mechanisms between classroom processes and desired outcomes (Martinez et al., 2016, p. 15).

Participant observation in this project went on throughout the whole data collection period as unobtrusively as possible so that the participants remained in a natural learning environment. It covered student engagement in learning activities both in class and online, such as their management of learning, their reactions to and performance in learning tasks, the linked change over time,

and any other critical events. For example, I noted down that *students repeatedly made queries about the final exam towards the end of the semester* as a sign of student attitude towards learning that would influence their behaviour. Field notes were kept immediately after each class or any critical event. To increase the reliability of the notes to be taken and make sure all essential factors were included in every piece of note, I used a structured field note form (Appendix Six) adapted from Spradley's checklist of field note content (Spradley, 1980, p.78 cited from Cohen et al., 2011, p.467). This checklist gives prominence to acts and activities, so it serves well my purpose of observing student engagement in learning activities and their reactions. A reflection section was added to the original checklist. As part of the observational data, the field notes were primarily used to assist in explaining and interpreting the data from focus group discussions.

3.7.1.3 Thematic Analysis

Thematic analysis was employed to uncover participants' collective voices, consensus, and diversities in their learning experiences from focus groups. Focus groups and thematic analysis match each other well. Focus group discussions excel in developing themes (Cohen et al., 2011). Thematic analysis helps examine different perspectives and highlight similarities and differences (Braun & Clarke, 2006; Nowell et al., 2017), which are distinctive features of focus group discussions (Barbour, 2013).

A holistic idea of the data was obtained before I set into coding. I took field notes while moderating the group discussions and did debriefing to participants

at the end of each discussion. After the discussion recording was transcribed verbatim automatically, I proofread all the transcripts and added paralinguistic information to each of them according to the recording. By immersing myself in it, I managed to get familiarized with data, which was essential for obtaining the feeling of the whole (Nowell et al., 2017).

A coding scheme was developed, from four transcripts of the focus groups (two from Pre-FG and two from Post-FG), as a foundation for coding and a trail of evidence to guarantee credibility in data analysis (Nowell et al., 2017). Categories and sub-categories in the scheme were either derived deductively from the evaluative framework of the Kirkpatrick Model or generated inductively from the text data. The Model's first three criteria, *affective reaction*, *learning* and *behaviour*, were utilized as higher-order codes to help organize data (Nowell et al., 2017). The fourth criterion in the Model, *results,* was not included in the coding scheme because it was examined with numerical data. The sub-categories under these three criteria were derived from the flipped practice literature or the text data. All the lower-order codes were data-driven. The pre-defined categories work well for comparison purposes (Vaismoradi et al., 2013) and help capture consensus and diversities among students both within FC and between FC and BC. However, Nowell et al. (2017) warn that using pre-defined categories tends to risk losing the richness of data. It should be noted that although I used the Kirkpatrick framework to guide coding, it did not mean I was committed to staying within the framework. Instead, I tried to stay open to the data to capture emerging themes. After establishing the higher-order codes, I coded the transcripts openly, giving full and equal attention to data items and

identifying interesting aspects in the data. Then I mapped the lower-order codes from open coding on the framework to check fitness. Two categories, *improvement* and *factors,* emerged from the data. It is worth mentioning that the coding scheme was not rigidly set but evolved as I further engaged with the data and my understanding developed. For example, the sub-category *previous experience* evolved from “miscellaneous” codes that did not seem to fit into any category (Braun & Clarke, 2006). However, as I engaged with the data further, I came to realize it was an important factor that influenced students’ attitudes towards and behaviour in learning activities in IEC.

Thematic analysis in this project incorporates manifest and latent data analysis, as themes were derived from deductive and inductive categories and matched the research questions. For analytic credibility, I wrote memos for all the codes at each level to identify emerging impressions that may form the basis of themes. I kept adjusting the codes throughout the analysis to ensure each theme was supported coherently and sufficiently by data. The qualitative data were coded with Atlas.ti version 9.

3.7.2 **The Quantitative Phase**

The quantitative phase aimed to understand the overall effects of FC students’ engagement by examining both their effort in the learning activities throughout the semester and the outcomes of their academic performance in the learning tasks and the final examination. Data for this phase were primarily collected from LMS, a common practice for educational evaluative studies (Papamitsiou

& Economides, 2014). Learning records from all the participants in the two classes (n = 53) were collected.

3.7.2.1 Student Learning Records

Quantitative data to examine FC students' enhanced engagement in learning activities and improved academic performance were mainly collected from Chaoxing, the institutional LMS where the learning resources of IEC were presented to students and most learning activities were mediated. Data on essay writing were collected from iWrite. Student learning records of video watching, quizzes, forum discussion, individual exercises, essay writing, group work and the final examination were collected from both classes and compared for evidence of enhanced performance in FC. A summary of the learning activities in IEC and the corresponding learning records, together with the task time and the task types in both classes, is presented in Figure 3.7 for a rapid grasp of the implementation procedures.

Regarding student engagement in **pre-class instructional videos**, the ratio of the actual time a student spends watching a video to the length of the video ($R_{w/l}$) was used as an indicator of engagement. Compared with the video-watching time, which is very much affected by the length of a video, and the access rate, a more valid indicator of disengagement than engagement, according to Ellis et al. (2017), the $R_{w/l}$ tells the extent to which a student completes a video. When $R_{w/l} = 1$, it indicates a student watches a video 100%. $R_{w/l} > 1$ suggests a video is repeated; the larger the $R_{w/l}$, the greater the extent to which a video is repeated. Conversely, $R_{w/l} < 1$ means that a student does not

finish watching a video. Thus, $R_{w/l}$ is more accurate as an indicator of engagement than the time spent on a video or the number of accesses made to a video. All the videos were further divided into three sub-groups: grammar videos, writing videos and text videos. The $R_{w/l}$ of all the three sub-groups in the two classes were compared.

Learning activities	Learning records	Task Time & Task Type			
		Pre-class	During-class	After-class	End-of-semester
Recorded instructional videos	Each student's video completion rates ($R_{w/l}$) for each sub-type of videos	●		Δ	
Quizzes	each student's completion rate and quiz scores	●		▲	
Online forum discussion	Each student's postings per sub-type discussion ($R_{p/d}$) and the received likes	○	○	○ Δ	
Classroom feedback	Not available		● ▲		
Lecture	Not available		▲		
Individual/ pair work	Each student's completion rate and exercise grades		●	▲	
Classroom group discussion	Not available		●		
Group work projects	Each student's completion rate of projects, attempts of peer assessment and grades for each project		● ▲	● ▲	
Essay writing	Each student's completion rate, revision attempts and essay grades			● ▲	
Final examination	Each student's final examination score and the scores of each sub-section				● ▲

"●" stands for a required activity in FC;
"○" stands for an optional activity in FC;
"▲" stands for a required activity in BC;
"Δ" stands for an optional activity in BC.

Figure 3. 7 Summary of the learning activities in IEC and the corresponding learning records

When measuring student engagement in **forum discussions**, the number of postings each student placed in the discussion forums and that of the likes they received from their peers were collected. The forum discussions were divided into four sub-groups: student-led discussion, teacher-led exercise discussion, teacher-led topic discussion and student exemplar work discussion. As each discussion included different topics, each student's average postings per discussion ($R_{p/d}$) was used to measure student engagement in the online forum.

Every student's completion rates of their quizzes, individual work and essays were collected to measure student engagement in these **independent exercises**. Student attempts to revise their essays were also collected to help understand their effort in essay writing. In measuring student engagement in **group work projects**, every student's completion rate of all the projects and their attempts to peer assessment were collected.

Students' grades for quizzes, individual exercises, essays and group work projects throughout the semester, and those in the final examination, were collected to measure their academic performance and examine the enhanced performance in FC. Each **quiz** had about 15 objective items; all were automatically rated and transferred to the hundred-mark system. Each **individual exercise** had about ten items made up of translation, paraphrase and text-related questions. A human teacher rated every item on an A to E scale, where A was automatically converted to the full mark, B to 80%, C to 70%, D to 60% and E to 50%. An item was rated zero if it was not answered. In addition, students wrote two **essays** during the semester via iWrite, a corpus-based automatic essay rating system. Essays were rated in terms of language

(35%), content (50%), and organization (15%). **Group work projects** were assessed by both the teachers and the students, where teacher assessment accounted for 60% of the total grade and peer assessment for 40%. However, as focus group discussions indicated, most students did not follow the assessment criteria in peer assessment; only the grades assigned by the teacher were used for data analysis.

Student scores from the **final examination** were collected as an indicator of the overall learning outcome of the semester. Instead of GPAs that include daily student performance, the final examination grades were used to minimize repeated measurement. Vocabulary and grammar items in the final examination were selected from the textbook content to measure knowledge retention. They were all objective items and were rated automatically. Translation, cloze, reading and writing items measured students' language and higher-order thinking skills. The keywords in each translation item were selected from the texts students learned in the semester. Human teachers rated translation items based on loyalty to the original language and language accuracy. Cloze and reading items were randomly selected from the question bank of College English Tests and were automatically rated. The 200-word essay was rated by human teachers with the same essay rating criteria in iWrite mentioned above. The IEC final examination tested students' reading and writing proficiencies but not listening and speaking proficiency. No additional standardized tests examined students' listening and speaking proficiency at the end of the semester when this project was carried out. Therefore, data on students' listening and speaking proficiency were not available.

3.7.2.2 Statistic Analysis

Normality tests, together with the skewness and Kurtosis and the scatterplots of the data, were conducted first to decide the appropriate statistical tests that followed. Independent sample t-tests or Mann-Whitney U tests, in case data were not normally distributed, were run to check whether there were statistically significant differences (Cohen et al., 2011) in student engagement in learning activities between FC and BC and where the differences lay. Paired sample t-tests (or the Wilcoxon tests in case of the non-normal distribution of data) were used to determine the within-class difference (Cohen et al., 2011) in student engagement when there were two sub-types of a learning activity. When there were more than three sub-types of a learning activity, Friedman tests were conducted to determine within-class differences in student engagement (Cohen et al., 2011). The Wilcoxon tests were then conducted as post hoc tests to determine where the differences existed. In such cases, Bonferroni correction was applied to adjust the significance level $p = .05$ by dividing the p-value by the number of pairwise tests performed (Armstrong, 2014).

In terms of the student grades, missing data were replaced by series means. The independent t-tests or Mann-Whitney U tests, in case data were not normally distributed, were administered to determine whether there were statistically significant differences in student grades for the learning tasks and the final examination between FC and BC. Correlation tests were run to check the association between student performance in the learning tasks and their grades and the direction and magnitude of the relationship. Multiple regression analysis was further run to check whether student engagement in the learning

tasks predicted their academic outcomes and, if so, to calculate the weighting of the independents on the dependents (Cohen et al., 2011). The quantitative data were analysed with SPSS 28.

3.8 Validity and Reliability

By employing both qualitative and quantitative methods to evaluate the effectiveness of learning activities in the learning process and outcomes, the mixed-methods design of this project optimizes the opportunities to acquire credible evidence for improvement (Collins & Onwuegbuzie, 2013). Merging data can be the most challenging part of this design because it requires the same measures or assessments on both the quantitative and qualitative datasets (Creswell, 2013), a critical validity indicator of a convergent parallel mixed-methods design (Creswell, 2014).

This consistency issue in measurement was solved by the evaluation framework and the data integration strategy for this project. Creswell (2014) suggests conceptual or theoretical frameworks are important to the design of mixed-methods research. By attending to the Kirkpatrick Model's evaluative criteria, this project's quantitative and qualitative data complemented or elaborated on each other systematically. I built parallel constructs by combining the objective quantitative data on student performance in the learning tasks with the self-reported qualitative data depicting student engagement that contributed to their performance in these tasks. By doing so, data from the two phases were investigated side by side, and findings were grounded in the data. More importantly, findings were less biased when self-reported data were examined

together with objective data. The evaluative framework and the data integration strategy added to this mixed-methods research validity or credibility, a qualitative term corresponding to the quantitative concept of *validity* (Yilmaz, 2013).

The initial equivalence of the participants in the two classes and the descriptive validity (Collins & Onwuegbuzie, 2013) helped mitigate this project's sampling weakness. Convenience sampling and a small participant size may compromise the generalizability of a project. The self-evaluative nature of this project denotes that I studied the course and the classes of students I was teaching. However, it is worth mentioning that students, though of the same major, were randomly assigned to the two classes under investigation, so selection bias was minimized. Admittedly, a sample size of no more than 28 participants in each group and less than 53 in total was not ideal for a rigorous quantitative educational study (Creswell, 2013). Nevertheless, the initial equivalence of the participants (as presented in Section 3.2) reduces the number of sampling units necessary for data saturation and statistical significance (Bazeley, 2013; Collins & Onwuegbuzie, 2013; Guest et al., 2006; Hennink & Leavy, 2014), hence reducing the consequences from the sampling weakness. In addition, I tried to maximize the descriptive validity, which refers to the factual accuracy of the account, to build up the credibility of this study (Collins & Onwuegbuzie, 2013). Towards the end of each focus group discussion, I played back to the group the important issues that had been covered and allowed students time to comment, revise or expand on their contribution to guarantee descriptive precision when reporting the findings. In

addition, interview transcripts were shared with participants for member-checking. I kept field notes of my observations and reflections to accumulate evidence and establish the credibility of conclusions. I tried to maintain transparency by giving thick descriptions of the context, participants, actions and events and attending to divergent results and negative cases so that naturalistic generalizations could be made (Collins & Onwuegbuzie, 2013). Though the generalizability of the findings may be limited to practitioners in EFL flipping pedagogy and course evaluation, the global ensemble of teaching English as a foreign language, flipping pedagogy and quality enhancement may well expand the chance of case-to-case transfer.

My stance as an insider researcher also adds to the validity of the findings. According to Trowler (2016), insiders are culturally literate, which leads to good judgement, an essential element for successful case study research. With a good knowledge of the context, insiders are more likely to capture complexity in a university's unique multiple cultural configurations. The most direct benefit is that insider researchers have better access to naturalistic data and respondents. As the lecturer of IEC, I worked closely with my students and stayed responsive to their needs in learning. As a result, trust and empathy were established between us; all the students in the two classes were ready to participate in the project. Student voice managed to be maximized. I informed my participants of the evaluation process and tried to raise their awareness of the utility of both the evaluation process and the outcomes. Student ownership was, therefore, increased. The maximized participant voice and their increased

ownership, in turn, enhanced the integrity of the data and validity of the research findings (Scriven, 2003; Yilmaz, 2013).

That said, I understand the negative sides of being an insider. It may become difficult for an insider to notice some dimensions of the lived practice that are apparent to an outsider because these dimensions are normalized (Bell & Waters, 2014; Trowler, 2016). Additionally, power relations may be involved in researching students one is teaching. To avoid these potential negative consequences, I elected to collect data via less obstructive methods.

3.9 Ethical Considerations

Ethical approval of this project was granted by the Department of Educational Research at Lancaster University. This project involved collecting data from student participants via focus group discussions; power differentials between the teacher and the student participants and confidentiality between group members were the two major ethical issues that needed to be addressed.

The potential power relations arising from a teacher researcher researching students he/she is teaching was the biggest threat to ethics and data integrity in this project. The following measures were taken to guard against the potential risk: when introducing this project to students, I made it clear and emphasized to all the students in the two classes that participation was voluntary and not associated with any form of assessment for the course and that non-participation or withdrawal would not lead to any adverse consequences. In data collection, focus groups were administered instead of the one-on-one interview to avoid potential interview bias arising from the teacher-student

power differentials (Trowler, 2016). Considering that the participants in this project were involved in multiple data collection sessions, I added a sign-up procedure before each focus group discussion so that participants could choose how they were engaged in data collection. It turned out to be a wise decision, as some students did not sign up for focus groups though they consented to participate in the project. In the pandemic emergency, I was aware that any potential health risk should be guarded against when collecting data, so I stopped recruiting participants or rescheduling discussions when I realized prolonged data collection would increase the risk of exposing the student participants to the threat of COVID.

As focus groups involve more participants than interviews in one sitting, confidentiality is more at risk. At the beginning of each discussion, I reminded the participants not to reveal information about the discussion without the permission of the other group members.

3.10 Summary

This chapter commences with my rationale for formative course evaluation and the mixed-methods design, followed by a brief introduction of the sampling strategy and my epistemological stance. Then the Kirkpatrick Model, the evaluation framework adopted in this project, is presented, and the research context briefly described. The context section is positioned here because it incorporates sub-sections on the LMS and the course design of the classes under evaluation, which are closely related to the research design of this project. An overview of the convergent parallel mixed-methods design for this

project is given afterwards. Then data collection and analysis methods are described in detail. The chapter concludes with a discussion of the reliability and validity of the study and the ethical considerations I made when doing this evaluative project.

Chapter 4: The Qualitative Findings

4.1 Introduction

The qualitative phase explored students' affective reactions, perceived cognitive gains and behavioural engagement in the learning activities through the participants' voices. Kirkpatrick emphasizes that the four levels in the Model should be examined in sequence so that the evaluation results are understood with precision (Kirkpatrick, 1996; Kirkpatrick & Kirkpatrick, 2006). As such, the qualitative data were explored before the quantitative results to answer RQ 1 and its sub-QRs:

RQ1. To what extent are students in FC engaged in the learning activities?

RQ 1.1 To what extent do students in FC react affectively to the learning activities?

RQ 1.2 To what extent do students in FC perceive they learned from the learning activities?

RQ 1.3 To what extent are students in FC engaged in the learning activities behaviourally?

Data for this phase were mainly collected from focus group discussions. Field notes of participant observation were used to assist in explaining or interpreting the focus group discussions. Themes were derived from and centred on student engagement in the learning activities and are presented in order of the learning activities in FC. Each theme centred on one specific learning activity and

answered the sub-RQs in sequence. Data from BC is presented alongside FC for a better understanding of student engagement typical of FC. The italicised texts in this chapter are direct quotes translated from the participants' responses given in Chinese in the focus group discussions.

4.2 Student Engagement in Recorded Instructional Videos (theme 1)

As in a typical flipped classroom, FC began with students watching videos before class, which accounted for 20% of the points for students' daily performance in FC. There were usually one or two videos for each period. After adjustments were made based on student suggestions in Pre-FG, each video lasted no more than 15 minutes. The videos fell into three sub-types to impart text-related language knowledge to students and prepare them for hands-on classroom practices. During the semester of this project, the grammar videos for IEC covered grammatical knowledge on verbal tenses, articles, determiners, nouns, conjunctions, imperative sentences, inversion and relative clauses. The writing videos introduced to the students the genres of narration, exposition and argumentation, the techniques of compare and contrast, the style issues of sentence fragments and comma splices, and the rhetorical device of parallelism. The text videos presented the students briefly with the text to be covered in the coming period, mainly to address the language points used in the text. Videos were not a requirement of BC but were available to its students as an online learning resource. Five per cent of the points for daily performance were allocated to video watching in BC to encourage students to use the video resources to assist in learning.

4.2.1 **Student Affective Reactions to Recorded Instructional Videos**

Overall, students in FC were satisfied with the video resources, grammar videos, in particular, describing them as *helpful*, *useful* and *covering essential knowledge content of the text*. They took watching videos as the primary source of knowledge. Some participants held that this pre-class learning task equipped them with *sufficient knowledge to answer quiz questions*, *prepared them for individual exercises* in class, and improved *classroom efficiency*. Some mentioned its unique strengths of repeatability and controllability that were impossible with a classroom lecture delivered by a human teacher. They reported they learned better from videos because they could *control the pace of learning* and *time spent on it*. Towards the end of the semester, some participants concluded that video-watching was one of the most *practical* learning activities of the flipped class.

However, participants in FC differed in their perception of the difficulty level of the videos. While some claimed *video content was no more difficult than high school level*, some felt *videos spoke too fast for them to catch up*, some confessed *they could not learn as much by watching videos as by reading printed words,* and there were also students who expressed *a need for mother tongue support*. As students in FC attached importance to videos, they experienced anxiety when they failed to understand the content and had to rewind the videos repeatedly. Some went *restless* when unable to stay organized and finish watching videos before class.

Due to the difference in course design, participants in the two classes differed in their perceptions of the roles videos played. Students in BC took videos as supplements to lectures. They thought videos were *handy* when they failed to follow the lecture and missed important class content or doing individual exercises. They added that l*earning grammar via videos was more effective than lectures* because *videos are condensed and short, so* they were more likely to notice the key points. However, they still preferred interactive lectures. To them, videos were *less interactive than classroom lectures*. Consequently, they kept *drifting away from the one-way videos* despite their compactness.

> *Grace: Videos cover the same subject content as lectures but are condensed and much shorter. But they appear dull and rigid and, therefore, less impressive. I forget video content within days.*
>
> *BC-Pre-FG-2*

It is worth noting that some students in FC reported the same problems with the one-way transmitted video. In addition to the challenge of staying focused, some felt that video content appeared *less impressive* to them, resulting in *short memory* of what was learned.

4.2.2 **Student Perceived Gains from Recorded Instructional Videos**

Participants in FC reported they *learned language knowledge of vocabulary, grammatical rules and text-related content from videos*. However, learning effects varied among participants, especially in the first half of the semester.

While some claimed *they could complete pre-class quizzes and in-class exercises and participate in discussions with what they had learned from videos*, some reported they *had difficulty understanding the video content even after they repeated videos for times*. As a result, they *made many mistakes in quizzes and exercises, especially about vocabulary and grammar*.

By the end of the semester, participants in FC reported improved language skills. They could better understand videos delivered in English and paraphrase without relying too much on the teacher's explanations. They *no longer blindly took everything down* from video recordings. Most students learned to choose what to write down as they were more capable of *identifying what was critical* for themselves. These changes in note-taking from trying to copy the content to a meaningful understanding indicated a transformation from surface learning to deep learning on the learners' part (Ellis et al., 2017). Moreover, they developed better learning skills. They were *more efficient* as they *could make better use of LMS* and became *technology-savvy*. More students started to *take electronic notes* by using *Notability* or *Microsoft OneNote*.*Their learning autonomy developed* accordingly.

As a result of the different roles videos played, students in BC differed tremendously in their perceived gains from videos. Students in BC rendered videos *relatively limited in use* and barely mentioned any gains from videos in their discussions.

4.2.3 **Student Behavioural Engagement in Recorded Instructional Videos**

Students in FC demonstrated great agency and developed techniques for handling videos. Almost all the participants reported that they *took control* when watching videos.

> *Sissi: I control the video speed. I usually speed up the video by 1.5 times but will pause to read the textbook and look up the dictionary.*
>
> *Irene: I am a little inefficient. I need to repeat it at least twice before I understand the content. And I'll pause a long sentence many times to understand it; otherwise, I forget the previous part of the sentence when moving on.*
>
> *Allen: The same as they do. I need to repeat the videos. I will rewind the videos to take notes.*
>
> *Alisa: I usually watch a video twice. I go through the video without a stop for the first time to have a general understanding of it and take notes during the second time, pausing and rewinding it where necessary.*
>
> *FC- Pre-FG-1*

The time and effort students spent varied by the length or content of the videos. Many students claimed that they would *spend much more time on grammar videos.* Nonetheless, they were *less likely to repeat a video if it was long*.

As time went on, students could regulate their learning better. At the beginning of the semester, some would *forget to watch videos*; by the time of the Pre-Focus, students had become more organized. Almost all students could *finish watching videos before class*. Nonetheless, some students lacked self-control, were vulnerable to distractions when watching videos and failed to complete the task repeatedly. To enhance efficiency and effectiveness, students would *familiarize themselves with vocabulary and text content* before watching videos. They would choose the time and places to watch videos to avoid interruption. Some collaborated with roommates for that purpose.

> *Renata: If possible, our dorm will find a time and watch the video together.*
>
> *Zoe: Yes, so we don't interrupt each other. When someone needs to pause to take notes, we would stop and wait or do something else related.*
>
> *FC- Pre-FG- 2*

Towards the end of the semester, students managed their learning more efficiently. They established the habit of *using Internet resources or reference books* to assist their understanding of the video content. They would *review videos* when encountering problems in quizzes or assignments. Some would *review videos regularly to combat short memory*. However, students still varied in their technological skills. While some learned how to control videos and use LMS within the first few weeks, some students never managed to watch videos

on their computers simply because they did *not know how to run Flash* to allow videos to be played on their computers.

As students in FC and BC used videos for different purposes, they differed tremendously in how they dealt with them. Though videos were not a requirement of BC, most students still watched videos, but only as supplementary resources after lectures. Most of them made use of videos to *assist themselves in completing quizzes and assignments*; some would turn to videos to *make up notes that they had missed out in lectures*. Some students reported they would rewind the videos and take notes to understand them well. However, more admitted they just ran over videos but *kept drifting away*.

4.3 Student Engagement in Quizzes (theme 2)

Quizzes were designed to test students' knowledge retention and understanding of the video content. Students in FC were to do pre-class quizzes after watching the video(s). In BC, students took quizzes after the lecture. Quizzes in BC were, at first, taken in class but moved out of class after the Pre-FG as students complained that *limited class time affected their performance in quizzes.*

4.3.1 Student Affective Reactions to Quizzes

Students in both classes tended to use quiz outcomes as indicators of learning effects and were, therefore, emotionally affected by quiz scores. They felt *down and anxious about their poor performance and low grades* and *were motivated* when they *made progress or achieved higher scores*.

Lynn: I did pretty poor in quizzes previously. I am exalted now that I can have progressively higher scores.

FC- Pre-FG- 2

Mari: I am dissatisfied with everything about myself. I am doing so poorly and still making many mistakes despite my effort.

FC- Pre-FG- 4

Sue: It is not about the form of the task but the score. My average points go down because of the low quiz score.

Tim: Yes, that is what makes me anxious.

Sophie: But it brings me a sense of achievement when my scores progressively increase.

Tim: Yes, this is when I am motivated most.

BC- Pre-FG- 1

Except for their similar response to quiz scores, students in FC and BC reacted to quizzes per se quite differently. Though anxious about their performance, students in FC were uniformly positive towards quizzes. Students from FC shared that quizzes were *effective for learning* because they were *closely related to the video content* and helped them *check the extent they learned*. They held that quizzes *motivated them to watch videos in time*, helped them *realize what they failed to learn*, *promoted them to review or make further*

inquiries and *expanded their exchange with peers*. In the post-FG, some students concluded *quizzes were the most conducive to learning*; some went so far as to take the pre-class *video + quiz as the principal part of IEC* instead of the face-to-face classroom learning.

In contrast, students in BC expressed more negative feelings related to quizzes. Because of the time limit, they felt *nervous when doing quizzes in class.* Some felt resistant to in-class quizzes because they were reminded of *the dreadful experience of sitting exams*. Some students attributed their poor performance to the LMS interface, complaining that it was *inconvenient to use* and *uncomfortable to read on* and claiming that they could have *done quizzes better on paper*. However, students' discomfort with the LMS relented when they were habituated to doing tasks online and performed better in quizzes, especially when they had enough time to finish the task after quizzes were moved out of class. Some students started to embrace doing tasks online, seeing it as *a trend for education in the future*.

4.3.2 Student Perceived Gains from Quizzes

Quizzes enhance learning because they not only feed back learning outcomes but also indicate the effectiveness of videos. Students in FC confirmed that quizzes enhanced their language knowledge and developed their learning skills. They were more familiar with the LMS and made better use of tools for learning. In contrast, students in BC rarely mentioned any gains from doing quizzes, which might very well result from their negative feelings about quizzes that affected their performance.

4.3.3 Student Behavioural Engagement in Quizzes

Students in both FC and BC demonstrated significant autonomy in doing quizzes. To get higher scores, most students would *review what they had learned before taking quizzes or take quizzes as soon as they learned the text* when *their memories of the content were still fresh*. Most of them could *finish quizzes in time*. Some would go further to *review or make further inquiries to solve their problems encountered in quizzes*. Besides, students in both classes *paid close attention to quiz scores*. Students' exceptional agency in quizzes can be partly explained by their perception of quizzes as a gauge of learning effects. However, it should be noted that in both classes were students who lacked autonomy and *failed to finish quizzes in time repeatedly*.

When doing quizzes before class, students in FC made better use of their devices to contain the inconvenience of the LMS interface.

> *Evan: You don't have to scroll the screen up and down.*
>
> *Gordon: Yes, I use two devices, one to read the question items and the other to answer the questions. So, it does not trouble me much if questions and choices are not displayed on the same page.*
>
> *FC-Pre-FG-4*

In contrast, more students in BC were struggling with the LMS. Quite a few students in BC were unfamiliar with the LMS, especially at the beginning of the semester. They *could not stay organized when doing quizzes via LMS*,

especially when they were doing quizzes in class. They *always handled the interface by mistake*, which *distracted their attention from quiz questions*. Some could never remember to save their answers and kept *starting over again when their answers were lost*. In addition to the technical issues, some students in BC reported they *forgot to take quizzes* after they were moved out of class.

Besides, students in FC interacted more with classmates about quizzes. When they had trouble answering quiz questions, they would like to *discuss their problems with peers* in addition to referring to other resources. However, some students in FC made every attempt to *find answers to quizzes*. Some *checked quiz questions before watching videos* so that they could *have a clearer focus on video content*; some *took quizzes over videos* so that it was much easier to find answers.

4.4 Student Engagement in Online Discussion (theme 3)

Forum discussions fell into four sub-types. An online discussion forum was available for students to discuss problems they encountered in learning a specific section and share ideas on given topics. These problem discussions were further divided into two sub-types: one was led by students themselves, where they posted their problems and discussed them with peers. The other was teacher-led, where the lecturer posted common errors from student work for students to discuss. The lecturer also led topic discussions in the forum, where topics related to the text were posted for students to share and exchange ideas. Lastly, exemplary student group work for each unit was posted in the forum to improve student performance, with teacher comments alongside.

Students were encouraged to write their comments on the exemplary work and reflections on their performance. All four sub-types of forum discussions were optional to serve different student learning needs. This design of optional participation helped capture, more truthfully than required activities, how the students in the two classrooms involved themselves in the learning activities for different purposes. Unlike video-watching and quizzes, where student engagement was much influenced by class design, there appeared to be fewer overall between-class differences in students' engagement in forum discussions.

4.4.1 Student Affective Reactions to Online Forum Discussions

Participants were generally positive about the forum discussions. They shared that *forum discussions solved most of their puzzles* and *enhanced learning*. A large portion of the participants from FC deemed forum discussion *the most useful and efficient learning activity in IEC*.

> *Rachel: The investigation is a process of verifying my understanding, and explaining my understanding to others enhances memory.*
>
> *FC- Pre-FG- 2*

> *Niamh: I find it very meaningful. To avoid misleading my classmates, I will try to refer to authoritative resources. I am more meticulous in learning than before.*
>
> *FC- Pre-FG- 3*

Sophie: Sometimes, I am not quite sure about the answer to the question either. That prompts me to make further references. The point is that when I answer that question with what I have found out, both of us learn.

Tim: Yes.

Sue: And it is more impressive when we get the answer through discussion and investigation.

Ella: Exactly.

Sue: The more complicated the question, the more helpful discussion and exploration are to enhance understanding.

BC- Pre-FG- 1

Forum discussions not only enhanced student memory and understanding of what they learned but produced a sense of belonging, which is a critical contributor to student engagement (Kahu & Nelson, 2018). Most students admitted they *benefited most from reading peers' postings in the forum* even *though they didn't ask any questions.* As for those who sought help in the forum, they confirmed that *peers were so responsive that they felt supported.*

Positive as most students were, there were still different voices. Helpful as students thought of forum discussions, some held it was secondary to the teacher's explanations. Many students, especially those in BC, indicated a distinct preference for teacher-led discussions, as they included *problems*

common to them. Students' awareness of their shared problems gave rise to their *doubt about the accountability of their peers' answers*, which may help explain their preference for teacher feedback.

> *Nora: After all, we are students. I would like to use the forum as a channel more to exchange ideas with my classmates [than solve problems].*
>
> *Bella: The answers produced by us may not be as reliable as those by teachers.*
>
> *Ivan: No, they are not. I will not rely on my classmates' explanations too much.*
>
> *Nora: It appears they are not quite sure about their own understanding.*
>
> *Ivan: Those long explanations are just downloaded from the internet and may not be reliable.*
>
> *BC- Pre-FG- 4*

It should be noted that students, in general, preferred to have answers that were *simple and to the point* rather than *long* and *complicated*. Some were *impatient to read long answers*; some felt it *difficult to make sense of the long and inclusive explanations*. There were still students who considered forum discussion of *limited use*, as *answers in it were similar,* or *the questions and answers were difficult to understand and follow*. It is also worth mentioning that

students in both classes expressed their *point-earning mentality* in joining in forum discussions, though some were quite *annoyed by their peers' point-earning attempts*.

4.4.2 Student Perceived Gains from Online Forum discussions

Students in FC and BC consistently testified that forum discussions enhanced their language knowledge. They *had their questions straightened out*, *made up what they had failed to learn* and *enhanced their memory and understanding* of what they had learned. Moreover, they were motivated to *explore reliable answers*, where their learning skills were improved. Most importantly, they learned to *compare and judge peers' postings*, indicating higher-order thinking skills (Bergmann & Sams, 2014).

4.4.3 Student Behavioural Engagement in Online Forum Discussions

While there appeared little distinct between-class difference in how students participated in forum discussions, there were apparent cross-case differences within the class. On the whole, students in both classes demonstrated great initiative and collaboration in student-led forum discussions. Most of them joined discussions for two primary purposes: to solve the problems they encountered and *earn points by answering peers' questions*. That said, many of them admitted they joined forum discussions *more to answer than to ask questions*. There were signs of a point-earning mentality among students. Most students regulated their participation in discussion by *their points in LMS* and measured their performance by *comparing points with peers*. Though driven by points, they were *serious with what they posted*. They would *try to produce*

reliable answers by *making further references and inquiries* before posting their answers in the forum. Some students *paid close attention to forum discussions* and would *log in to LMS regularly to follow the postings* in which they were interested.

Students varied in their way when reading the postings. Some would *go through every posting and take down what was helpful to them*; some just *read the latest postings* as they found out peers would *post similar answers to the same questions*. Some students were considerate and self-restrained. They would rather not earn points than repeat the questions or answers. They *read the discussions*, *gave likes to the postings they approved* and *left without posting anything*. As a result, students would *pay particular attention to the postings with more likes*. Many students insisted on *finishing quizzes before visiting the forum* even though they knew that they could find answers to quizzes in the forum and that some of their peers did that for higher quiz grades.

Some students had clear orientations when visiting the forum. Some *visited the forum only when they could not solve their problems*. Some *paid particular attention to teacher-led discussions* or *joined in teacher-led discussions only*. That said, students were less interactive and engaged in teacher-led discussions. Most of them just came to *post their answers or ideas and left, not reading peers' postings or caring whether their postings were repeated or correct*. As a result, they did not learn much from the discussion. There were usually several questions in a teacher-led discussion. *Some students would post their answers in separate threads to earn more points*. There were

students, though very few, who *hardly joined the online forum.* The reasons for non-participation were many and various. Limited perceived benefit from forum discussions, lack of interactive skills and inclinations, and peer influence were the primary reported causes.

4.5 Student Engagement in Classroom Teacher Feedback (theme 4)

Classroom activities in both classes started with teacher feedback on outstanding problems from the other learning activities. To students in FC, teacher feedback was the only learning activity the teacher took charge of in class.

4.5.1 Student Affective Reactions to Classroom Teacher Feedback

Students in FC held that *teacher feedback was indispensable to enhance learning.* Most of them held that teacher feedback extended *what they had learned from videos and forum discussions*, *solved the critical problems beyond them, and guided them to approach their problems properly*. By helping them understand why they made mistakes, students were *motivated to reflect*, which was essential to autonomous and deep learning (Barton & Ryan, 2013). Some students added they were *in the habit of relying on teacher guidance,* which, they believed, was *authoritative* and, therefore, *more reliable than inquiry or collaborative learning.* That belief resulted in their preference for teacher guidance even though they could find authoritative resources and comprehensive answers to the questions by themselves.

Evan: We are in the habit of following and relying on teachers. After all, this is what we have been doing during all these years of schooling.

Carl: Honestly, most of the time, I am waiting for teacher feedback.

Evan: Yes, because my problems are usually the same as my classmates.

Micky: I am not assured of peers' explanations. I don't think they know better than I do.

Evan & Carl: Yes, we prefer teachers' explanations.

Gordon: Yes, I understand better and learn more from teacher feedback.

Evan: Very often, we cannot make out the subtle differences by looking up dictionaries or discussing with peers.

FC-Pre-FG-4

Due to the differences in class design, students in BC perceived the role of teacher feedback quite differently. Most students in BC held teacher feedback *helpful* in that it *refreshed their memory of the knowledge content learned from lectures*. Many Students in BC *preferred to solve their problems from face-to-face teacher feedback in class rather than through peer collaboration in the online forum*.

4.5.2 Student Perceived Gains from Classroom Teacher Feedback

Classroom teacher feedback enhances students' language knowledge and higher-order thinking skills. Students in both classes claimed *their memory and understanding of language knowledge were enhanced* through teacher feedback, which was *conducive to producing correct language forms*. In addition, they learned to *compare language forms*, *conclude language rules* and *reflect on their mistake*, which are indicators of higher-order thinking skills.

However, it is worth mentioning that some students claimed they *gained nothing from in-class teacher feedback*. These mainly included students who *attached little importance to language knowledge* and those who *did not participate in the online forum discussions*, as classroom feedback involved solving problems left unsettled in the forum. Besides, teacher feedback solved the problems common to most students; therefore, it was impossible to cater for individual student needs. Some students in FC reported that *they failed to understand teacher feedback* as *what the teacher explained was unrelated to their problems*.

4.5.3 Student Behavioural Engagement in Classroom Teacher Feedback

Students demonstrated limited autonomy in teacher feedback. Some students claimed that to understand the feedback better, they would *make preparations by reviewing quizzes and assignments or postings in the forum before class*. Some would *take notes of what was being explained in class*. Nevertheless, classroom observation indicated that most students just *sat there and listened*.

When asked if they had any further questions to ask, there was hardly any response from the students.

4.6 Student Engagement in Individual Exercises (theme 5)

Individual exercises were assigned for each period of IEC. They were to develop students' language skills in reading comprehension, translation and paraphrasing by putting into practice what students had learned from the text. In FC, a period of 20 minutes of individual exercises followed the classroom teacher feedback and was followed by a group discussion to address the problems students encountered in the exercises. In BC, the same individual exercises were to be completed after class without a time limit for the task.

4.6.1 Student Affective Reactions to Individual Exercises

There were many common affective reactions to individual exercises among students in FC and BC. On the whole, students in both classes were positive about individual exercises. Both commented that the exercises were beneficial because they *motivated them to learn the text carefully*. Both were comfortable with individual exercises because they were accustomed to working independently, and *the workload was moderate*. Of the three exercise types, students in both classes *favoured translation exercises* but considered *paraphrasing the most challenging*. Both *disliked the unfriendly LMS interface*, describing it as *inconvenient to read on* and *harmful to the eyes*. However, they admitted, at the same time, that their *dreadful feeling toward the interface would relent if they could perform the task better and obtain higher points*. That is to

say, their feelings and attitude toward LMS were much affected by their performance and the points they could obtain from the task.

Despite the many commonalities in the two classes, there were still distinct between-class differences due to the different course designs. Students in FC perceived the in-class individual exercises as *the most efficient and effective learning activities* in IEC. The efficiency can primarily be attributed to their long-established *habit of working individually*, which enabled them to *adjust to the task quickly*. The 20-minute exercises *prepared students for effective engagement in group discussions*. They deemed *doing exercises a more effective way to learn than listening to lectures*.

> *Gordon: Lecturing is not as effective. To finish the exercises within the time limit, we must read the text very carefully and be familiar with the words and expressions.*
>
> *Micky: Lecturing is passive, and students can slack off.*
>
> *FC-Post-FG-3*

However, students in BC had more negative feelings towards the individual exercises. As the task was to be completed after class and timely feedback and support were not always possible, they encountered more difficulties when doing the exercises. Although there was usually two to three days' duration for the task, some students were still *anxious to catch the deadline*. Some students *did not attach much value to the exercises*; some *disapproved of digital assignments*.

4.6.2 Student Perceived Gains from Individual Exercises

There are both between-class and within-class differences in students' perceived gains from individual exercises. Most students in FC had a good command of the knowledge presented in the videos, so they could finish the 10-item exercises within 20 minutes. While some students in FC reported their enhanced impression of the text content and improved language knowledge and skills through exercises, some reported difficulty doing paraphrases on their own or correcting errors in their work without support. Student divergences in their perceptions of improved language knowledge and skills were more substantial in BC. Distinct differences were found in vocabulary, paraphrasing and reading, primarily because students *couldn't find reliable resources to solve their problems*, they found it *impossible to identify errors in their work by themselves,* or they were *unsure of their answers* or *unable to express themselves clearly*. In brief, students in FC performed better in text reading comprehension. However, students in both classes had shared problems in correcting their errors and doing paraphrases.

4.6.3 Student Behavioural Engagement in Individual Exercises

Students in both classes demonstrated significant autonomy in individual exercises; however, their approach to learning was somewhat performance-oriented. Students appeared to attach more importance to doing the exercises than learning from them and tried their best to *get higher scores*. Some students *would leave* their *exercises altogether behind after submission*. They *paid no attention to teacher feedback* on their performance, let alone *corrected*

the mistakes or errors in their work, even if these were marked out for them. Some *had no idea of teacher feedback* on individual exercises at all.

Students in FC would *refresh their memory of the text content* by *reviewing the text before class* so that they could complete the exercises within the time limit in class. All of them *were highly concentrated when working on the task*. Most *finished the task within the time limit*, which could hardly be done without full preparation and familiarity with the subject content. However, several students could never complete the work in time. Students in FC would *take the initiative to improve their work* by *exploiting what they learned from the following group discussion*. They were motivated to do so partly because improved work could *bring them higher grades*.

To improve performance, students in BC would also prepare before the individual exercises. They would *review the text*; some would *go through the video and take notes*. However, their performance in the task was not as consistent as that in FC. Some of the problems in BC were technical issues. While many students *became familiar with LMS* and learned to *use different devices to work efficiently*, some students *loathed digital assignments* and could *never handle LMS properly*. Nevertheless, more problems were with learning autonomy. While many students *stayed organized* and *completed the exercises on time, some* would always *miss the deadline*. Many students could *complete the exercises independently within 30 minutes based on what they learned from the lectures or the videos*; some would *spend hours* not knowing how to answer the questions and where to find the answers. Technically, towards the end of the semesters, some students appeared to be *in the habit of*

doing tasks via LMS, so much so that they *never remembered their offline tasks*.

4.7 Student Engagement in Group Discussion/Interactive Lecture (theme 6)

In FC, individual exercises were followed by 30 minutes of face-to-face classroom discussion for students to discuss their finished exercises and posted opinions. The students either volunteered or were invited to clarify, complement and justify their ideas, and the instructor attended to the shared problems and extended discussions. In BC, 60 minutes of classroom time was assigned to lecture in a dialogic manner. The lecture covered the content in videos and classroom discussions in FC.

4.7.1 Student Affective Reactions to Group Discussion/ Interactive Lecture

Students in FC diverged in their attitude to classroom discussion. Many were positive, describing it as *effective*, *helpful* and *necessary*. They regarded classroom discussion as *an extension of the previous learning activities, which made learning go deeper and broader*. Learning became *effective when they discussed with peers and followed the teacher to compare and integrate different ideas and correct or improve their understanding*. On the other hand, some held classroom discussion was *less effective*. They felt it *difficult to follow it* because they were *too busy taking notes* or *could not understand what their classmates were discussing*. Some were *too shy to bring forth their questions* in

front of others, so their problems were not always adequately addressed; some *did not see the value of interacting with peers*.

Tracy: I had no experience discussing tasks with peers in high school. I preferred to work alone, for my ideas can differ from others.

Lora: Then that is good for brainstorming.

Tracy: But I feel uncomfortable thinking differently from others, and it is not easy to adapt to. And it can be embarrassing to ask questions others may have already understood.

FC-Pre-FG-3

Micky: We used to sit exams in every class in high school; I've got no experience of discussions with my classmates in class and have no idea how to do it effectively. I need time, maybe quite long, to adapt to it, but our class time is limited. It can be wasteful to spend time doing something ineffective in class. So for efficiency, the teacher needs to hold the stage in class.

Carl: And when discussions are just among students without teacher guidance, we are just checking answers without exchanging ideas

FC-Pre-FG-4

Despite the differences mentioned above, the participants consistently emphasized the teacher's role in managing discussions to solve shared student problems and guarantee effective learning. Most valued teacher guidance much more than peer interaction or independent inquiry.

> *Niamh: I would like the teacher to select students. I can only stay attentive and engaged in a discussion where the teacher selects students to share their ideas.*
>
> *Tracy: Me too, because I am not the kind of student who asks questions unless required.*
>
> *Niamh: When the teacher leads the discussion, it is more likely to discover our shared problems that can otherwise be overlooked. It is always hard for me to identify the errors in my work, and I feel like students tend to make similar mistakes, so without a teacher there, we probably cannot realize our problems through discussion with peers.*
>
> *FC-Pre-FG-3*

In contrast to the students in FC with diverse attitudes toward classroom discussion, students in BC were consistently satisfied with lectures throughout the semester. Most were positive about how the lecture was delivered, commenting that it was *substantive*, *systematic*, *experiential* and eas*y to adapt*. Many believed they *learned better in lectures than via videos*.

Sue: I feel like the lecture offers more than the videos do. Though both cover the same subject content, lectures extend beyond the text.

Ella: Yes, these extension is critical to learning.

BC-Pre-FG-1

Landy: I am more attracted by lectures and learn more from them. I kept drifting away when watching videos and had to repeat them many times.

BC-Pre-FG-2

Some reported that even though *lecturing was of moderate effect* and *lacked novelty to them*, and *they were less engaged in interaction*, they were *comfortable with lectures* because *they were familiar with lectures*.

L: I am given to this way of having classes. It is what we have been doing since the first day of school. It works, even if not very effective.

Michelle: Yes, this is what I am accustomed to. I mean, I listen to the teacher, take notes and try to understand the content. It makes me feel I am having a real class.

Regina: I adapted to this class the first day. Familiarity means I don't need much time to adapt to it. There is no big difference

from how we used to have our classes, but there is not much novelty either.

BC-Pre-FG-3

Only when students were answering questions, either voluntarily or selected by the teacher, did they experience some negative feelings in the lecture. Most of them said they felt *anxious* and *nervous*. Some students were not confident with their classroom performance because it was difficult to follow the lecture delivered in English, making it challenging *to stay focused in class*. Lack of necessary language skills to engage in the lectures was consistently reported in both the Pre- and the Post-FG. Some students admitted they *had a limited understanding of the lecture*. As a result, they had trouble finishing quizzes and individual exercises after class. Some expressed their *need for the mother tongue* to assist their understanding of the lecture; some students had difficulty taking notes as they *couldn't hear the teacher when they were writing*. Lack of learning skills was also part of the causes of students' note-taking problems. Some students reported that they *had no idea what to write down*.

4.7.2 Student Perceived Gains from Group Discussion/ Interactive Lectures

Participants in FC reported increased language knowledge, such as *expanded vocabulary size* and *knowledge of language rules, improved understanding of text content,* and *enhanced language skills in translation* and interactive skills. More importantly, some reported the development of the higher-order thinking capacity of *summarizing*, *analysing*, *comparing* and *integrating* ideas.

Most students in BC expressed that they *learned well by receiving knowledge* transmitted to them and *taking notes*. Many claimed they could understand *the lecture delivered in English within the first two weeks*. There were signs of improved learning skills. Some students learned to *make pre-class preparations for lectures* and *make use of other learning resources or APPs to facilitate learning*.

4.7.3 Student Behavioural Engagement in Group Discussion/ Interactive Lecture

Most students in FC demonstrated a fair amount of autonomy during discussions. They would *write down what they thought was important*, try to *improve their answers by taking the perspectives of their peers or the teacher* and *compare and integrate ideas*. Conversely, some students had difficulty judging the quality of their work. They would follow peers' ideas blindly and give up their stances all too quickly; some could not identify the errors in their work unless there was an apparent contradiction between their answers and those of their peers, indicating deficiencies in language proficiency. Towards the second half of the semester, students' skills in managing the discussion developed. To ensure effective participation, some would *make preparation by reviewing their notes and exercise items beforehand*. Students varied in their interaction with peers in group discussions. Some students had limited involvement in discussions throughout the semester because they were *busy catching up with their exercises or taking notes*. While some students were *willing to share their understandings*, some never asked questions and *preferred to handle their*

problems independently, no matter whether they could do it or not. Many just *checked answers when the teacher did not lead a discussion*.

In BC, student autonomy in engaging themselves in the lecture was mainly demonstrated by their effort to prepare before class and make up what they missed in the lecture after class. Students would *familiarize themselves with the vocabulary in the text* and *the text content before class* and *make use of the recorded videos* to *make up for what was missed in the lecture*. Though most students claimed they were attentive during lectures and tried to *take notes*, some admitted they *strayed away,* especially when they *thought they had already known the subject content* being lectured. Some *dozed off* from time to time during lectures, which never happened in FC.

4.8 Student Engagement in Group Work (theme 7)

Group work was the most complex of all the learning activities in IEC. For each group work, students were to complete a project in collaboration with their group members and do peer assessment, where they assessed their group members' contribution to their project and the completed work by the other groups. During the semester of this evaluative project, students undertook five projects in groups: a role-play of leaving home for university, a PPT representation on social manners, a 500-word proposal for AIDs patients, a survey on university students' study habits, and a debate on the national birth control policy. By the time of the Pre-FG, both classes had finished the first two group work projects and were on their way in the third one. In FC, there were 30 minutes in each period for students to work on their group work project face-

to-face in class. In BC, students had 15 minutes in class to check progress and coordinate their work in the group. Of all the learning activities discussed in the focus groups, group work attracted participants' attention most, regardless of the differences in class design. However, unlike the learning activities discussed beforehand, where apparent differences were found between FC and BC, between-class divergences in group work were minor compared with those between groups or individual students within a class.

4.8.1 Student Affective Reactions to Group Work

There are apparent similarities in student attitudes towards, perceptions of or feelings about group work in FC and BC. In both classes, students' affective reactions toward group work projects were divergent within the class. In each class, some students were serious with their work. Some tended to make a minimum effort; they did not care much about their performance and just completed their work for the assessment purpose. Many students shared the perception that *group work improved their interactive skills* but had limited effects on *subject knowledge growth or academic achievements*, so it *didn't pay off, given the time and effort spent on it*.

> *Ivan: We collect materials for group work projects from the Internet and translate them into English. But I actually learn nothing from doing these. Like the proposal for AIDS patients, I searched technical terms on the internet, and there were quite many I needed for the project, but I cannot remember any of*

them. I will gain very little unless I try to memorize those words on purpose afterwards.

Bella: Yes, we collect resources in Chinese, and many of us use automatic online translation, so there is not so much help to our English proficiency, but still, we need to spend time searching for resources and organizing what we have found.

Nora: Group work is less effective for my English proficiency than individual exercises.

Bella: less than lectures are.

Ivan: …Unless I learn every new word in the work. But that means I need to spend extra time.

Nora: On the whole, it is not cost-effective given the time and effort spent on it.

BC-Pre-FG-4

Irene: It is really time-consuming. But I don't know what I have learned from doing it. It may help develop the skills of negotiation, an important skill, but it does not help much in academic development.

Sissi: I learned more from reporting on manners but very little from role-play. Most of us didn't have new ideas for the play.

Alisa: Yes, most of our ideas are similar.

Allen: Maybe it is not necessary to do so much group work. Of all the group tasks I have completed, very few times do I feel like I have learned something. Not the academic ones, but the skills of presentation or things like that.

FC-Pre-FG-1

Students in both classes suggested in Pre-FG that *the workload of group work was too much*. It led to an adjustment to group work from one small task in each period to one larger project for each unit.

In terms of interactions with peers, *interesting* as it was to work with others and *friendly* and *cooperative* as peers were, it was challenging for most students to make effective interactions. *Limited class time for group work*, *time conflict between group members* after class, the *difficulty level of the tasks*, the *physical or social distance* between group members, or *gender differences* were frequently mentioned as hindrances to effective peer interaction in group work projects. The group size was another factor that affected student experience. Many students complained that *working in a group with 5 to 6 members was tiresome* and hoped to *work in smaller groups*. Some students were *so vexed or frustrated* at ineffective cooperation with peers that they claimed they *preferred to work individually*, which could be more reliable and controllable.

However, students further suggested that it was not so much whom they were working with as what they were working on that affected their engagement.

Students, on the whole, *preferred more interactive and creative tasks but* were unwilling to invest time and effort in preparation. They held that *collecting materials for the project was the easiest task* and *the least beneficial to learning*. Nonetheless, it was also *of limited benefit if the tasks were too challenging*. Students considered *tasks involving language output*, such as presenting the completed work orally, more challenging. There appeared to be *polarized differences* in students' attitudes towards presenting their work in class. Many *dreaded doing presentations in class*, associating it with negative feelings and experiences, while a few expressed their *eagerness to do that*, supporting it as effective in promoting learning. Participants in both classes admitted that their *inability to add new ideas to the topic* under discussion should account for part of their inertia in interaction. Most expressed the *need for a capable team leader, reliable group members, and more teacher support* to improve their experience in group work.

Peer assessment received the most negative comments of all the sub-tasks in group work. Participants in both FC and BC expressed their *dislikes for peer assessment*. Many students were *not interested in reading peers' work*. Most *did not see the value of peer assessment*, superficially assuming that only the assessed could benefit from the assessment. Therefore, they thought *peer assessment brought them nothing but an extra workload*. Besides, they *doubted the reliability of peer assessment*.

> *Ivan: I don't like being assessed by peers or assessing peers' work. I don't know what appropriate grades are, so I grade all groups similarly as long as something is submitted there. Even*

if I find someone didn't do his/her work, I don't give him/her a low grade. I felt uneasy about doing that.

Nora: There is always someone that doesn't do the assessment. Not sure what happens there.

Ivan: Our points go down if the assessor fails to do his/her job. I found our work was not assessed today and had to remind our assessor. That is really troublesome. It is even more troublesome to read peers' work.

Mediator: So you are not interested in reading peers' work?

Ivan: I would read peers' work more carefully if there is only one piece to assess. It is too much to read 5 or 6 pieces each time.

Tina: I lost my judgement when working on so many pieces of work.

Nora: Besides, everyone has his/her own criteria; some always give higher grades, and some much lower.

Bella: I guess that's why each piece has more than one assessor. Because we use different criteria, if each piece of work is to have only one assessor, work of similar quality may have very different grades.

Nora: Peer assessment is affected by a lot of subjective factors.

Ivan: So peer assessment is of no big use.

Bella: We rely more on teacher assessment.

All: Yes, that's true.

BC-Pre-FG-4

Still, there are minor differences in students' affective reactions to group work between FC and BC. Though both classes preferred tasks that were interactive/cooperative between group members, FC approved of tasks that required careful investigation, sound reasoning and language competence.

Sissi: Besides videos, the debate is the second most beneficial task to me. But it was pretty demanding.

Irene: I was utterly worn out after the debate.

Sissi: But it was the most rewarding.

Zoe: I think the debate is the most beneficial. It brings comprehensive practice, from finding resources on the topic to integrating all the resources into arguments. I've got a good understanding of the birth control policy and its effect on the country and the people.

Allen: I think videos are the most useful because they offer knowledge. Anyway, the debate is interesting.

Niamh: Debate is the most beneficial to me. I read a lot, wrote my ideas down to develop arguments and practised speaking. It is from these abundant practices that learning takes place. And I actually have had no impression of the other tasks.

FC-Post-FG-2

While participants from FC thought *the debate was the most beneficial* of all group work tasks, many students in BC felt it was too challenging for them. Instead, they *liked creative tasks more*.

Leon: I like role-play; I learned something from doing it and hope to do more work like this.

Alan: It was fun to add different elements into one work. We used sound effects in our play.

Nora: We video recorded our play and made a small video of it.

…

Alan: Debate is too difficult; I just couldn't understand what the other side was saying. Everyone is like I am just saying what I want to say, and what the others are saying has nothing to do with me.

Nora: I couldn't even follow my team members. We literally don't have the language proficiency for debate.

Alan: Both sides were just talking about their own ideas, and it is all up to luck which side was to win.

Tina: Even if I understood what the opposite side was talking about, I read through the text I stored on my cell phone beforehand when making an argument.

Alan: The ridiculous part was there were occasions I understood well and knew how to defend, but I couldn't express myself.

Nora: The same with me.

Tina: Debate could be a meaningful activity if we had the necessary English proficiency, but…

Alan: I couldn't even understand what the others were talking about.

BC- Post-FG-2

Though both classes preferred face-to-face communication to online interaction, BC, nevertheless, appeared to encounter more problems in peer interaction, many of which were not academic but communicative. Group collaboration in BC was *affected by more factors, such as students' initiatives and learning styles*, which were not mentioned as obstacles in FC.

4.8.2 Student Perceived Gains from Group Work

Just as there were more between-class similarities than divergences in students' affective reactions to group work in FC and BC, so were there in students' perceived gains. Many participants in both FC and BC reported expanded generic knowledge and enhanced interactive/collaborative skills, though much less knowledge and understanding of the subject content. They learned to *exploit group members' strengths* rather than slog away alone. Learner autonomy also developed in both classes. Students were able to regulate their tasks and coordinate between group members. Towards the second half of the semester, some groups learned to *keep working journals* in their online discussion groups, *recording their progress and adjustments so that everyone in the group could follow the project*. Many students developed the skills of effective online asynchronous communication.

> *Morgan: If it is impossible to communicate face to face, you need to organize your ideas and send them in a lump in the discussion group, then wait for your group members' response. It usually does not work if you try to exchange ideas bit by bit and get real-time responses because your group members may not be online as you are. You may get no response if you say anything that way in the online discussion group.*
>
> *BC-Post-FG-1*

In particular, group work, among all learning activities, contributes most to developing students' higher-order thinking skills.

Vicky: I was assigned to collect the survey results, so I used a table to record the data and reorganize them in order. I did the job, but the idea of organizing and integrating data was from Alisa.

FC-Pre-FG-2

Jane: Group interaction is a process of comparison. When I am reading our group members' contributions, I will try to find the similarities and differences between my work and theirs and make adjustments or supplements accordingly. I have also made many reflections, especially when I read work better than mine.

BC-Pre-FG-5

Students in both classes claimed that group work *broadened their vision* and *developed their way of thinking*. They learned to *design their tasks* and compare and evaluate work from their peers and their own. FC also reported changes in attitude in Post-FG. Many proved that they were *impressed with what they had learned from doing the tasks and* started appreciating *the benefit of learning by doing*.

It should be noted that the gains mentioned in focus groups were not corresponding in the two classes. For example, students in FC did not mention their gains in language skills in focus group discussions, and students in BC did not talk about any changes in attitude. Nonetheless, we cannot jump to the

conclusion that group work produced no effect in these aspects in each class as it is difficult to control what is being discussed in focus group discussions (Cohen et al., 2011).

4.8.3 Student Behavioural Engagement in Group Work

Similarly, diversity in learner autonomy in group work exists more across individual students than between classes. Most students prioritised *convenience* when interacting with peers, so they chose to *work with roommates or peers in the nearby dorms* when joining a group. Students' English proficiency influenced their performance. Some students could *use resources in English directly*; some had to *rely on their mother tongue*. Some groups *did everything in their mother tongue* and then *translated their final work into English*. Many students would try to perform their tasks well, but there were underperformance and inertia in each project. Some students *used whatever they found on the internet* and *would not bother to check the appropriateness or make any reorganization or revision*, so they produced work too irrelevant or too low in quality to be used in their project. There were occasions when group members *did not hand in their work in time*, resulting in their group's late submission of the whole project.

Students in BC, on the whole, regulated their tasks better. Most groups *submitted their work in time*. However, it should be noted that this better regulation may be just a result that group work in BC was done after class, and students could have enough time to work on their projects. Student autonomy in FC, however, was more diverse. Some students in FC were highly motivated.

They would *check and improve their work voluntarily* before sending it to peers, *follow and respond to teacher feedback closely* and try to learn from every opportunity. In contrast, some students were *unaware of teacher feedback at all* and could *never follow instructions for tasks* throughout the semester. Some claimed *they lost their agency* when working in groups because they tended to *rely on their group members*. Some noted that after group work was adjusted to a two-week project, they *forgot to finish their assigned tasks* and *lost enthusiasm for their work* as time went on.

Though most of them claimed they *worked harmoniously* with their group members, students in both classes admitted they *did not interact effectively or efficiently*. Primarily, it was because *they did not know how*. As students generally had zero experience in group projects, *getting their work started turned out particularly challenging* for most groups, especially at the beginning of the semester. In case the tasks were difficult to handle or group members were unfamiliar with each other, there could be *long lulls of silence*. Interaction between peers became *even less when the project schedule was tight*.

There were minor conflicts between group members from time to time. The most serious one resulted from group members (in FC) *misunderstanding the date of the debate* and *not making full preparation* and ended up with Micky *withdrawing from their group*. However, their group still managed to participate in the debate the next day. Most often, students *solved their problems by themselves via negotiation and compromise*. Interactions in BC appeared to be more problematic. Some students complained that *staying discreet to avoid offending others* when interacting with peers *was painful*. There were occasions

when students *revised peers' work without permission*, *made unfavourable comments on peers' work* without giving any reason or *refused to react to peers' feedback* without explaining why. In response, most students would *ignore peers' offensive words and behaviours for peace's sake*. It appears there lacked good communication among students for effective collaboration. Besides, some students in BC *never join online discussions of their groups*.

Students in the two classes collaborated with group members in quite similar manners. All groups *divided labour* within their groups. On most occasions, they *drew for tasks*. Many *took up different roles in different groups,* while some would *play a fixed role* in each group because they *excelled at a particular skill*. Groups *made working plans* for the whole project and tried to *follow their plans*. One shared problem with planning in both classes is that *their discussion centred on dividing labour* among group members and *setting up timetables for sub-tasks* rather than on how to address the tasks. Consequently, group members *produced work that did not match well*. Most students would share their work with group members when working on the tasks. Many would *read group members' contributions*, *exchange ideas, integrate their materials,* and *adjust or improve their work according to* peer feedback *before submission*. However, throughout the semester, there were always groups that just *pieced together whatever was available* and did *not revise their work*, especially *when the workload or the working schedule was tight*. Some group members *left the whole project behind when they finished their assigned tasks* and therefore *had no idea at all what their final work looked like*. There were occasions of circumventing tasks in both classes. In FC, this was more often a group

decision. However, in BC, it happened to individual students when they were *reluctant to improve their work*. Some went so far as to *shed their responsibility on peers* who suggested the improvement.

At the beginning of the semester, all groups *preferred to work face to face* and tried to find time and places to *work together*. They became more skilled in doing tasks online as the semester went on. One consequence was that students in FC would not fully use their classroom time. However, due to time conflicts between group members, most students *worked individually after class* rather than with peers. It turned out too late to fix it in the next class when they found someone had failed to complete his/her tasks as planned. Regulating the project effectively appeared always to be a challenge to FC.

Peer assessment posed as the most problematic in group work. In both classes, only a few students read peers' work carefully and did the *assessment by referring to the given criteria*. Most students *browsed through peers' work as quickly as possible*, *paying more attention to the form/pronunciation* than the content. Some would *check the beginning part only*. Most students *graded peers' work indiscriminately*. They usually *gave their own group members the highest grades* because individual performance affected the grade of the whole group. They *gave the same grade* to all the other groups or grades just a little lower in case of non-performance or inferior quality because peer assessment was not anonymous, and they *did not want to offend their classmates*. Seldom did students give comments along with grades, though encouraged to do so. Some students never took part in peer assessment.

4.9 Summary

In this qualitative phase, seven main themes were generated via thematic analysis of the data collected from focus group discussions and participant observation. Each main theme elaborates on student engagement in a specific learning activity in three sub-themes, depicting students' affective, cognitive and behavioural engagement respectively. Six themes captured differences to various extents in student engagement between FC and BC in some or all of the three dimensions, while the theme on group work projects depicted more between-class similarities than divergences across the three dimensions. The following is a summary of FC students' affective, cognitive and behavioural engagement in IEC based on a comparison with student engagement in BC.

4.9.1 FC Students' Affective Engagement in the Learning Activities

Students in FC were generally positive about the learning activities in IEC except for group work. They took videos as their primary knowledge resources and favoured grammar videos in particular. They used quizzes to measure the learning effects of video content. Students held that online forums helped solve their problems and motivated them to learn; they felt supported when interacting with peers. They consistently held that teacher feedback was more authoritative and reliable, solving the problems beyond them and guiding them on how to approach their problems. They adapted to individual exercises most quickly and held them the most effective and efficient to enhance their language knowledge and skills. Students in FC attributed the benefit of individual exercises partly to classroom group discussion, which, they believed, made

learning go deeper and broader. Among all the activities, group work attracted the participants' discussion most in the focus groups but is the only learning activity students did not receive favourably. Most participants perceived that the time and effort they spent on the work did not pay off.

On the whole, students' affective reactions were more divergent towards interactive activities than individual work. It was probably because interactive activities were not only affected by more factors but were inherently more complex. The qualitative data indicate that students' attitudes, perceptions and feelings towards the learning activities were also affected by their previous learning experiences, which largely contributed to their language proficiency and learning mentality. Most of the students learned English by doing a large amount of paper-and-pencil exercises and had little experience doing interactive tasks. Thus, they felt it challenging to make effective interactions, especially at the beginning of the semester. Students were eager to immerse themselves in doing paper-and-pencil exercises and were emotionally affected by their performance in the exercises. As the flipped classroom had more interactive activities, it explains why students in FC needed more time to adapt to the class. Though most students felt adapted to FC by midterm, some were not until the end of the semester. They felt that learning activities in FC lacked systematicity and could not find connections between the activities.

However, FC students were, on the whole, less negatively affected by the learning activities than BC. That was partly because students in FC had more face-to-face interactions and timely teacher feedback, thanks to the class design. Besides, they were less troubled by LMS, though they used LMS more.

In addition, they demonstrated less inclination to rely on teachers though they too held teacher support essential.

4.9.2 FC Students' Perceived Cognitive Gains in the Learning Activities

Participants in FC reported the most perceived gains in language knowledge and learning skills, mainly in Post-FG. Almost all the learning activities contributed to their improvement in these two skills. Individual exercises and classroom discussions were most conducive to enhancing language skills. The interactive activities were reported to promote more students' higher-order thinking skills. Interestingly, although students received group work least favourably, they reported more gains in higher-order thinking skills from group work than the rest of the learning activities in IEC. Learning activities in FC were more interactive, so there was more reported development in higher-order thinking skills. Students also reported changes in attitude towards learning; they came to agree that learning by doing was more effective. In addition, students in FC demonstrated more initiative in learning and better skills in using technology.

However, it is worth noting that though class design did make a difference in students' perceived gains from certain learning activities, some divergences were among individual students/student groups within the class rather than between the classes. For instance, students' perceived gains from forum discussions, teacher feedback, individual exercises and group work were similar in the two classes. However, the differences in student perceptions of the extent they benefited from these activities were apparent within the class.

4.9.3 FC Students' Behavioural Engagement in the Learning Activities

Students in FC demonstrated significant autonomy in most learning activities. Student autonomy appears to be associated with the extent to which they are familiar with the tasks. The more familiar the tasks, the more self-directed they were. Students demonstrated the greatest initiative in doing quizzes and individual exercises, which they were long accustomed to. Learner autonomy in the interactive activities of online forum discussion, classroom discussion and group work differed more across individual students than between classes. Students were least engaged behaviourally in teacher feedback, probably because the communication in it was primarily one way. In addition to the task type, student autonomy was also affected by their language proficiency and learning skills, especially their insufficient language output skills and limited competence in tool use.

Many students regulated their effort in IEC as per their LMS points. Some constantly compared points with peers. However, watchful as they were to their points, they cared less about their performance leading to these outcomes. They loathed doing peer assessment and paid little attention to the model work provided for them. They made more efforts to do tasks than to learn from the tasks. There were cases of task circumvention in watching videos or doing group work in addition to non-performance or inertia in discussion activities.

Compared with students in BC, students in FC were better prepared for and more engaged in class; they had less trouble with LMS and were more open to new tools. However, more students in FC were less organized in learning and

took longer to adapt to IEC. That was probably because FC was more different from the class students used to attend than BC. It is worth noting that, according to students in FC, the longer the gap between the submission date and the feedback, the less likely they would remember to check the feedback. It indicates that students' regulation of learning was also influenced by class management on the teacher's part.

Chapter 5: The Quantitative Findings

5.1 Introduction

The quantitative phase aimed to understand the overall effects of FC students' engagement in learning activities by examining their effort invested in the learning activities throughout the semester and the outcomes of their academic performance in the learning tasks and the final examination. The results are corroborated with qualitative findings. RQ2 and its sub-RQs are to be answered in this chapter:

> RQ 2. What are the overall effects of the learning activities on student engagement in the learning process and learning outcomes in FC?
>
> > RQ 2.1 To what extent is student engagement in the learning activities enhanced in FC?
> >
> > RQ 2.2 To what extent are students' learning outcomes enhanced in FC?

5.2 Enhanced Student Engagement in the Learning Activities

Statistical data on student engagement in video watching, quizzes, individual exercises, essay writing, online forum discussion and group work projects are reported in this section. A comparison between FC and BC was made to examine the differences in engagement between the two classes and seek evidence of enhanced engagement in FC. Student engagement in the sub-types of a specific learning activity was also checked. It was to help understand

how students in each class dealt with the learning activities and obtain further evidence of between-class differences. RQ 2.1 is answered in this section.

5.2.1 Enhanced Engagement in Recorded Instructional Videos

Throughout the semester, there were altogether 45 instructional videos posted in LMS for IEC. All the videos fell into three sub-groups: text videos (n = 26, accounting for 58% of all the videos), grammar videos (n = 11, 24%) and writing videos (n = 8, 18%). The ratio of the actual time a student spent watching a video to the length of the related video ($R_{w/l}$) was used to measure student engagement in video watching. Figure 5.1 shows a marked difference between FC and BC in student engagement in the three sub-types of videos. A further examination of the mean scores shows that all the means of $R_{w/l}$ of the three video sub-groups are > 1 in FC, indicating that students in FC watched the videos more than once on average. In contrast, all the means of $R_{w/l}$ in BC are < 1, indicating that students in BC generally did not finish watching the videos. Further comparison of the mean scores indicated differences in student engagement with videos between the two classes and within each class.

The normality test showed the data were not normally distributed, so the Mann-Whitney Wallis statistic was conducted to determine the differences (Cohen et al., 2011) in video watching between FC and BC in the three sub-groups of writing videos, grammar videos and text videos. The results in Table 5.1 show statistically significant differences in video watching between FC and BC in all the three sub-groups, each with a strong effect size indicating a big difference in FC and BC in effect (Cohen, 1988). FC significantly outperformed BC in their

engagement in writing videos (U = 71, z = -4.97, *p* = .00, r = .68), grammar videos (U = 77, z = -4.86, *p* = .00, r = .67) and text videos (U = 95, z = -4.54, *p* = .00, r = .62).

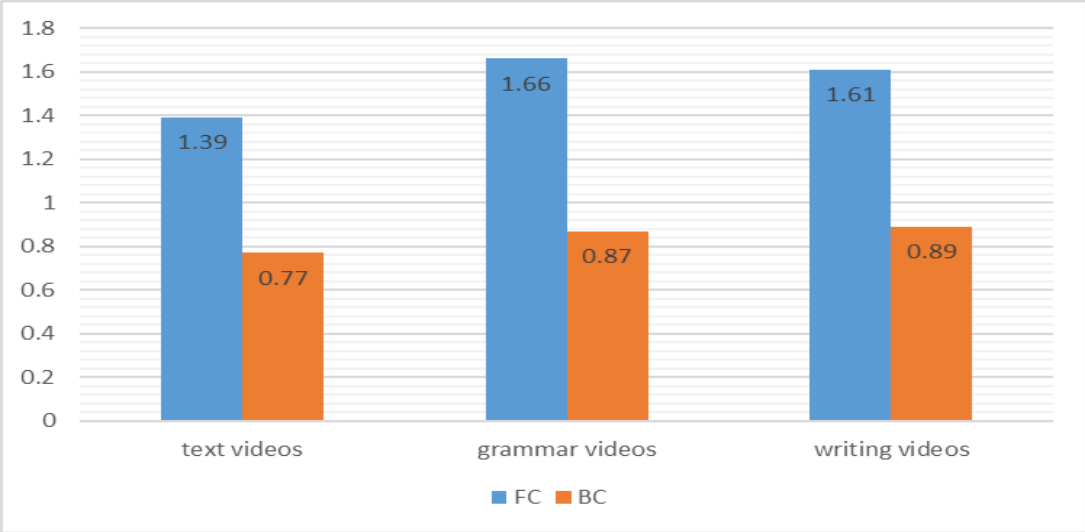

Figure 5. 1 Rate of time on video watching to video length ($R_{w/l}$) to indicate student engagement in the three sub-types of videos

Videos ($R_{w/l}$)	$M_{FC}(SD_{FC})$ (n = 25)	$M_{BC}(SD_{BC})$ (n = 28)	Mann-Whitney U score	Z-score	*p*	*r*
Writing Video	1.61(0.40)	0.89 (0.45)	71.00	-4.97	.00	.68
Grammar Video	1.66 (0.57)	0.87 (0.50)	77.00	-4.86	.00	.67
Text Video	1.39 (0.47)	0.77 (0.34)	95.00	-4.54	.00	.62

Table 5. 1 Mann-Whitney tests to compare student engagement in video watching between FC and BC

The Friedman tests were run to investigate further whether there was any statistically significant difference between student engagement in the three sub-groups of videos within each class. A statistically significant difference (shown in Table 5.2) was found between student engagement in the subgroups of videos in FC ($\chi^2 = 15.44$, *p* = .00), with a relatively low coefficient of concordance (Kendall's W = .31). The Wilcoxon tests post hoc tests showed a statistically significant difference in student engagement between text videos and writing videos (Z = -3.40, *p* = .001) and that between text videos and

grammar videos (Z = -3.51, p = .000) at the significance level of p = .017 while no statistically significant difference in student engagement between grammar videos and writing videos (z = -0.26, p = .798). In contrast, though the Friedman test indicated a statistically significant difference between BC students' engagement in the three sub-groups of videos (χ^2 = 7.79, p = .02) with a low coefficient of concordance (Kendall's W = .14), the Wilcoxon tests indicated no statistically significant difference between sub-groups of videos at the significance level of p = .017 (Z = -0.46, p = .649; Z = -2.35, p = .019; Z = -1.89, p = .059).

	Chi-Square	df	Asymp. Sig.	Kendall's W[a]	Wilcoxon Signed Ranks Test Z (2-tailed Asymp. Sig.[*])		
					G-W[b]	T-W[c]	T-G[d]
FC (n = 25)	15.44	2	.00	.31	-0.26 (.798)	-.3.40 (.001*)	-3.51 (.000*)
BC (n = 28)	7.79	2	.02	.14	-0.46 (.649)	-2.35 (.019)	-1.89 (.059)

a. Kendall's Coefficient of Concordance
b. Comparison between grammar videos (G) and writing videos (W)
c. Comparison between text videos (T) and writing videos
d. Comparison between text videos and grammar videos
* Significance level p = .05/3 = .017 after Bonferroni correction as the Wilcoxon Test is repeated for three paired groups

Table 5. 2 Friedman tests to compare student engagement in three sub-groups of videos

The quantitative results corroborated the focus group findings in the following aspects: first, students in FC invested significantly more time and effort in videos. The focus group findings explained that it was because FC students took videos as their primary sources of knowledge. However, students in BC used videos as supplements to lectures, so they would not watch videos when they felt they had learned well from lectures. The log data in LMS show that some students in BC never watched videos. It should be added that the

qualitative results indicate FC students watched videos not only before class to learn what is new but were more likely to turn to videos for revision purposes to reinforce what they had learned than BC students. That again may confirm that FC students took videos as the primary sources of knowledge. Second, students in both classes paid particular attention to grammar videos. The statistics added to the qualitative findings that students spent more time on grammar and writing videos than on text videos. However, the two classes differed in that FC students spent significantly more time on grammar and writing videos than text videos, but such differences were not significant in BC. This finding may confirm the qualitative results that, while students in both classes attached particular importance to language knowledge, students in BC learned knowledge primarily from lectures and used videos only as supplements, so many of them were likely to just run over the videos.

5.2.2 Enhanced Engagement in Independent work

As quizzes (n = 22), individual exercises (n = 12), and essay writing (n = 2) were to be completed independently by students, they were grouped together in this section for the sake of convenience to report data. Figure 5.2 shows the differences between FC and BC in students' average completion rates of these independent exercises. The mean scores of the completion rate show that students in FC completed all their quizzes ($M_{\%}$ = 100, SD = 0.00) and essays ($M_{\%}$ = 100, SD = 0.00) expected of them, while students in BC did all the individual exercises ($M_{\%}$ = 100, SD = 0.00).

The Mann-Whitney tests (Table 5.3) indicated there was a statistically significant difference between FC and BC in students' completion rate of quizzes (U = 225.00, Z = -3.27, *p* = .001) with a relatively big effect size (r = .45) and individual exercises (U = 280.00, Z = -2.46, *p* = .014) with a medium effect size (r = .34). However, student engagement in essay writing was not significantly different between the two classes in either completion rate (U = 337.50, Z = -0.95, p = .13) or revision attempts (U = 247.00, Z = -1.84, p = .066).

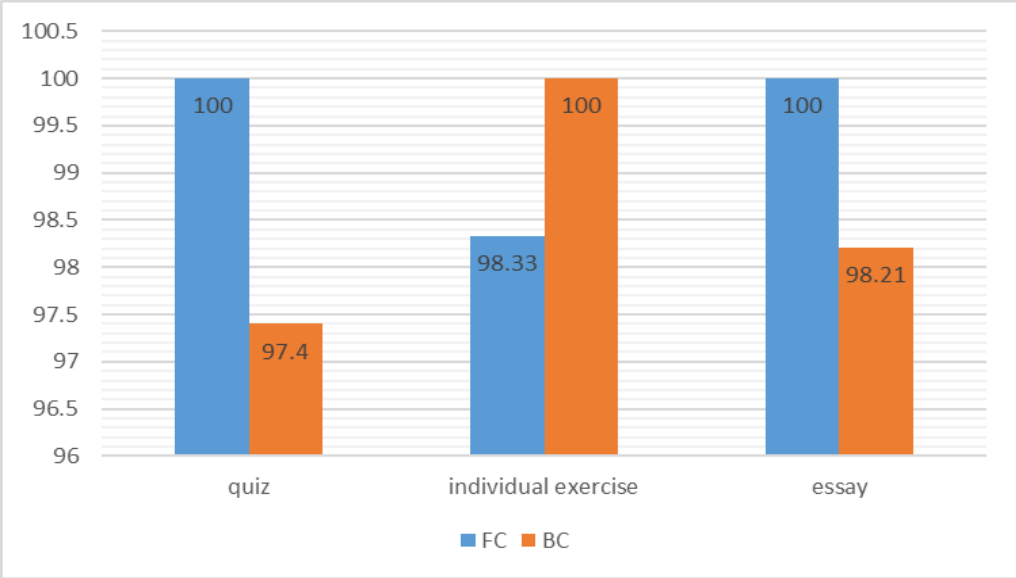

Figure 5. 2 Students' average completion rate of independent exercises in IEC

Completion rate & Essay attempts	$M_{FC}(SD_{FC})$ (n = 25)	$M_{BC}(SD_{BC})$ (n = 28)	Mann-Whitney U score	Z-score	*p*	*r*
Quiz (%)	100 (0.00)	97.40 (4.00)	225.00	-3.27	.001	.45
Individual exercises (%)	98.33 (3.40)	100 (0.00)	280.00	-2.46	.014	.34
Essay (%)	100 (0.00)	98.21 (9.45)	337.50	-0.95	.345	.13
Essay attempts (n)	7.80 (4.29)	10.32 (5.74)	247.00	-1.84	.066	.25

Table 5. 3 Mann-Whitney tests to compare student engagement in independent work between FC and BC

Friedman tests (in Table 5.4) indicated statistically significant differences in students' completion rates between the three tasks in both FC ($\chi^2 = 10.00$, p = .007, Kendall's W = .20) and BC ($\chi^2 = 15.94$, p = .000, Kendall's W =.29). Nonetheless, the Wilcoxon post hoc tests showed the only statistically significant difference at the significant level of p = .017 was between students' completion rate of individual exercises and quizzes in BC (Z = -2.86, p = .004).

	Chi-Square	**df**	**Asymp. Sig.**	**Kendall's W[a]**	**Wilcoxon Signed Ranks Test** Z (2-tailed Asymp. Sig.[*])		
					IE- Q[b]	**E- Q[c]**	**E-IE[d]**
FC (n = 25)	10.00	2	.007	.20	-2.24 (.025)	-0.00 (1.00)	-2.24 (.025)
BC (n = 28)	15.94	2	.000	.29	-2.86 (.004*)	-1.81 (.070)	-1.00 (.317)

a. Kendall's Coefficient of Concordance
b. Comparison between individual exercises (IE) and quizzes (Q)
c. Comparison between essays (E) and quizzes
d. Comparison between essays and individual exercises
* Significant level p = .05/3 = .017 after Bonferroni correction as the Wilcoxon Test is repeated for three paired groups

Table 5. 4 Friedman tests to compare student completion of independent tasks

The quantitative results indicate that students in FC were consistently actively engaged in the quizzes, individual exercises and essays, while students in BC were significantly less engaged in quizzes. The findings corroborate the qualitative results that students in FC were positively engaged in pre-class quizzes while some students in BC failed to finish quizzes for various reasons. The quantitative data add to the qualitative findings that even though the students in BC encountered more difficulties doing individual exercises after class, they all completed the exercises as required. In contrast, some students in FC did not submit their individual exercises, which, as the focus group

discussions suggested, probably resulted from the in-class time limit for the individual exercises in FC.

5.2.3 Enhanced Engagement in Online Forum Discussion

FC and BC had forum discussions in the same number and types over the semester. There were altogether 51 discussions for the online forum in four sub-types in each class: student-led discussions (n = 25), teacher-led exercise discussions (n = 14), teacher-led topic discussions (n = 6) and student exemplar work discussions (n = 6). Each student's average postings per discussion ($R_{p/d}$) was used to measure student engagement in the online forum. Figure 5.3 visualizes student engagement in each sub-type of online forum discussions.

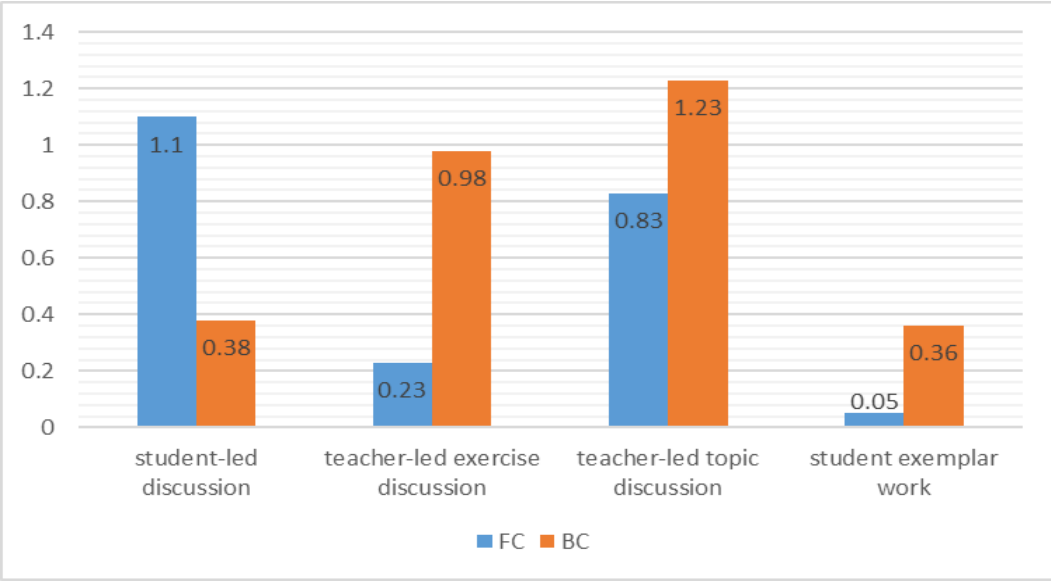

Figure 5. 3 Students' average postings on a topic ($R_{p/d}$) to indicate student engagement in the online forum discussion in IEC

A careful examination of the mean scores for $R_{p/d}$ (in Table 5.5) indicates that students in FC and BC differed in their engagement in the four sub-types of online forums. Students in FC were more interactive with peers. They were most engaged in student-led discussion (M = 1.10, SD = 0.56) and received far more likes from peers (M = 70.24, SD = 28.21). Students in BC tended to follow the teacher. They contributed more postings in teacher-led topic discussion (M = 1.23, SD = 0.45), teacher-led exercise discussion (M = 0.98, SD = 0.72) and student exemplar work (M = 0.36, SD = 0.48). Both paid the least attention to exemplar work from peers (M_{FC} = 0.05, SD_{FC} = 0.10; M_{BC} = 0.36, SD_{BC} = 0.48).

$R_{p/d}$ & likes (n)	**Student-led Discussion**	**Teacher-led Exercise Discussion**	**Teacher-led Topic Discussion**	**Student Exemplar Work**	**Total Postings**	**Likes (n)**
M_{FC}(SD_{FC}) (n = 25)	1.10 (0.56)	0.23 (0.26)	0.83 (0.33)	0.05 (0.10)	0.71 (0.32)	70.24 (28.21)
M_{BC}(SD_{BC}) （n = 28）	0.38 (0.39)	0.98 (0.72)	1.23 (0.45)	0.36 (0.48)	0.64 (0.42)	8.54 (6.52)

Table 5. 5 Descriptive statistics of student engagement in the online forum

The T-test for independent samples (in Table 5.6) showed that the mean score of $R_{p/d}$ for the total postings in FC (M = 0.71, SD = 0.32) did not differ significantly (t = 0.67, df = 51, *p* = .50, d = .19, CI = -0.35, 0.73) from that in BC (M = 0.64, SD = 0.42). The Mann-Whitney tests (in Table 5.7) showed that the two classes were statistically significantly different from each other in their engagement in each sub-type of online forum interaction. Students in FC placed significantly more postings in student-led discussion (U = 81.0, Z = -4.80, *p* = .000) with a big effect size (r = .66). Meanwhile, they received significantly more likes from peers (U = 0.00, Z = -6.24, *p* = .000) with a very big effect size (r = .85). In contrast, students in BC posted statistically significantly

more in teacher-led exercise discussion (U = 104, Z = -4.40, *p* = .000, r = .60), teacher-led topic discussion (U = 151, Z = -3.59, *p* = .000, r = .49) and student exemplar work (U = 229, Z = -2.48, *p* = .013, r = .34), with at least medium effect sizes (Cohen, 1988).

$R_{p/d}$	M(SD)	CI	t (df)	*p* (2-tailed)	Cohen's *d*	CI (*d*)
FC (n = 25)	0.71 (0.32)	-0.14, 0.27	0.67 (51)	.50	.19	-0.35, 0.73
BC (n = 28)	0.64 (0.42)					

Table 5. 6 T-test for independent samples to compare total forum postings between FC and BC

$R_{p/d}$ & likes (n)	M_{FC} (SD_{FC})	M_{BC} (SD_{BC})	Mann-Whitney U score	Z-score	*p*	r
Student-led Discussion	1.10 (0.56)	0.38 (0.39)	81.00	-4.80	.000	.66
Teacher-led Exercise Discussion	0.23 (0.26)	0.98 (0.72)	104.00	-4.40	.000	.60
Teacher-led Topic Discussion	0.83 (0.33)	1.23 (0.45)	151.00	-3.59	.000	.49
Student Exemplar Work	0.05 (0.10)	0.36 (0.48)	229.00	-2.48	.013	.34
Likes (n)	70.24 (28.21)	8.54 (6.52)	.000	-6.24	.000	.85

Table 5. 7 Mann-Whitney tests to compare student engagement in sub-types of forum discussion between FC and BC

The Friedman tests (shown in Table 5.8) were run to further investigate whether there was any statistically significant difference in student postings between the four sub-types within each class. A statistically significant difference was found in FC (χ^2 = 64.79, *p* = .000), but with a high coefficient of concordance (Kendall's W = .86). The Wilcoxon tests post hoc tests showed statistically significant differences in FC students' postings between teacher-led exercise discussion and student-led discussion (Z = -4.37, *p* = .00), student exemplar work discussion and student-led discussion (Z = -4.37, *p* = .00), teacher-led topic discussion and teacher-led exercise discussion (Z = -4.35, *p* = .00), student exemplar work discussion and teacher-led exercise discussion (Z = -

3.63, p = .00) and student exemplar work discussion and teacher-led topic discussion (Z = -4.39, p = .00) at the significance level of .008. There was no significant difference between teacher-led topic discussion and student-led discussion (Z = -2.00, p = .45). The results indicated that students in FC engaged similarly in student-led discussion and teacher-led topic discussion. However, they placed significantly fewer postings in teacher-led exercise discussion and the least in student exemplar work.

In BC, a similar statistically significant difference was found in student postings between the four sub-types of online forums (χ^2 = 59.46, p = .000), still with a high coefficient of concordance (Kendall's W = .71). The Wilcoxon post hoc tests showed statistically significant differences in student postings between teacher-led exercise discussion and student-led discussion (Z = -4.27, p = .00), teacher-led topic discussion and student-led discussion (Z = -4.60, p = .00), student exemplar work discussion and teacher-led exercise discussion (Z = -4.08, p = .00) and student exemplar work discussion and teacher-led topic discussion (Z = -4.63, p = .00) at the significance level of .008 while no statistically significant difference between student exemplar work discussion and student-led discussion (Z = -0.36, p = .72), and teacher-led topic discussion and teacher-led exercise discussion (Z = -2.37, p = .02) at the significance level of .008. In brief, students in BC engaged in teacher-led topic discussion and exercise discussion to a similar extent, but they were statistically significantly less engaged in student-led discussion and student exemplar work.

	Chi-Square	df	Asymp. Sig.	Kendall's W[a]	Wilcoxon Signed Ranks Test Z (2-tailed Asymp. Sig.[*])					
					TchE -Stu[b]	TchT -Stu[c]	EW-Stu[d]	TchT-TchE[e]	EW-TchE[f]	EW-TchT[g]
FC (n = 25)	64.79	3	.000	.86	-4.37 (.00)*	-2.00 (.45)	-4.37 (.00)*	-4.35 (.00)*	-3.63 (.00)*	-4.39 (.00)*
BC (n = 28)	59.46	3	.000	.71	-4.27 (.00)*	-4.60 (.00)*	-0.36 (.72)	-2.37 (.02)	-4.08 (.00)*	-4.63 (.00)*

a. Kendall's Coefficient of Concordance
b. Comparison between teacher-led exercise discussion (TchE) and student-led discussion (Stu)
c. Comparison between teacher-led topic discussion (TchT) and student-led discussion
d. Comparison between student exemplar work discussion (EW) and student-led discussion
e. Comparison between teacher-led topic discussion and teacher-led exercise discussion
f. Comparison between student exemplar work discussion and teacher-led exercise discussion
g. Comparison between student exemplar work discussion and teacher-led topic discussion

* Significance level $p = .05/6 = .008$ after Bonferroni correction as the Wilcoxon Tests were repeated for six paired groups

Table 5. 8 Friedman tests to compare student engagement in the four types of online forums

The quantitative data complemented the focus group findings that, though there appeared no significant overall between-class difference in student engagement in forum discussion, students in FC and BC were strikingly different in their preference for and engagement in the four sub-types of discussion. Students in FC were significantly more interactive with their peers in forum discussions. They were most engaged in the student-led discussion, which may explain why they received many more likes from their peers than did students in BC. They were as engaged in teacher-led topic discussions but far less engaged in the teacher-led exercise discussion. That was probably because most of their problems had been settled in discussions led by themselves and the classroom face-to-face group discussions. They rarely placed postings in the forum that demonstrated exemplary work from peers. Conversely, students in BC were more teacher-led. They significantly

outperformed their peers in FC in all teacher-led discussions, triangulating the focus group findings that students in BC were more teacher-centric. As such, they interacted far less with their classmates in the online forums. Like their peers in FC, students in BC paid the least attention to the exemplar work from their peers, which triangulated the qualitative findings that students, on the whole, were not making good use of feedback (e.g. learning from the exemplary work, reflecting on their own performance and identifying areas for improvement in their work) that aimed to improve their performance. It is worth noting that all the forum discussions were optional; that is, the students mobilized the forums voluntarily. Thus, the data collected from forum discussions, compared with those from required activities, can tell more truthfully how the students reacted differently to the four sub-types of activities and, in turn, how the class design made a difference in student engagement in learning.

5.2.4 Enhanced Engagement in Group Work

Students in both classes did five group work projects in the semester. The mean scores of student completion rate in Table 5.9 indicate that students in BC completed all the projects as expected and did more within-group peer assessments. However, the results of Mann-Whitney tests (Table 5.9) show there was no statistically significant difference between FC and BC in students' completion rate of the projects (U = 336.00, Z = -1.06, p = .29, r = .15), students' attempts in between-group peer assessment (U = 318.50, Z = -0.56, p = .57, r = .08) or within-group peer assessment (U = 251.00, Z = -1.77, p = .08, r =.24), with small effect sizes.

Completion rate & peer assessment	$M_{FC}(SD_{FC})$ (n = 25)	$M_{BC}(SD_{BC})$ （n = 28）	Mann-Whitney U score	Z-score	*p*	*r*
Completion rate %	99.2 (4.00)	100 (0.00)	336.00	-1.06	.29	.15
Between-group assessment (n)	9.96 (5.04)	9.14 (7.71)	318.50	-0.56	.57	.08
Within-group assessment (n)	11.48 (4.51)	13.43 (5.12)	251.00	-1.77	.08	.24

Table 5. 9 Mann-Whitney tests to compare student engagement in group work between FC and BC

That said, students in each class treated peer assessment in different ways. A t-test for paired samples (Table 5.10) showed no statistically significant difference (t[24] = -1.58, *p* = .13, CI = -3.50, 0.46) between FC students' attempts of between-group peer assessment (M = 9.96, SD = 5.04) and those of within-group peer assessment (M = 11.48, SD = 4.51), with a modest effect size (*d* =.32, CI = -0.88, 0.24). In contrast, the Wilcoxon test indicated a statistically significant difference (Z = -2.57, *p* = .01, r = .34) between BC students' attempts in between-groups peer assessment (M = 9.14, SD = 7.71) and those in within-group peer assessment (M = 13.43, SD = 5.12). Students in BC assessed their own work significantly more than their classmates' work in other groups.

Peer assessment attempts	M(SD) (n = 25)	CI	t (df)	*p* (2-tailed)	Cohen's *d*	CI (*d*)
Between-group assessment (n)	9.96 (5.04)	-.3.50, 0.46	-1.58 (24)	.13	.32	-0.88, 0.24
Within-group assessment (n)	11.48 (4.51)					

Table 5. 10 T-test for paired samples to compare student engagement in peer assessment in FC

Peer assessment attempts	M (SD) (n = 28)	Z-score	*p* (2-tailed)	*r*
Between-group assessment (n)	9.14 (7.71)	-2.57	.01	.35
Within-group assessment (n)	13.43 (5.12)			

Table 5. 11 Wilcoxon test to compare student engagement in peer assessment in BC

The quantitative findings triangulated the focus group results that students in FC and BC participated in group work very similarly on the whole. Students' completion rates of group projects were high in both classes, with students in BC having a full completion rate. There was no significant difference in students' attempts at peer assessment between the two classes. Both engaged more in assessing work within their own groups than the work from the other groups. However, the quantitative data revealed that such an orientation was statistically significant in BC, indicating again students in BC were more introverted. Class design in BC can affect the way students do their group work. As class time was mostly taken up by lectures, group work in BC mostly went online after class, depending much upon students' initiative. Hence, assessing peers' work from the other groups was more likely to be left behind by students due to their lack of interest in peers' work or their intention to reduce workload, as the focus group discussion indicated.

5.3 Students' Enhanced Academic Outcomes

Student grades from quizzes, individual exercises, essays, group work projects and final examinations were collected and compared for evidence of enhanced learning outcomes in FC. Factors that predict student academic outcomes in the final examinations were further investigated. RQ 2.2 is answered in this section.

5.3.1 Improved Academic Outcomes

Statistical analyses of student grades indicated no significant difference in students' grades for the final examination and essays between the two classes.

However, the two classes differed in grades for the other learning tasks. FC performed better than BC in quizzes and individual exercises but did not do as well in group work projects. Figure 5.4 visualizes student grades in all the tasks and the final examination in the semester. The t-tests for independent samples (Table 5.12) showed the mean score for quizzes in FC (M = 84.14, SD = 5.63) was statistically significantly higher (t = 2.64, df = 51, *p* = 0.11, CI = 0.85, 6.23) than that in BC (M = 80.58, SD = 4.14) with a moderate effect (*d* = .73, CI = 0.17, 1.28). The mean score for individual exercises in FC (M = 88.36, SD = 4.44) was statistically significantly higher (t = 3.62, df = 51, *p* = .001, CI = 1.74, 6.08) than that in BC (M = 84.45, SD = 3.41) with a fairly strong effect (*d* = .96, CI = 0.39, 1.53). However, the mean score for group work projects in FC (M = 86.53, SD = 2.47) was statistically significantly lower (t = -2.56, df = 51, *p* = .014, CI = -2.76, -0.33) than that in BC (M = 88.08, SD = 1.93) with a moderate effect (*d* = .70, CI = 0.14, 1.26). The t-tests also showed there was no statistically significant difference in the mean scores for essays (t = -0.45, df = 40.98, *p* = .656, CI = -1.96, 1.25) between FC (M = 80.60, SD = 3.36) and BC (M = 80.96, SD = 2.23), but with a weak effect (d = .12, CI = -0.42, 0.66). Neither was there a statistically significant difference in the mean scores for the final examination (t = -0.35, df = 51, *p* = .726, CI = -5.88, 4.13) between FC (M = 71, SD = 9.25) and BC (M = 71.88, SD = 8.82), with a weak effect (d = .10, CI = -0.44, 0.64).

When further examining the sub-sections of the final examination, the t-tests for independent samples (Table 5.12) and Mann-Whitney tests (Table 5.13) did not show a statistically significant difference between the two classes in all the sub-

sections. Figure 5.5 presents student grades in each sub-section of the final examination.

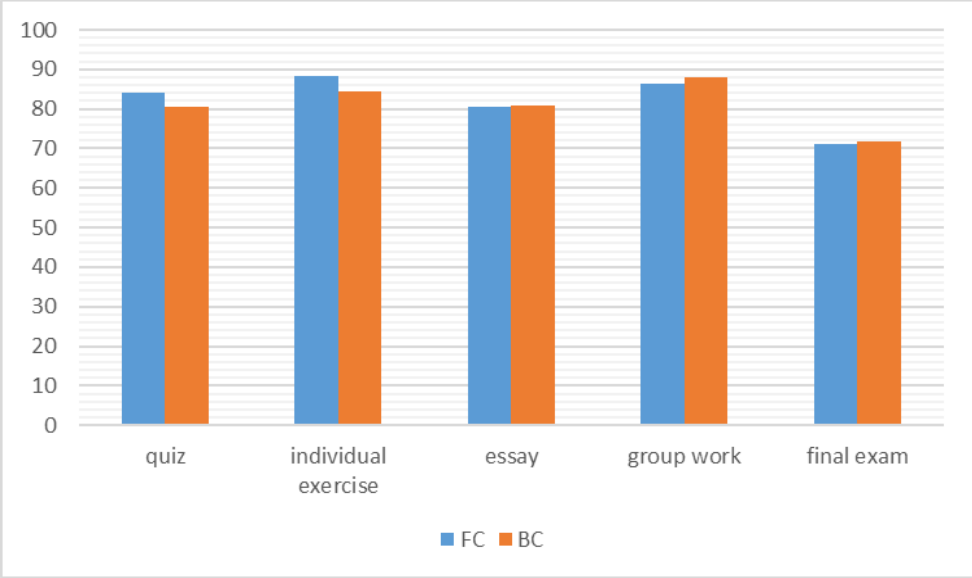

Figure 5. 4 Student grades for all the tasks and examinations in the semester

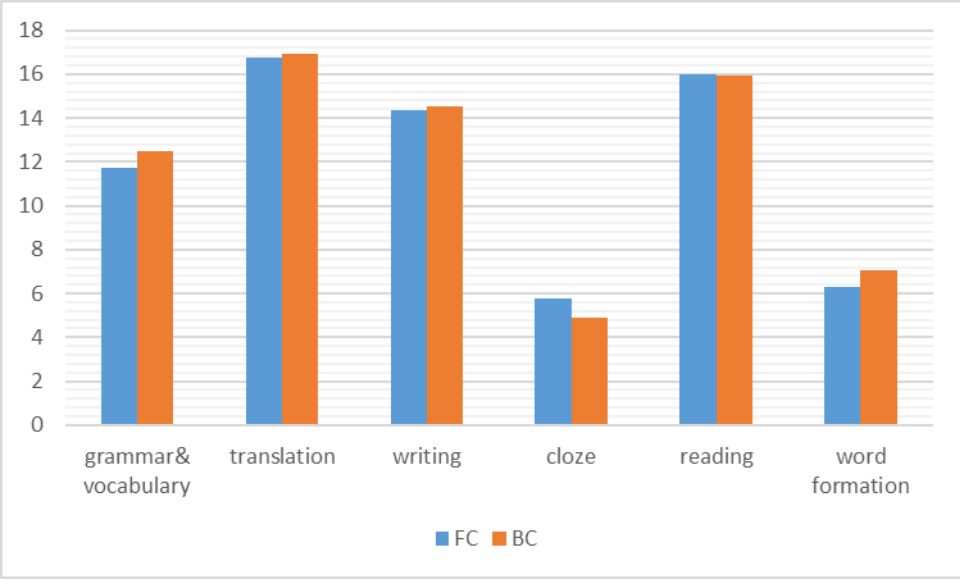

Figure 5. 5 Student grades in each sub-section in the final examination

The quantitative results showed no statistically significant difference in students' final examination outcomes between FC and BC, in either the total or sub-

section scores. Nonetheless, it should be noted that FC scored statistically significantly lower in essay writing than BC in the placement test, though the two classes did not significantly differ in the total scores (shown in Table 3.1). Therefore, it may be inferred that FC improved in essay writing over the semester. As learning records show, FC performed as well as BC in both essays and significantly outperformed BC in quizzes and individual exercises during the semester.

Grades	$M_{FC}(SD_{FC})$ n = 25	$M_{BC}(SD_{BC})$ n = 28	CI	t (df)	*p* (2-tailed)	Cohen's *d*	CI (*d*)
Quiz	84.14 (5.63)	80.58 (4.14)	0.85, 6.23	2.64 (51)	.011	.73	0.17, 1.28
Individual exercise	88.36 (4.44)	84.45 (3.41)	1.74, 6.08	3.62 (51)	.001	.96	0.39, 1.53
Essay	80.60 (3.36)	80.96 (2.23)	-1.96, 1.25	-0.45 (40.98)	.656	.12	-0.42, 0.66
Group work	86.53 (2.47)	88.08 (1.93)	-2.76,-0.33	-2.56 (51)	.014	.70	0.14, 1.26
Final examination	71.00 (9.25)	71.88 (8.82)	-5.88, 4.13	-0.35 (51)	.726	.10	-0.44, 0.64
Gr. & Vo.	11.72 (1.93)	12.50 (2.44)	-1.99, 0.43	-1.30 (50.28)	.200	.36	-0.19, 0.90
Translation	16.78 (2.01)	16.93 (1.73)	-1.18, 0.88	-0.29 (51)	.774	.08	-0.46, 0.62
Writing	14.38 (2.21)	14.52 (1.89)	-1.27, 0.99	-0.25 (51)	.807	.07	-0.47, 0.61

Table 5. 12 T-tests for independent samples to compare student grades between FC and BC

Grades	$M_{FC}(SD_{FC})$ (n = 25)	$M_{BC}(SD_{BC})$ （n = 28）	Mann-Whitney U score	Z-score	*p*	*r*
Cloze	5.80 (1.91)	4.89 (2.47)	270.00	-1.44	.15	.20
Reading	16.00 (2.27)	15.96 (3.11)	338.00	-0.22	.83	.03
Word formation	6.32 (2.87)	7.07 (2.46)	300.00	-0.90	.37	.12

Table 5. 13 Mann-Whitney tests to compare student grades between FC and BC

5.3.2 Contributing Factors of Academic Outcomes

The Pearson test and Spearman test (in Table 5.14) indicated that FC students' grades in their final examinations were correlated statistically significantly with

their grades for the placement tests (p = .005, r = .542) and their individual exercise grades (p = .004, r = .560). The Pearson test and the Spearman test (in Table 5.15) showed that BC students' grades in their final examinations were correlated statistically significantly with their grades in the placement test (p = .011, r = .476), quizzes (p = .008, r = .494) and individual exercises (p = .001, r = .583). The correlation coefficient is used as the effect size in correlation data, where a coefficient of correlation between .5 and .8 is interpreted as a strong positive effect, and one between .3 and .5 is interpreted as a moderate positive effect (Cohen et al., 2011).

Correlation (ρ)	**r M (SD)**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1. Final examination grades	71.00 (9.25)	1										
2. Placement test grades	75.64 (8.28)	.542**	1									
3. Video attempts (Rw/l)	1.55 (0.45)	.088	-.432*	1								
4. Forum attempts (Rp/d)	0.71 (0.32)	.138	.027	.322	1							
5. Essay attempts	7.8 (4.29)	.200	-.066	.324	-.321	1						
6. Peer assessment	21.44 (8.27)	.293	.467*	-.097	-.054	.282	1					
7. Quiz grades	84.14 (5.63)	.226	.246	-.017	.161	.300	.068	1				
8. Individual exercise grades	88.36 (4.44)	.560**	.340	.240	.124	.405*	.373	.384	1			
9. Essay grades	80.60 (3.36)	.357	.134	.256	-.012	.458*	.354	.237	.167	1		
10. Group work grades	86.53 (2.47)	.129	.365	-.130	.017	.154	.099	.233	.011	.088	1	
11. Independent work completion rate (ρ)	0.99 (0.02)	-.065	-.099	.109	.044	.252	.207	.107	.206	.327	-.266	1

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

Table 5. 14 Correlation tests of student performance in learning activities in FC

Correlation r (ρ)	M (SD)	1	2	3	4	5	6	7	8	9	10	11
1. Final examination grades	71.88 (8.82)	1										
2. Placement test grades	78.40 (9.40)	.476*	1									
3. Video attempts (Rw/l)	0.84 (0.41)	-.102	-.230	1								
4. Forum attempts (Rp/d)	0.64 (0.42)	.013	-.283	.584**	1							
5. Essay attempts	10.32 (5.74)	-.148	-.244	.200	.399*	1						
6. Peer assessment	22.57 (10.34)	-.026	.226	.173	-.061	-.101	1					
7. Quiz grades	80.58 (4.14)	.494**	.449*	.115	.320	.016	-.031	1				
8. Individual exercise grades	84.45 (3.41)	.583**	.210	.335	.248	-.071	.212	.650**	1			
9. Essay grades	80.96 (2.23)	.075	-.004	.183	.184	.011	.336	.421*	.108	1		
10. Group work grades	88.08 (1.93)	.-.048	.031	.251	.213	.115	.356	-.031	.216	.108	1	
11. Independent work completion rate (ρ)	0.99 (0.03)	.143	-.094	.209	.453*	.006	.302	.313	.556**	.268	.204	1

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

Table 5. 15 Correlation tests of student performance in learning activities in BC

Multiple linear regression was then applied to model the influence of the individual exercise grade and the placement test grade on students' final examination outcomes in FC. The two variables entered the model in a stepwise sequence (Table 5.16). The adjusted R square in the model for FC (R^2 = .40, F[2,22] = 9.12, *p* = .001) indicated moderate goodness of fit of the model (Cohen et al., 2011). The beta weighting of the variable *Individual exercise grades* (β = .42, t = 2.53, *p* = .019) and that of the variable *Placement test grades* (β = .40, t = 2.37, *p* = .027) in FC indicated a moderate effect size of the two predictor variables (Cohen et al., 2011) for FC students' final examination grades. Multiple linear regression in BC (Table 5.16) showed that the individual exercise grade and the placement test grade entered the model in a stepwise sequence, but the quiz grade did not. The adjusted R square in the

model for BC (R^2 = .428, F[2,25] = 11.12, p < .001) indicated moderate goodness of fit of the model (Cohen et al., 2011). The beta weighting of the variable *Individual exercise grades* in BC (β = .51, t = 3.40, *p* = .002) indicated a strong effect size of the predictor variable for BC students' final examination grades. The beta weighting of the variable placement test grades in BC (β = .37, t = 2.49, *p* = .02) indicated a moderate effect size of the predictor variable for students' final examination grades.

variances	Final examination grades					
	FC (n = 25)			BC (n = 28)		
	β (standardized)	95% CI	VIF	β (standardized)	95% CI	VIF
Constant	-40.77	-101.53, 19.99	-	-65.63	-131.64, 0.38	-
Individual exercise grades	.42*	0.16, 1.61	1.131	.51**	0.52, 2.1	1.046
Placement test grades	.40*	0.06, 0.83	1.131	.37*	0.06, 0.63	1.046
R^2 (adjusted)	.40			.43		
F	(2,22) 9.12**			(2,25) 11.12**		
D-W	1.545			1.797		

***p* <.05; ***p* <.01**

Table 5. 16 Multiple linear regression to predict contributing factors of the final examination outcomes

The regression models in Table 5.16 show that in both classes, students' placement test grades and their individual exercise grades can significantly predict their final examination outcomes, with individual exercise grades exerting a bigger influence on the final examination outcomes. Table 5.14 indicates that FC students' individual exercise grades are positively correlated with their essay attempts (*p* = .045, r = .405) which are positively associated with their essay grades (*p* = .021, r = .458). It indicated that FC students' enhanced performance in individual exercises was related to their improved performance in essay writing. Further, as students wrote and revised essays out of class on their own, it can be inferred that FC students' improved

language skills had much to do with their learning autonomy. Table 5.15 shows that BC students' individual exercise grades are positively correlated with their quiz grades ($p < .001$, $r = .65$) which are positively associated with their essay grades ($p = .026$, $r = .421$). It can be inferred that BC students' acquired language skills were closely related to the subject content they learned in class. It corroborated the qualitative findings that BC students learned primarily from lectures.

It should be noted that the multiple linear regression models also indicated that students' previous language proficiency was another important predictor of their learning outcome in IEC. Due to differences in class design, students' grades for placement tests correlated with student performance in quite different ways in the two classes. In BC, students' placement test grades correlated with their quiz grades ($p = .016$, $r = .45$), indicating that BC students' language knowledge learning had much to do with their previous language proficiency. In FC, students' placement test grades negatively correlated with their video watching rates ($p = .03$, $r = -.43$), indicating that students with lower language proficiency made more use of videos. However, FC students' placement test grades were positively associated with their peer assessment attempts ($p = .019$, $r = .467$), indicating that students with higher language proficiency were more interactive in group work. This result corroborates with the focus group findings that language proficiency was one of the factors that influenced students' engagement in group work. In brief, the different correlation patterns between student performance in the learning activities in FC and BC indicate that class design may affect student performance in the course.

5.4 Summary

This chapter examined student engagement in the learning activities and their obtained scores for evidence of enhanced engagement and learning outcomes in FC. The quantitative findings were triangulated with the qualitative findings. The results show FC students took great initiative in pre-class activities: they watched videos repeatedly, attaching particular importance to grammar and writing videos; they completed pre-class quizzes a hundred per cent and then joined in peer-led online discussion forums actively. Overall, they were more interactive with their classmates and less teacher-centred than students in BC. In terms of academic outcomes, students in FC performed equally well in the final examination as their peers in BC, and they caught up with peers in BC in essay writing. Throughout the semester, they significantly outperformed their peers in BC in quizzes and individual exercises. However, the quantitative data showed that FC students' group work grades were not as good as those in BC. The time limit for group work tasks could be an obstacle to FC students' performance. The quantitative data further revealed that students' grades in the placement test and the individual exercises were significant predictors of their grades in the final examination. That suggests that students' previous language proficiency and the language skills they acquired over the semester determined their performance in the final examination. The quantitative data, when corroborating the qualitative, indicates that class design can be a factor that influences student engagement in learning.

Chapter 6: Discussion

6.1 Introduction

Chapter Four, reporting the qualitative phase of this project, explored FC students' engagement in the learning activities and answered RQ1; Chapter Five, reporting the quantitative phase, investigated students' enhanced performance and outcomes in FC and answered RQ2. In this chapter, the findings from the previous two chapters will be brought together to illuminate what affected student engagement in the learning activities of FC. A joint display of the qualitative findings and the quantitative findings is presented in Appendix Seven. Then critical reflections on the design and the implementation of flipped classrooms and the formative course evaluations are made to generate evidence for improvement. RQ 3 and RQ 4 are answered in this chapter:

RQ 3. What affects student engagement in learning activities in FC?

RQ 4. What are the reflections and areas for improvement regarding flipped EFL classrooms and formative course evaluation?

RQ 4.1 What are the critical reflections on the design and implementation of flipped EFL classrooms?

RQ 4.2 What are the critical reflections on formative course evaluation of the flipped EFL classrooms?

6.2 Factors Influencing Student Engagement in the Learning Activities in FC

In this section, the qualitative and the quantitative findings are brought together for insights into the factors that influence student engagement in the learning activities in FC. Student engagement is complex and multifaceted; it is fundamentally situational, arising from the interplay of the context and the individual (Wilson et al., 2022). A large body of research on student engagement has indicated that the factors influencing student engagement are many and various. According to Kahu's conceptual framework for student engagement (Kahu, 2013; Kahu & Nelson, 2018), these factors fall primarily into three categories: students' psychosocial characteristics, structural factors, and sociocultural influences. In this section, the key factors influencing student engagement in FC will be discussed by referring to these three categories in Kahu's engagement framework. From the triangulated qualitative and quantitative data, four factors are sorted out as primarily enhancing or hindering FC students' engagement in learning activities. These four factors are the flipping class design, student characteristics, foreign language learning assessment and China's educational system. The first two factors, which fall into the category of psychosocial characteristics, are directly driven from the converged data; the structural factor of assessment policy and practice and the sociocultural factor of China's educational system are inferred based on the findings of the two corroborated datasets. RQ 3 is answered in this section.

6.2.1 Class design

According to Kahu's conceptual framework, psychosocial influences immediately affect student engagement (Kahu, 2013; Kahu & Nelson, 2018). These influences include the university dimension, the student dimension and the interacting relationships between the two. Individual student engagement occurs dynamically at the complex intersection of students' characteristics and backgrounds and the institution and its practices. Teaching practices, students' characteristics and previous experiences, and student relationships with the staff and peers are examples of widely accepted factors in this category. The qualitative and quantitative findings in this study corroborate to confirm that the flipped class design and practices increased FC students' interaction with peers and improved their performance in the learning activities, thus, contributing to enhanced student engagement in learning.

6.2.1.1 Arrangement of Learning Activities

The series of learning activities arranged in FC appeared to help to improve student engagement and learning outcomes. The converged results of this project indicated that students in FC were actively engaged in learning the core subject content via recorded instructional videos. Students of lower language proficiency, in particular, spent more time on the videos. These findings are consistent with the existing flipped practice literature that videos offer students more opportunities for review and mastery (Elmaadaway, 2018; van Alten et al., 2019). Videos benefit students of lower performance more (Lombardini et al., 2018) because the cognitive load can be reduced in processing the learning

materials by enabling students to control the frequency and the pace of video watching (Abeysekera & Dawson, 2015). The quiz results of this study confirm that videos can be more effective than lectures in learning the core subject knowledge, as the learning records in LMS show that students in FC not only significantly outperformed students in BC in the overall quiz score but did better in every quiz throughout the semester. It may well be that, compared with lectures, videos were shorter and controlled by students at their own pace of learning.

The quantitative and qualitative results triangulate that students in FC took quizzes seriously. The qualitative data explain that students used quizzes to measure how they learned the core subject knowledge from the videos. The pre-class instructional videos and quizzes made students conscious of their problems, motivating their involvement in online forum discussions, as the participants reported in focus group discussions. The quantitative data further illuminate that students in FC engaged most in the discussion forum led by themselves and designed for them to share and discuss their problems encountered in videos and quizzes. Recent meta-analysis studies have suggested that quizzes in flipped classrooms positively affect learning outcomes (Hew et al., 2021; van Alten et al., 2019), but pre-class discussions can be more effective than quizzes in enhancing student performance (Strelan et al., 2020). That is probably because discussions provoke interaction and reflection, which are deeper approaches to learning. However, quantitative statistics in this project did not indicate any significant correlation between video watching, quiz performance and forum discussion in FC. Focus group

discussions revealed it might be because many students took part in forum discussions to answer peers' questions instead of asking questions. By enabling students to share and work out their problems together, forum discussion contributed to establishing a learning community in FC. Abundant research on student engagement suggests that learning communities foster learning (Kahu, 2013; Kahu & Nelson, 2018; Wu et al., 2017). Empirical evidence has supported that interacting with peers positively affects almost all aspects of student development, including general education, personal-social development and intellectual skills (Kuh et al., 1997; Lai, 2021; Yu et al., 2018). Therefore, the more students interact with peers in educationally purposeful ways, the more likely they are to engage in their learning (Zhoc et al., 2018). The participants in the focus groups testified that by interacting with peers, they not only enhanced learning but developed a sense of belonging.

It is worth mentioning that the increased peer interaction in the online discussion forum should be primarily attributed to the flipped class design. Students in BC were inclined to be led by the teacher rather than interact with peers when treated with the same learning activities but arranged in a different pattern. Students' long-established habit of learning in a teacher-centred pattern might have further explained why they paid the slightest attention to the forum of exemplary peer work, as peer interaction of this kind was rare in their previous schooling. It should be noted that as all the online forum discussions were optional, that is, the students mobilized online forums on their own initiative; the differences in student engagement in online forums between FC

and BC can be more telling as evidence that the flipping class design enhances peer interaction.

The results of this study support the existing flipped practice literature that pre-class learning activities prepare students for active classroom participation (Al-Zahrani, 2015; Chuang et al., 2018; Munir et al., 2018). Videos and quizzes, coupled with forum discussions in FC, effectively equipped the students with essential subject knowledge for classroom activities. It is particularly beneficial for foreign language learning. According to the cognitive load theory, learning becomes effective when there is a sufficient cognitive capacity for working memory (de Jong, 2009). When students' cognitive load to process language forms is reduced, they have more cognitive capacity to process language content. Learning, hence, becomes more meaningful. This study showed that students in FC not only completed their individual exercises in less time but with significantly better quality. In addition to pre-class preparation, FC students' improved performance in individual exercises can be partly attributed to the face-to-face group discussion in class. Qualitative data indicated that by interacting with peers, students understood better the questions under discussion and improved their answers with what they learned from group discussion. That confirms the extant findings that flipped classes foster deep learning (Kim et al., 2017; Munir et al., 2018), which features meaningful understanding (Richardson, 2005). The sociocultural perspective of learning holds that all higher forms of learning and cognitive development are social in nature (Lantolf, 2009). Empirical evidence from a myriad of research on flipping pedagogy supports that interaction in the flipped classroom not only enhances

student engagement but promotes higher-order thinking processes (Kim et al., 2017) such as reasoning (Betihavas et al., 2016), critical thinking (Munir et al., 2018), and creativity (Al-Zahrani, 2015).

An additional bonus of the flipping approach in this study is that it improves students' understanding of effective language learning. Students in FC appreciated learning by doing or enquiry and held that learning a language in an authentic context is more effective than in lectures. Compared with students in BC, they were less enthusiastic about vocabulary books and mock-test papers. However, it should be admitted that the different class design in the flipped class poses some challenges to adaptation. Qualitative results show that most students in FC took longer to become adjusted.

6.2.1.2 Time Limit for Tasks

The results of this study further indicate that the time limit for tasks in FC can influence student engagement and performance in learning activities. Despite the opportunity for face-to-face interaction in class, students in FC did not perform as well in group work as in BC. That may result from the limited time on task. In FC, there was a 30-minute time limit for group work in each period, which, as some students revealed, was insufficient to complete their tasks. Early work on language learning discovers that learners' cognitive processes are closely linked to the availability of planning time (Kormos & Trebits, 2012). The time limit may also explain students' lower completion rate of individual exercises in FC. However, FC students' grades in individual exercises were still significantly higher. That is probably because group work involves more higher-

order thinking skills, thus is more complicated than exercises focusing on language skills. Research on educational psychology indicates that sufficient time on task is conducive to in-depth information processing and higher-order thinking processes (de Jong, 2009). By contrast, group work in BC was done after class, so sufficient time on tasks was likely to contribute to BC's better performance in group work considering the two classes were working on the same project.

On the other hand, this study found that some students in FC did not make full use of their class time for group work, indicating that students' inferior performance might not just result from limited time for the task but from their ineffective use of time. The existing research on flipping pedagogy suggests that group work is the most difficult to handle of all the learning activities in a flipped class (Betihavas et al., 2016; Chuang et al., 2018; Elmaadaway, 2018; Lombardini et al., 2018; Munir et al., 2018) despite its value to enhance learning through interaction and collaboration (Akçayıra & Akçayır, 2018; Lin & Hwang, 2018; van Alten et al., 2019). The qualitative findings of this study indicate that group work contributed most to the development of higher-order cognitive skills, but its value was not legitimately received by students in both classes, regardless of the different class designs. The quantitative results added that group work appeared irrelevant to the course. Neither students' attempts nor grades for group work were correlated with the other learning activities and their final examination scores for IEC. That indicates explanations other than class design and teaching practice need to be brought in to understand students' engagement in group work.

6.2.2 Student Characteristics

Kahu argues that student engagement is rarely influenced by institutional factors or student factors separately, but most probably by the interplay of the two dimensions, and that student engagement occurs when the two dimensions align (Kahu & Nelson, 2018). The results of this study indicated that FC students' engagement in the learning activities depended much upon the compatibility between what a task expected students to do and a complicated combination of skills, attitudes and values on the part of students. This finding is consistent with existing literature that students' reactions to learning are associated with the nature of the task, students' skills, personality, interests and past experiences (Kahu, 2013).

6.2.2.1 Student Skills

Consistent with the findings by Betihavas et al. (2016), Lombardini et al. (2018) and Munir et al. (2018), this study found student inertia or resistance to group work. Similar to the findings of previous studies on peer collaboration, students reported they were discouraged by their unpleasant experiences when working with peers, such as unequal distribution of responsibility within a group, ineffective interactions with peers (Elmaadaway, 2018) or unaccountable group members (van Alten et al., 2019). The tools students used for doing the tasks, LMS, for instance, were sometimes blamed by them for their underperformance. In addition, strikingly similar to Chen's findings from a collaborative online writing project (Chen et al., 2021), students in this project interacted in their groups more in a simple cooperative pattern where group

members worked separately and then pieced their work together than in a collaborative manner with sustained peer interactions and shared responsibility.

The qualitative data provided evidence that students lacked skills for effective communication and task management because they lacked collaborative learning experiences, which was supported by a body of educational studies on Chinese students (Barratt-Pugh et al., 2018; Chen et al., 2021). In addition, the qualitative and quantitative results in this study converged to show that students' English language proficiency can affect their performance and engagement in learning. It should be noted that the participants in this study were English majors, so English language proficiency for them was not just a communicative skill but a subject skill. All these findings pulled together, this project suggests that students' skills in the subject field, communication and task management affect their engagement in collaboration. However, it is interesting to note, as the qualitative data show, that students' interactive skills developed as the semester went on, but they did not do group work with much difference. Hence, lack of skills explains just part of students' superficial engagement in group work.

6.2.2.2 Learning Attitude

Previous studies suggest that students' self-efficacy and the value they attach to an activity are two significant determinants of learning motivation, affecting students' choices of learning activities and their engagement in the chosen activities (Lai, 2021). The review research by Betihavas et al. (2016) concluded that students in flipped learning did not perceive the value of interaction.

Similarly, this study found that students did not buy into the value of working with peers, which can be the underlying cause of their inactivity in group work. It is worth noting that despite the difference in class design, students in both FC and BC in this project thought about and dealt with group work in much the same way. Many believed they did not learn much of use from working in groups, so it was not time-effective. Students' reaction to group work brings about the issue of compatibility between what a learning activity aims to achieve and what students expect to gain from the activity, which is what Kahu and Nelson (2018) argue to be the alignment between the institutional dimension and the student dimension.

The results of this project show that while students in FC doubted the value of group work, many were motivated to join in the student-led online discussion forum and received classroom group discussion quite positively. Qualitative data suggested that one possible explanation was that students were extrinsically motivated by obtaining higher grades in learning tasks. By joining online forum discussions and classroom group discussions, students in FC could get developmental feedback to improve their performance in quizzes and individual exercises for higher grades. According to Richardson (2005), students who aim to obtain high grades tend to use a strategic approach to learning, switching between the surface approach for assessment purposes and the deep approach to meaningful understanding. While students were eager to get higher grades, group work results were beyond individual students' control as their efforts were likely to be compromised by their peers. When unsure of the returns their efforts could bring, it was natural that students used

a surface approach to learning. Most of them just completed their part of the task for assessment purposes.

A further look into the learning objectives of these three interactive activities may add more insights into the underlying attitude that shaped students' motivation for engagement: online forum discussions helped students solve their problems with subject content; classroom group discussions enhanced students' mastery of language skills; group work developed students' collaborative and higher-order thinking skills. Students in FC were quite motivated to enhance their language knowledge and skills, as evidenced by the statistical results of their significantly invested time and effort in grammar and writing videos and student-led online forum discussions. The misalignment between learning objectives and student expectations is evident here: group work aims to develop students' interactive and higher-order thinking skills, but students were keen to enhance their language knowledge and skills. Students' obsession with subject knowledge and skills is not unique to students in FC but common to Chinese students in general, which has much to do with students' previous educational experiences that are mediated by assessment policy and practices and shaped by China's educational system. These factors belong to structural and socio-cultural influences and will be discussed in the following sections.

Here we will look into the impact of these educational practices and contexts on students as they shape students' belief in effective learning. In this study, students' belief in developing language skills by doing repeated paper-and-pencil exercises may well result from their confidence in the effectiveness of

doing so. After all, to most of them, this is how they managed to compete with most of their peers and finally go to university. Many participants suggested they should *be tested regularly* and *ranked accordingly to know how they learned*. Many students in this study acknowledged they gauged learning with test scores, so much so that they felt at a loss as to what they had learned without being tested and lost their goals for endeavours. It should be noted that what the students meant by “being tested” is to do test papers in the form of standardized tests; quizzes, to them, are too informal to be tests.

A wealth of research shows that excessive test-taking experiences have a cumulative negative impact on students, resulting in decreased motivation to give genuine effort, increased use of inappropriate learning strategies, and, to some, growing disillusionment about tests (Nichols & Dawson, 2012). Early work also concludes that summative assessments, in particular, compromise students’ beliefs about the value of learning, leading to performance comparison among students at the expense of deeper learning (Hughes, 2011; Nichols & Dawson, 2012). That explained students’ enthusiasm about comparing grades and points with peers and their neglect of feedback conducive to their performance in the learning activities. Nevertheless, the results of this study show that students in FC were more engaged in learning activities and less performance-oriented than BC. One explanation can be that there was more interaction between students in FC than in BC, which promoted active and deeper learning.

6.2.2.3 Student Values

The influence of face-issue values on peer interaction between students is discussed in this section. The face-issue mentality has its root in China's social-cultural context. It is discussed in this section because the face issues affect individual students differently, shaping how they engage in collaboration to various extents. A body of educational research concerning Chinese students concludes that Chinese students tend to avoid making critical comments to save face in peer feedback (Chen et al., 2021; Yu et al., 2018). This study added empirical evidence to the existing literature.

Peer assessment was affected most by students' face-issue mentality in this project. When peer assessment was not anonymous, students were more likely to assess peers' work indiscriminately to avoid offending peers. It should be noted that many students did apply the assessment criteria provided for them, but on most occasions, much more leniently, so there were differences, but tiny ones, in grades for work of different quality. Nicol et al. (2013) explained that students' application of assessment criteria was affected by their experiences, resulting in their using two sets of criteria simultaneously: the explicit set provided by the teacher to calibrate their practice and the implicit set deriving from their own experiences. Students in this project were aware that their practices undermined the role of assessment, but they believed what they did was natural and appropriate.

As such, many doubted the value of peer assessment. There is no denying that some students used face-issue as an excuse to circumvent the assessment

task. LMS records showed that some students assessed peers' work with the same grade throughout the term, disregarding the quality of the work and the given criteria. All these findings indicated that the effects of peer feedback and peer assessment were undermined by students' conceptions of social conventions. However, what should be noted is that though students in FC and BC reacted to peer assessment in much the same way, there was statistically significant inertia in BC with between-group peer assessment, indicating a difference possibly resulting from class design.

6.2.3 Assessments in Foreign Language Teaching

According to Kahu's conceptual framework of engagement (Kahu, 2013), the structural factors, though not exerting a direct influence, have a recognized impact on student engagement by interplaying with the psychosocial influences. The findings of this project indicate that the current assessment practice in the discipline of foreign language teaching can be a critical factor that shapes FC students' engagement in learning activities. It explains students' obsessive effort to develop subject knowledge and skills by doing paper-and-pencil exercises and their lack of interest in collaborating with peers for group tasks.

Existing research has proved that assessment practices influence student understandings of learning, which produce a substantial impact on what students value to learn and how they approach learning (Fletcher et al., 2011; Nichols & Dawson, 2012; Zhang & Bournot-Trites, 2021). Assessment in HE serves to provide information about student learning, teaching quality and program accountability. Educationists share that to ensure students graduate

with critical academic and professional skills, assessments should aim to evaluate student attainment of higher-order skills, such as critical thinking, creativity, collaboration and leadership (Barratt-Pugh et al., 2018; Fletcher et al., 2011). In reality, the assessment practice in the discipline of foreign language teaching is not quite in concordance with these high-end purposes. These critical skills for life success are usually assessed formatively in everyday classroom-based learning tasks as low stakes. High-stakes summative assessments like final examinations or national/international standardized tests focus more on the test-takers' language proficiency: their knowledge and skills of the target language. Group work tasks appear unrelated to high-stake tests in both form and content. Most high-stakes tests contain, in a large proportion, standardized items of multiple choices and blank filling. Essay writing in these tests may be the most related to group work tasks. However, essay writing in high-stakes tests accounts for a limited proportion and focuses more on language form, where deficiency in higher-order thinking skills is often offset by rich vocabulary, precise diction and accurate grammar. Students in general, and Chinese students in particular, deem high-stakes assessments as having consequential implications for their future. Therefore, students in FC insistently held that time and effort for learning should be devoted to what was useful for those critical tests. The emphasis on language skills in high-stakes assessments shaped, in a way, FC students' conceptions of learning activities, which, in turn, influenced the way they approached these activities.

6.2.4 China's Educational System

Examining student engagement within the broad social, political and cultural context offers crucial insight into why students become engaged or disengaged in learning (Kahu, 2013; Kahu & Nelson, 2018). It will be impossible to fully understand the participants' lack of skills and confidence in interactive activities and their obsession with higher grades in this study without looking into China's wider socio-cultural context where its educational system is embedded.

6.2.4.1 Strict Testing System

Abundant evidence in the existing literature supports that much of the learning and teaching practices in China are shaped by its high-stakes testing systems (Barratt-Pugh et al., 2018; Zhang & Bournot-Trites, 2021) that pervade current China's entire educational system. In China, stringent criterion-referenced tests are used to select students at every stage of schooling, from primary school to doctorate programmes. Under such an assessment context, to win better resources that open up more chances for further education, a student has to rank high, which means s/he has to do better than the others. Through rounds of selection by high-stakes tests, no more than 40% of the student population has the opportunity for higher education (Liu & Li, 2020). In a sense, these tests that aim to select higher-achieving students for further education are critical to Chinese students as they can have irreversible consequences on their life trajectories. As a result, the importance of test grades and grade ranking is deeply rooted in Chinese students' minds at the very beginning of schooling.

Educational assessment environments influence approaches to teaching and learning. Nichols and Dawson (2012) find that high-stakes testing systems change what is taught in schools: in most classrooms, content that will be tested is emphasized; content that will not be eliminated. Meanwhile, teachers spend large amounts of time preparing students for tests and emphasize their importance to students. Previous research on China's educational system and qualitative findings in this study prove that some students do not have listening and speaking training at all in high school English classes because tests of these two subsets of skills are not mandatory in University Entrance Examination in their regions (Zhang & Bournot-Trites, 2021). It should be noted that in China, this is not a practice by individual teachers or in specific classrooms but instead arises from shared regional or provincial practices. Existing research finds that less competitive schools are more likely to be test-centric (Zhang & Bournot-Trites, 2021). Some students in this study further revealed that to prepare for the University Entrance Examination, they did endless mock test papers and were only taught what was in these papers in high school English classes. That brings some clues to students' lack of interactive experiences and skills. Interactions, especially those between peers, rarely occur in high school English classes because they are of limited use to those high-stakes tests. The perceived criticality of the tests by all the stakeholders underlies the examination-centric teaching and learning approach that screens all irrelevant elements to tests. To quote one of the participants in the focus group, "*We couldn't afford that time [to develop interactive skills] in class.*"

That explains, to some extent, students' preference to be spoon-fed in learning. It is about the time cost. It is more time-efficient to get straightforward instructions and answers from the teacher so that students have more time to drill and achieve automaticity for tests. Unlike Nichols's findings based on the testing system in the United States, where teachers' preoccupation with tests resulted in students' missed opportunities for authentic learning experiences (Nicol et al., 2013); in China, such preoccupations are driven more by students themselves and parental expectations.

6.2.4.2 Limited Educational Resources

Underlying the fierce competition in education is the limited educational resources in the country. China now has a vast student population of about 0.3 billion, while only about four per cent of its Gross Domestic Product (GDP) goes to the education sector each year (Liu & Li, 2020). There is no exaggeration to say that in China, behind every student's competition at school is the contest between family background and resources. It leads up to the issue of educational equity, as students from families of higher social and economic status are sure to have more access to better educational resources, driving under-privileged students into a disadvantageous position in this competition. In the face of an ever-increasing student body and limited educational resources, the current testing system, as a means of selecting students, may stay in effect for a long time. That is because, to the government, it is the most cost-efficient way to allocate its limited educational resources to the most talented students; meanwhile, the only way to orchestrate the available resources with equity (Barratt-Pugh et al., 2018). Without this strict testing system, children from

families of low economic status can be even more disadvantaged in this competition. However, as long as the current educational system continues to exist, students' approach to learning will be difficult to change, as their approach to learning is continually reinforced by the same system and, in addition, has deep cultural roots.

6.2.5 Brief Summary

Abundant empirical evidence in educational research has concluded that students use different approaches to learning. The results of this study indicate that the flipping design does enhance student engagement in learning activities in the flipped EFL classroom under investigation. However, the extent to which students engage in the learning activities is very much influenced by their conceptions of the activities and their skills, which are shaped by their previous educational experiences, institutional factors and the socio-cultural context. It should be noted that the factors discussed in this section are by no means inclusive but the primary ones that influence student engagement in FC based on the findings of this study.

Another important factor that moderates students' learning experience in IEC should be mentioned here. The use of learning technology, LMS, in this study, played a role far beyond an assistive tool of content delivery; it, as the published studies support, enhances student interaction with peers, teachers, and learning materials (Chang & Hung, 2019; Macfadyen & Dawson, 2012). It assisted, in particular, in establishing a learners' community via the online forum, thus promoting learner-centred active learning (Pardo, 2014). Essential

as LMS is to IEC, this study indicates it was not a primary factor influencing student engagement in the FC. Recent studies also support that learning technology moderates student engagement in flipped classrooms but has no direct impact on it (Lai et al., 2021). In this study, the two predictors of learning outcomes, the placement test and the individual exercises, have little connection with technology use. Besides, the qualitative results of this study indicate that, though students lacked previous experience and skills in using LMS, their affective reactions towards LMS were actually more situational. Student dissatisfaction with LMS relented when they performed the tasks well and when they became familiar with doing tasks online. Nevertheless, the results of this study do support the existing findings that ease of use and perceived usefulness are important parameters of students' technology acceptance (Chang & Hung, 2019; Papamitsiou & Economides, 2014). Henderson et al. (2015) argue that, in addition to the individual agency, how a course is structured, its content is created and delivered, and assessments are designed all frame students' use and perception of digital technologies. That may help explain to some extent the differences in affective reaction to LMS between FC and BC, where students in FC demonstrated better acceptance than BC.

6.3 Reflections and Areas for Improvement

The primary purpose of course evaluation is to strategize evaluation findings and develop action plans to enhance teaching and learning practices. Evaluations underpinned by pragmatic epistemology advocate reflection on practices to solve problems and enhance the credibility of evaluation findings.

Based on the results and discussions in the previous sections and chapters, this section presents critical considerations for better designing and implementing flipped EFL classrooms and formative course evaluation. RQ 4 and its sub-questions are answered in this section.

6.3.1 Designing and Implementing Flipped EFL Classrooms

Kirkpatrick and Kirkpatrick (2006) suggest two prerequisites for changes to take place: to decide on the change to be made and to get the acceptance of those involved. That implies that having students accept the flipping approach and getting involved in the flipped classroom is essential for enhancing student engagement in flipped learning. Change management experts agree that changes in habits, practices and behaviours are brought about when conditions are created to allow the participants to think and feel positive about the change (Macfadyen & Dawson, 2012). To win student support for flipped learning, the design and implementation of learning activities need careful consideration and planning. Based on the findings of this study, the following should be considered for improvement purposes.

6.3.1.1 The Design and Implementation of Group Work

The empirical evidence of this study shows that the biggest challenge of the flipped IEC lies in the design and implementation of group work. Collaborative learning is an essential component of the flipping pedagogy (Strelan et al., 2020). While its contribution to effective learning is widely acknowledged (Akçayıra & Akçayır, 2018), the design and implementation of collaborative tasks have always been a challenge (O'Flaherty & Phillips, 2015). Literature

indicates that group work breeds the most student dissatisfaction among all the learning activities in a flipped classroom (Betihavas et al., 2016; Lombardini et al., 2018; Munir et al., 2018). In this study, students' lack of cooperative learning experiences, interactive skills and confidence in collaborative learning hampered their collaboration in group work. As such, scaffolding student collaboration becomes essential to enhance student engagement. Munir et al. (2018) suggest that procedures of collaborative activities should be adequately explained to and learned by students so that they know how to go about the task. However, this study further indicates that student engagement in collaboration was moderated by their briefs in collaborative learning in addition to know-how. Richardson (2005, p. 674) argues that "educational interventions will not be effective in changing students' approaches to studying unless they […] bring about changes in students' perceptions." That implies helping students understand what is important to learning, and why they are doing that can be as critical as teaching them how to do it. Many scholars researching a flipping pedagogy suggest this should be done at the beginning of the semester upon implementing the flipping approach (Al-Zahrani, 2015; Chen Hsieh et al., 2017).

Lai (2021) supports with statistical evidence that the interplay between students' perceived value of group tasks and task difficulty is critical to student interaction and engagement in group work. Bishop and Verleger (2013) suggest that the flipped class design should be student-centred. The results of this study suggest that student characteristics should be fully considered in designing group work and that time for tasks, task difficulty, and workload need adjusting

accordingly. Group work tasks in EFL classrooms should be fit for enhancing students' language skills and develop their higher-order thinking skills. Meanwhile, tasks should match students' competence to be doable within the time limit in class. This study further reveals that students preferred to work in smaller groups so that every member could get adequate practice and coordination remained manageable within the group. The recent large-scale meta-analysis indicates that smaller groups do have a stronger effect size on student performance though no empirical evidence shows that group size affects learning outcomes (Strelan et al., 2020).

Another critical issue with group work is that, although a body of studies find cooperation between students in flipped class enhances learning outcomes (Akçayıra & Akçayır, 2018; Munir et al., 2018), recent meta-analysis studies indicate that group work is not a significant moderator of learning performance and outcomes (van Alten et al., 2019). Strelan et al. (2020) add that it does not matter much whether students engage in group work or individual tasks as long as the activities are student centred. Though the qualitative data in this study showed that students developed most of their higher-order thinking skills via group work, the quantitative results did not find any correlation between group work and the other learning activities in terms of either student engagement or learning outcomes. It implies that more fine-grained empirical studies are needed to explore and understand how group work enhances learning to better inform its design and implementation.

6.3.1.2 Teacher Feedback

van Alten et al. (2019) conclude from their meta-analysis of the flipping pedagogy that teacher feedback is an essential component in an active flipped classroom to guide and facilitate the learning process. This study confirms that teacher feedback is particularly important in flipped learning to ensure sustained active engagement and learning quality. In well-designed flipped classrooms, learning activities are related one after another, so the learning outcomes of prior activities can affect student engagement in the subsequent activities. Timely feedback prepares students for better participation in learning activities.

However, this study indicates that while students claimed teacher feedback was important to them, the messages sent by the teacher were generally not well received by students. It was mainly because, according to students in FC, feedback was not delivered promptly. The time lag between submission and feedback usually took days when assignments were completed outside the class. Students said *their impression of the assignment went vague*, and *enthusiasm about the task dimmed* when they were waiting for feedback. Some students forgot about the assignment and therefore did not check teacher feedback at all. This study supports that the flipping approach allows more timely feedback, as most learning tasks are done in class, which improves student performance on tasks and enhances learning outcomes. Hence, the flipped design is, to some extent, beneficial for sustained active student engagement.

Nevertheless, making teacher feedback effective for learning remained a challenge. A performance gap existed between what students actually did and the reference level. Literature suggests that more informational comments than grades should be included in feedback to maximise the effect of teacher feedback; it is crucial for performance-oriented students because comments as effective formative feedback motivate learning (Hughes, 2011; Nichols & Dawson, 2012) and link closely to improved student learning (Li & De Luca, 2012). The lesson learned from this project is that this change from grades to comments should take place gradually, as students highly instrumentally motivated by grades would likely be demotivated without external regulators. This study further indicates that when giving comments as feedback, it is not enough for improvement to take place by just telling students areas for improvement (Li & De Luca, 2012; Nicol et al., 2013), as teachers' comments are often mediated by students' interpretation (Boud & Molloy, 2013) and competence to work on the comments accordingly. Boud and Molloy (2013) argue that messages on performance gaps are feedback only when the messages are acted on. This study implies that for feedback to take effect, teachers need to show students how to do it, not just what to do. Unfortunately, the large body of principles of good practices of teacher feedback in the existing literature rarely covers this regard.

6.3.1.3 Classroom Management

This study indicates that effective classroom management can be vital to improving students' learning experiences in flipped classrooms. As a flipped

class is composed of learning activities arranged in sequence, these various learning activities pose extra demand on classroom management for teachers.

Hung (2017) argues that teachers must understand student needs to facilitate student engagement in learning activities in a flipped classroom. The results of this study indicate that students need reminders of learning tasks, incentives to stay on track, assistance with tool use, and scaffoldings for interaction. That implies that teachers need to play different roles in flipped classrooms. A body of flipped practice literature suggests that in flipped classrooms, teachers should change their role from content presenters to learning coaches (Akçayıra & Akçayır, 2018), guides, facilitators (Hung, 2017), assistants (Elmaadaway, 2018), resources (Lee & Wallace, 2018), instruction designer and media developer (Chen Hsieh et al., 2017), and so on. The question of when to shift between these multiple roles to meet different student needs then becomes important. One of the issues that makes effective classroom management challenging for teachers is that most students do not explicitly express their need for support; many are unaware of their problems. When the flipping pedagogy empowers students to exert more control over learning, it entrusts them with more responsibilities for themselves. Unfortunately, abundant evidence shows undergraduate students are far from capable self-directing learners (Akçayıra & Akçayır, 2018; Betihavas et al., 2016; Kirschner & van Merriënboer, 2013).

It implies that to understand student needs, teachers should constantly monitor student performance to be aware of students' problems at the earliest time and be responsive to their needs. That can be particularly important for classrooms

with students without flipped learning experiences, as they are more likely to encounter problems beyond their expectations. Reeve (2013) validates that teachers' initial support in a course can motivate students' agency later on during the course, which in turn enhances learning outcomes. Students' initial experiences in flipped classrooms can affect their perceptions and influence their engagement in flipped learning in the future. One of the strengths of the flipping approach is that it increases opportunities for teacher-student interactions and enables teachers to observe student performance closely and cater to student needs in time, which is conducive to effective classroom management.

Here due credit should also be given to LMS, allowing teachers to observe student involvement with learning materials and activities closely and adjust accordingly quickly. LMS is particularly helpful in identifying students who lag behind. The implication here is that LMS, as a teaching tool, should be integrated with pedagogical visions to catalyse change in student experience and learning outcomes. Teacher training to enhance teachers' technology use skills for pedagogical purposes is critical in this regard. Regrettably, current teacher training regarding technology use in teaching practice has been mostly technical rather than pedagogical.

6.3.2 Formative Course Evaluation

Formative course evaluation aims to improve teaching and learning practices by taking evidence-based remedial actions (Holland, 2019; Sozer et al., 2019). In this sense, it can be critical to the success of flipped classrooms.

6.3.2.1 Data Collection at Multiple Time Points

This study shows that formative course evaluation at multiple time points offers a holistic understanding of student engagement in FC, which contributes to an educated understanding of the pedagogical effects of the flipping approach. The evaluation taken in the earlier sessions of the semester serves to set the baseline of student engagement so that future changes can be captured. Besides, it allows teachers to be informed of student learning experiences as early as possible and adjust teaching timely to cater for student needs. As such, it offers the greatest scope for improvement in both learning and teaching practices. That can be particularly important to students without flipped learning experience, as they may need more support to benefit from the new approach. What is worth mentioning is that different from evaluations taken at the end of a semester that can hardly benefit the students who provide data, early evaluation of a course can benefit the students currently taking the course as well as the following cohorts (Holland, 2019; Sozer et al., 2019). The results of the end-of-term evaluation, when examined together with the findings of the preceding evaluations, allow teachers to understand students' learning outcomes in relation to their performance throughout the learning process, thus increasing the credibility of evaluative findings.

6.3.2.2 Convergence of Qualitative Data and Quantitative Data

The multiple time-point data collection offered a channel to capture data for both the learning process and learning outcomes; the mixed-methods design produced rich data. In this project, the qualitative data situated and explained

the quantitative findings, while the quantitative findings produced a succinct summary of the research results. The two datasets triangulated and complemented each other, enhancing the credibility of evaluation findings.

Recent research suggests that course evaluation for improvement purposes should be informed by student learning experience and adopt a student-centric approach (Steyn et al., 2019). The qualitative approach to data collection can trigger more elaborative responses about what is going on in a class. In this study, the qualitative approach managed to capture the unique nature of individual experiences in flipped learning embedded within the wider socio-cultural context and depicted the complex array of factors influencing student engagement in the learning process. The focus group discussions allowed student voices to be heard from their perspectives. Besides, the suggestions for improvement from students were closely related to their experiences in such a level of detail that allowed the teacher to make well-informed adjustments to improve the course design and teaching practices. Notwithstanding these strengths of qualitative findings, they were too intricate for us to understand the overall effect and the general trend. The quantitative findings offered the succinct answer to the effectiveness of the learning activities.

An added value of the qualitative phase in this study was that it maximized the process use of evaluative practices. Literature on course evaluation suggests that interactions with peers in focus groups provoke students' self-reflection (Holland, 2019) and, therefore, improve their perceptions of learning (Sozer et al., 2019). Scholars in course evaluation studies add that for course evaluation to take effect, it must be informed and owned jointly by the teachers and the

students taking the course (Fisher & Miller, 2008; Freeman & Dobbins, 2013). The qualitative approach in this study served these goals by allowing students to exchange ideas and make suggestions for their flipped classroom design and implementation.

However, this study shows that collecting data via focus group discussion is a double-edged sword. On the one hand, the open data from the participants' voices allow researchers to capture what they may not have noticed or expected. On the other hand, as discussions are under the control of the participants, there is the possibility that student discussions may not be sufficiently relevant to the issues under research. The implication is that discussion prompts should be fully prepared beforehand and adjusted timely during discussions. Both can be challenging to novice researchers.

6.3.2.3 Data Collected from LMS

The quantitative data for this study were collected from LMS. LMS plays a dual role in this study: mediating online learning activities for IEC and assisting data collection for this study. The strengths of LMS for data collection, this study shows, is that it tracks student access to learning activities and resources in different learning scenarios and allows collection of a variety of data via observation without intervening students. The detailed collection of the events in the learning experience offers a better background for data analysis and prediction. The caveat here is that LMS is a powerful tool to collect data, but how the recorded data are aggregated, analysed and reported is still at the discretion of investigators, which may well affect the research findings.

This study suggests that the selection of indicators of learning outcomes may affect the research findings. In this study, students' placement test scores and individual exercise scores are significant predictors of their learning outcomes in IEC. This finding can be a result that the final examination scores were used as the indicator of student learning outcomes. Placement tests, individual exercises and the final examinations in IEC all focused on English language knowledge and skills. As a matter of fact, when GPA, which counts in students' daily performance, was used as the indicator of learning outcomes instead, students' essay grades also entered the regression models as a significant predictor of learning outcomes in this study. Meanwhile, the lack of correlation between group work performance and the overall learning outcome can also result from the final examination being used as the indicator of learning outcomes. The final examination for IEC scarcely examined students' higher-order thinking skills but focused on language skills. That means the final examination did not appropriately measure what students learned and trained in group work tasks. Therefore, it implies that indicators should be carefully chosen so that the effectiveness of the learning activities can be appropriately measured.

The quantitative results show FC significantly outperformed BC in quizzes and individual exercises, but there was no significant difference in the final examination between the two classes. Such findings are no surprise and agree with the conclusions from the recent meta-analysis of the effects of flipping pedagogy. The flipping effect is the strongest when student performance is measured immediately after a flipped session but much weaker when the

measure is the final course mark or examination (Strelan et al., 2020). That is because students perform better when tested immediately on materials they have learned, but learning effects fade over time (Strelan et al., 2020).

In addition, literature on both the flipping pedagogy and the course evaluation suggests the intervention duration can be a moderator of the research results as students may need a more extended period to get used to a new approach, like flipped learning, to work effectively (Kirkpatrick & Kirkpatrick, 2006; van Alten et al., 2019). That implies there may be a difference in the overall learning outcomes when data covering a longer intervention duration are used. Hence, researchers' decisions as to what data are to be collected from LMS, over what time span and how data are to be analysed can affect the findings of a study. To bring the full potential of LMS for data analysis, how data are mined needs to be carefully thought out.

Collecting data from LMS is not devoid of disadvantages. The one most related to this study is that LMS tracks online activities only, so quantitative data for student participation in learning activities in the face-to-face context, like classroom discussion and face-to-face group work interaction, were not collected and triangulated with the qualitative data.

6.4 Summary

This chapter discusses the primary factors that affected students' engagement in FC and some critical reflections on designing and implementing the flipping pedagogy and formative course evaluation. The factors of student engagement are examined from three levels. The psychosocial level explains how the flipped

class design interplayed with students' skills in the subject area, interaction and task management, their attitude to effective language learning and their face-issue mentality, directly influencing their engagement in the flipped IEC. The structural level elaborates on the assessment practices in the discipline of foreign language teaching that shaped student attitudes to and practices in their learning of English. The sociocultural level probes into China's educational system in terms of the current testing system and educational resources that underpinned students' conceptions of and behaviour in learning. This chapter then reflects on the design and implementation of the flipping pedagogy and formative course evaluation. It suggests that the design and implementation of group work, timely teacher feedback and classroom management are critical issues in enhancing teaching and learning practices in flipped EFL classrooms. In terms of formative course evaluation, this study indicates that multiple time-points data collection, focus group discussions, and log data from LMS can produce rich data that allow researchers to perform robust investigations if issues of data collection and analysis discussed in section 6.3.2 can be carefully addressed.

Chapter 7: Conclusion

7.1 Introduction

This conclusion chapter begins with a brief overview of this research to answer the research questions in full based on the foregoing analysis and argument. It then briefly summarizes the contributions and limitations of this project.

7.2 Brief Overview of the Research

This convergent mixed-methods evaluative study explored and investigated to what extent the learning activities in a flipped EFL classroom enhanced student engagement in the learning process and the learning outcomes. The Kirkpatrick Model was employed to measure student engagement in learning activities from the affective, cognitive and behavioural perspectives and the subsequent learning outcomes. Qualitative findings and quantitative findings of this study converge to support the existing flipped practice literature that the flipping approach enhances student engagement in learning (Akçayıra & Akçayır, 2018; Kim et al., 2017; Lin & Mubarok, 2021; Turan & Akdag-Cimen, 2020). On the whole, students in the flipped class were more integratively motivated and less performance-oriented. They demonstrated impressive learner autonomy and were more interactive and less teacher-centred. Participants in this study were enthusiastic about enhancing their language knowledge and skills. Their reactions to interactive activities were more divergent than those towards individual tasks. Consistent with the findings of previous studies on flipped classrooms, many students in FC did not perceive the value of group work (Betihavas et al., 2016) even though more higher-order thinking skills were

reported from group work than the other learning activities. Some students in the flipped class were less organized and took a longer time to become adapted to the course. That said, the results of this study indicate differences in student engagement in some learning activities resulting more from individual learner differences than from class design.

In terms of learning outcomes, there was no indication of significant differences in final examination grades between FC and BC, which is in line with some of the existing findings on flipped learning (Betihavas et al., 2016; Jensen et al., 2015; Lombardini et al., 2018; McNally et al., 2017; van Alten et al., 2019). However, students in the flipped class performed significantly better in quizzes and individual exercises throughout the semester, and they progressed and caught up with BC in essay writing though they performed less well in group work. This study shows that students' placement test scores and individual exercise grades are two significant predictors of their final examination grades, indicating that students' previous language proficiency and their language skills developed in the course are important indicators of their overall learning outcomes in flipped EFL classrooms. Consistent with previous studies (O'Flaherty & Phillips, 2015; van Alten et al., 2019), collaborative group work, which scholars and educators highly embrace, turns to be not correlated to the learning outcomes. However, these results need to be interpreted with caution as decisions on data collection and analysis can, to some extent, affect research findings.

By referring to Kahu's conceptual framework of student engagement (Kahu, 2013; Kahu & Nelson, 2018), this study further reveals that flipping class design

can be an important contributor to student engagement and performance in learning activities. What should be noted is that it is the well-planned and organized multiple learning activities both within and outside class, rather than one particular activity, that engages students in active learning in the flipped class and enhances their learning performance and outcomes. Further, the class design does not take effect in vacuum conditions. In this study, it interplayed with students' skills and attitudes toward language learning, which were mediated by the assessment system in the discipline of foreign language teaching and shaped by China's educational system. That said, this study suggests, in concordance with the existing flipped practice literature, to further enhance the effects of the flipping approach from the pedagogical perspective, the design and implementation of group work still have much room for improvement (Chuang et al., 2018; Elmaadaway, 2018; Lombardini et al., 2018; Munir et al., 2018). Besides, teacher support and effective classroom management are crucial elements to enhancing student engagement in the flipped class (Lai et al., 2021; van Alten et al., 2019).

From the perspective of course evaluation, this study supports the existing literature that the mixed-methods design produces a richer context for data analysis and prediction, which contributes to a fuller understanding of the evaluation problems and increased credibility of the evaluation findings (Cohen et al., 2011; Mertens & Hesse-Biber, 2013). By collecting data at multiple time points, timely adjustments to the course are made possible (Fisher & Miller, 2008; Holland, 2019; Nguyen & Foster, 2018), and changes in student learning experiences are captured (Fisher & Miller, 2008). This study supports that, in

particular, LMS can be used as an effective tool for sustained data collection throughout a study. The caveat is that data collection, analysis, prediction and reporting from LMS have so far lacked standardization (Zhang et al., 2018). Hence, evaluative findings must be understood within the evaluative context and interpreted with caution (Papamitsiou & Economides, 2014).

7.3 Contributions and Limitations

This study fills a knowledge gap of what is happening during flipped learning. Focusing on student engagement in the learning activities, it revealed how different learning activities in a flipped EFL classroom contributed to the learning outcomes through specific learning mechanisms. Student voices of their learning experiences in the flipped classroom were brought into the study, improving our understanding of why and how some learning activities worked, and some did not. This increased knowledge of the learning process and its mechanism offered critical pedagogical insights into the design, implementation and evaluation of the flipping innovation.

This study innovatively integrated Kahu's conceptual framework of student engagement and the Kirkpatrick Model. By embedding the Kirkpatrick Model into Kahu's conceptual framework of student engagement, this study made up for the weakness of Kahu's conceptual framework in measuring the process of student engagement. In doing so, it brought empirical data for Kahu's framework, illustrating with empirical evidence how students' psychosocial factors interplay with institutional structural factors and the broader sociocultural factors to affect their engagement in learning. Further, it introduced the

qualitative approach to the Kirkpatrick Model to measure student engagement in the learning activities, which situated the data interpretation and prediction and enhanced the credibility of the evaluation findings. Doing so demonstrated that qualitative and quantitative data could work well within the Kirkpatrick Model to measure the effectiveness of the flipping pedagogy, hence, verifying the Model's applicability to e-learning (Kirkpatrick & Kirkpatrick, 2006).

Last but not least, as the unique strength of an evaluative study, the student participants benefited from their participation in the evaluation process. By discussing and reflecting on their practices in learning, the student participants had a better idea of their learning performance, developed more explicit goals of endeavours, deepened their understanding of effective learning and developed more ownership of the course under evaluation. All these are signs of autonomous learners.

One of the limitations of this study is that not all the qualitative data on the learning activities had corresponding quantitative data to triangulate them. The quantitative data of this project were collected from LMS, which tracked and recorded online activities only. As a result, face-to-face classroom activities of teacher feedback and group discussion had only qualitative data collected from focus group discussions and participant observation. As such, research design can be an area of improvement for future research.

A further limitation of this study can be the chosen indicator of learning outcomes. To avoid repeated measures of the learning outcomes, this study used students' final examination grades of the semester as the indicator of the

overall learning outcome. However, most test items in the final examination were from standardized proficiency tests focusing on learners' language skills. Consequently, higher-order thinking skills that students had been taught and trained in the course were not fully measured. Future research should use indicators that adequately measure course objectives and teaching practices so that pedagogical effectiveness can be appropriately understood.

The small sample size and context-specific weaknesses hamper the generalizability of this research. The about 50 participants were all first-year English majors with zero experience of flipped learning in a Chinese University, so some findings can be typical of them and may not be generalized to other student populations. Nonetheless, the dense description in the qualitative data may mitigate, to some extent, the limitation of the sampling weakness and help expand the case-to-case transferability of this study.

Appendix One: Ethical Approval

Educational Research | Lancaster University

11th November 2020

Dear Yan

Thank you for submitting your ethics application and additional information for **The effectiveness of learning activities in the flipped classroom**. The information you provided has been reviewed and I can confirm that approval has been granted for this project.

As Principal Investigator your responsibilities include:

- ensuring that (where applicable) all the necessary legal and regulatory requirements in order to conduct the research are met, and the necessary licenses and approvals have been obtained;
- reporting any ethics-related issues that occur during the course of the research or arising from the research (e.g. unforeseen ethical issues, complaints about the conduct of the research, adverse reactions such as extreme distress) to the Research Ethics Officer (Dr Murat Oztok or Dr Natasa Lackovic).
- submitting details of proposed substantive amendments to the protocol to **Dr Kyungmee Lee** for approval.

Please do not hesitate to contact your supervisor if you require further information about this.

Yours sincerely

Kathryn Doherty
Programme Co-ordinator
PhD in Higher Education: Research, Evaluation and Enhancement

Appendix Two: Letter of Invitation

Lancaster University

Faculty of Arts and Social Sciences and Management School Research Ethics Committee (FASS-LUMS REC)

ETHICS APPLICATION FORM FOR STAFF and PhD STUDENTS

Letter of Invitation

Dear Students,

I am Yan Shen. I lecture your Integrated English Course (IEC). I would like to invite you to participate in my PhD dissertation research. My research aims to understand to what extent learning activities in a flipped classroom enhance student engagement and their learning outcomes by evaluating the effects of the learning activities in our IEC class.

If you agree to participate, you will join in two focus group interviews, an hour or so for each, one in the first half of the semester, the other toward the end of the semester. In these focus group interviews, you will share and discuss your experiences in the learning activities with some of your classmates. Besides, I will observe how you take part in the learning activities both in class and in your LMS. I will also analyse your learning records in the LMS to study how you take part in learning activities and your learning outcomes by doing so.

Your contribution to the research will help to build up valuable insights into the areas of research interest. The data of the research will also be used to inform design and implementation of learning activities in the future flipped IEC classes to help us teach and learn more efficiently and effectively.

I am looking forward to your participation. Together, we enhance our performance in the IEC classes. Please don't hesitate to contact me if you have any queries about the research project: (y.shen13@lancaster.ac.uk). Take time to decide whether or not you wish to take part.

Sincerely,

Yan Shen

v02-19

Appendix Three: Participant Information Sheet

Lancaster University

Participant information sheet

Dear participant,

I am Yan Shen, a PhD student at Lancaster University and I would like to invite you to take part in a research study about the effects of learning activities in flipped classrooms.

Please take time to read the following information carefully before you decide whether or not you wish to take part.

What is the study about?
This study aims to explore and understand the extent learning activities in a flipped classroom enhance student engagement in learning and the consequent learning outcomes from an evaluative perspective. By doing so, the study hopes to bring about knowledge-based improvement to the design and implementation of learning activities in the flipped EFL classrooms and bring insights into teaching and learning enhancement.

Why have I been invited?
I have approached you because I am interested in understanding how students take part in learning activities in the flipped EFL classrooms. I would be very grateful if you would agree to take part in this study.

What will I be asked to do if I take part?

If you decided to take part, this would involve the following:

1. Participating in two focus group interviews, each of which will take about an hour, one in the first half of the semester and the other towards the end of the semester. Each group will have about 5 group members, and you will discuss your experiences in taking part in learning activities with other group members. The focus group will be audio-recorded and fully transcribed and anonymized.
2. Providing a permission to be observed when taking part in learning activities both in class and on the LMS throughout the semester. All the participants will be anonymized. Particular care will be paid to any identifiable information to protect anonymity.
3. Providing a permission to access your learning records of this course in MLS. Your learning activities recorded in the site will be very carefully treated and fully anonymised

What are the possible benefits from taking part?
If you take part in this study, your insights will contribute to our understanding of the extent the learning activities enhance student engagement and academic outcomes in the flipped EFL classrooms. This, in turn, will bring about knowledge-based improvement to the design and implementation of learning activities in the flipped classrooms and bring insights into teaching and learning enhancement. This study will allow you to share your experiences in the EFL learning activities with other learners, from which you will development a better self-knowledge of your own performance in the class. The evaluative nature of this

project will allow learners to have a clearer idea of the goals of learning English a foreign language, hence, it may help you develop clearer goals of endeavour.

Do I have to take part?
No. It's completely up to you to decide whether or not you take part. Your participation is voluntary. **If you decide not to take part in this study, this will not affect your studies and the way you are assessed on your course.**

What if I change my mind?
If you change your mind, you are free to withdraw at any time up to 2 weeks following your first focus group interview. If you want to withdraw, please let me know, and I will extract any ideas or information (=data) you contributed to the study and destroy them. However, it is difficult and often impossible to take out data from one specific participant when this has already been anonymised or pooled together with other people's data. Also, if you are involved in a focus group and then withdraw, your data will remain part of the study. This is mainly because your data is part of the ongoing conversation and cannot be destroyed. However, I will try to disregard your views when analysing the focus group data.

What are the possible disadvantages and risks of taking part?
It is unlikely that there will be any major disadvantages in taking part, though it will mean you will invest about 120 minutes altogether for two focus group interviews. You will also be observed when taking part learning activities during the semester, but this will take place unobtrusively.

Will my data be identifiable?

Particular care will be taken to protect participants from being identified. After the focus group and observation, only I, the researcher conducting this study will have access to the audio-recorded focus group data and other raw (pre-anonymised) data. My supervisor may check the anonymised data with me at appropriated time. I will keep all personal information about you (e.g. your name and other information about you that can identify you) confidential, that is, I will not share it with others. I will remove any personal information from the written record of your contribution. Besides, participants in the focus group will be asked not to disclose information outside of the focus group with anyone not involved in the focus group without the relevant person's expressed permission.

How will we use the information you have shared with us and what will happen to the results of the research study?
I will use the information you have shared with me only in the following ways:
I will use it for research purposes only. This will include my PhD dissertation and other academic publications. I may also present the results of my study at academic conferences. In addition, the data may be used to inform course design at the Departmental level, School and University level.

When writing up the findings from this study, I would like to reproduce some of the views and ideas you shared with me. I will only use anonymised quotes (e.g. from my interview with you), so that although I will use your exact words, all reasonable steps will be taken to protect your anonymity in our publications.

How my data will be stored
Your data will be stored in encrypted files (that is no-one other than me, the researcher will be able to access them) and on password-protected computers. No hard copies of any data will be generated in this study since all the data are digital. I will keep data that can identify you separately from non-personal information (e.g. your views on a specific topic). In

accordance with University guidelines, the data will be securely stored for a minimum of ten years. Lancaster University's Research Data Management Policy suggests that all researchers, PhD students included, deposit their data in Lancaster University's institutional data repository and make data freely available to other researchers with an appropriate data license. I will exclude all personal data from archiving.

What if I have a question or concern?

If you have any queries or if you are unhappy with anything that happens concerning your participation in the study, please contact myself:

Yan Shen Email: y.shen13@lancaster.ac.uk Tel: +8613917153355;

No. 7989, Wai Qing Song Rd, Qingpu District, Shanghai, 201701, P. R. China

or my supervisor: Dr. Kyungmee Lee

Email: k.lee23@lancaster.ac.uk Tel: +44 (0)1524 595158;

County South, Lancaster University, Lancaster, LA1 4YL, UK

If you have any concerns or complaints that you wish to discuss with a person who is not directly involved in the research, you can also contact:

Professor Paul Ashwin – Head of Department

Tel: +44 (0)1524 594443 Email: paul.ashwin@lancaster.ac.uk

County South, Lancaster University, Lancaster, LA1 4YD, UK.

This study has been reviewed and approved by the Faculty of Arts and Social Sciences and Lancaster Management School's Research Ethics Committee.

Thank you for considering your participation in this project.

Appendix Four: Consent Form

Lancaster University

Faculty of Arts and Social Sciences and Management School Research Ethics Committee (FASS-LUMS REC)

ETHICS APPLICATION FORM FOR STAFF and PhD STUDENTS

CONSENT FORM

Lancaster University

Project Title: Is learning enhanced by learning activities in the flipped classroom? A case study of a flipped EFL classroom from an evaluative perspective for improvement purpose

Name of Researchers: Yan Shen
Email: y.shen13@lancaster.ac.uk

Please tick each box

1. I confirm that I have read and understand the information sheet for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily	☐
2. I understand that my participation is voluntary and that I am free to withdraw at any time up to 2 weeks following my first focus group interview, without giving any reason. If I am involved in focus groups and then withdraw, my data will remain part of the study. I understand that as part of the focus group I will take part in, my data is part of the ongoing conversation and cannot be destroyed. I understand that the researcher will try to disregard my views when analysing the focus group data, but I am aware that this will not always be possible.	☐
3. If I am participating in the focus group I understand that any information disclosed within the focus group remains confidential to the group, and I will not discuss the focus group with or in front of anyone who was not involved unless I have the relevant person's express permission	☐
4. I understand that any information given by me may be used in future reports, academic articles, publications or presentations by the researcher/s, but my personal information will not be included and all reasonable steps will be taken to protect the anonymity of the participants involved in this project. 5. Anonymised data will be offered to Lancaster University's institutional data repository and will be made available to genuine research for re-use (secondary analysis)	☐
6. I understand that my name/my organisation's name will not appear in any reports, articles or presentation without my consent.	☐
7. I understand that any interviews or focus groups will be audio-recorded and transcribed and that data will be protected on encrypted devices and kept secure.	☐
8. I understand that data will be kept according to University guidelines for a minimum of 10 years after the end of the study.	☐
9. I agree to take part in the above study.	☐

_____ _____ _____

Name of Participant Date Signature

I confirm that the participant was given an opportunity to ask questions about the study, and all the questions asked by the participant have been answered correctly and to the best of my ability. I confirm that the individual has not been coerced into giving consent, and the consent has been given freely and voluntarily.

Signature of Researcher /person taking the consent_____ **Date** _____ Day/month/year

One copy of this form will be given to the participant and the original kept in the files of the researcher at Lancaster University

v02-19

Appendix Five: Focus Group Discussion Prompt Sheets

Lancaster University

Faculty of Arts and Social Sciences and Management School Research Ethics Committee (FASS-LUMS REC)

ETHICS APPLICATION FORM FOR STAFF and PhD STUDENTS

PROMPT SHEET OF FOCUS GROUP INTERVIEW (early-term)

(The questions in this prompt sheet will be used flexibly and reorganized around the topics in the interview)

1. Can you briefly describe what you are doing when taking part in the learning activities?
   a) Individual tasks, e.g. watching recorded pre-class lectures, taking quizzes?
   b) Pair work?
   c) Group tasks, e.g. group discussions, role play, etc.?
      i. How do you form the group?
      ii. How do you negotiate the tasks?
      iii. What do you do when disagreement arises in the group?
2. How do you evaluate the tasks?
3. How do you evaluate your performance in the tasks?
4. What are your emotional experiences in the course, esp. when taking part in the activities?
   a) How do you feel when working individually? With a peer? In a group?
   b) How do you feel when receiving feedback from peers? From the lecturer?
5. Do you have any suggestions for the learning activities?
   a) How should they be designed?
   b) How should they be implemented?

v02-19

PROMPT SHEET OF FOCUS GROUP INTERVIEW (end-of-term)

(The questions in this prompt sheet will be used flexibly and reorganized around the topics in the interview)

1. What are you doing to complete the tasks? And how is it working out?
   a) Individual work, e.g. when watching recorded pre-class lectures?
   b) Pair work?
   c) Group work, e.g. group discussion, role play, etc.?
2. Are you doing anything different from before esp. when taking part in these learning activities? And why?
   a) From what you used to do before taking part in this course?
   b) From what you did during the first half of the semester?
3. What are you going to do when taking part in these learning activities in the future?
4. Which learning activity do you think contributes most to your learning? Explain briefly.

Appendix Six: Field Note Form

Field Note	
Actors **Space** **Time** **objects**	FC Main Building 417A 11/26/2020 (feedback, individual exercises, group discussion, group work) Section 4.3
Goals	To have a good command of the expressions in the text; To be able to summarize the main ideas of the paragraphs, and role-play David and me.
Acts & activities &events	Students are always concentrated when doing individual exercises in class. Group discussion showed they had a good command of the text content but were still timid about expressing their ideas. Students had difficulty analyzing how the beginning and the conclusion echo. It was the first time they had discussions of questions of this kind. All groups finished their role-play in less than 10 minutes and started to work on their group project. Mari has not been participating in classroom activities. He was reviewing his quiz when others were having the group discussion. He seldom interacts with others in the class and does not perform well in his exercises. I need to talk to him in person to check out how to support him.
Feelings	
Reflections	What is the meaning of doing impromptu role-play/discussion in class? What do students benefit from that? There was a lot of noise. Most groups finish their work within minutes. Because they don't prepare, there is always not much to say. The discussions are impossible to go deep or broad. The time limit in class and impromptu performance can be causes of FC's underperformance in group work.

space: the physical setting;
actors: the people in the situation;
activities: the sets of related acts that are taking place.
objects: the artefacts and physical things that are there;
acts: the specific actions that participants are doing;
events: the sets of activities that are taking place;
time: the sequence of acts, activities and events;
goals: what people are trying to achieve;
feelings: what people feel and how they express this.
Reflections: the descriptions and analyses that have been done; the methods used in the observations and data collection and analysis; ethical issues, tensions, problems and dilemmas; the reactions of the observer to what has been observed and recorded – attitude, emotion, analysis, etc.; points of clarification that have been and/or need to be made; possible lines of further enquiry.

Appendix Seven: A Joint Display of the Qualitative Findings and the Quantitative Findings

Learning activities	Qualitative Findings	Quantitative Findings
Video	Both classes perceived the grammar videos as useful, but videos, on the whole, were less impressive and interactive;	$R_{w/l} > 1$ in FC; $R_{w/l} < 1$ in BC.
	FC students were satisfied with the videos, especially the grammar videos and felt video watching was efficient. They perceived videos' unique strengths of repeatability and controllability. Students had different perceptions of difficulty. The learning effects varied. Some reported anxiety. Students took videos as the primary resource of knowledge and reported improved language skills in listening, note-taking and paraphrasing. Students differ tremendously in technological skills. They demonstrated great agency.	FC significantly outperformed BC in their engagement in writing videos (U = 71, z = -4.97, p = .00, r = .68), grammar videos (U = 77, z = -.4.86, p = .00, r = .67) and text videos (U = 95, z = -.4.54, p = .00, r = .62). FC invested significantly more effort in writing videos (M = 1.61, SD = 0.40) and grammar videos (M = 1.66, SD = 0.57) than in text videos (M = 1.39, SD = 0.47). Video attempts in FC negatively were correlated with placement tests grades (p = .03, r = -.43)
	BC students used videos as supplements to lectures. They thought videos were handy and grammar videos were more effective than lectures. However, they still preferred lectures. Many just run over the videos; some never watched.	In BC, there was no statistically significant difference in student engagement between writing videos (M = 0.89, SD = 0.45, grammar videos (M = 87, SD = 0.50) and text videos (M = 0.77, SD = 0.34)
Quizzes	Both classes took quizzes as the indicator of learning effects. Students were emotionally affected by quiz results. They showed significant autonomy.	
	FC students were positive about quizzes. They felt quizzes were effective for learning because they were related to the knowledge content and motivated them to watch videos. Quizzes helped students realize the knowledge gap. They were more technology-savvy with LMS. Students had more interactions with peers about quizzes. Some circumvented quiz tasks. Some misunderstood video+quiz as the principal part of IEC.	FC students significantly outperformed BC students in their completion rate of quizzes (U = 225.00, Z = -3.27, p = .001, r = .45) and quiz results (t = 2.64, df = 51, p = .11, d = .73).
	BC students were more negative. They felt nervous due to the time limit and the dreadful experience of sitting	

	examinations. They were dissatisfied with the inconvenient LMS interface and had more problems with LMS. However, they felt better with LMS when quiz time was sufficient and started to embrace LMS.	BC students' quizzes grades were significantly correlated with their grades in the placement test (p = .016, r = .449) and the final examination (p = .008, r = .494)
Online forum	<u>There were fewer overall differences between FC and BC.</u> Both classes were positive to forum discussions. Students preferred simple and straight answers and demonstrated a distinct point-earning mentality. They joined forums to solve their problems and earn points by answering peers' questions. Students reported enhanced language knowledge. They were motivated to inquire and learned to compare and judge. Students demonstrated agency and collaboration. There were cases of non-participation. FC students were positive about forum discussions. Many deemed it the most useful and efficient. Most students reported that they benefited from reading peers' postings and felt supported. Students felt that forum discussion solved their puzzles and improved their understanding. BC: secondary to teacher explanations. Prefer to be led by teachers; doubt the reliability of peers' explanations;	The mean scores of $R_{p/d}$ for total postings in FC (M = 0.71, SD = 0.32) did not differ statistically significantly (t = .67, df = 51, two-tailed p = .50) from that in BC (M = 0.64, SD = 0.42, d = .19, CI = -0.35, 0.73). Students in FC placed significantly more postings in student-led discussion (U = 81.0, Z = -4.80, p = .000) and received far more likes from peers (U = .000, Z = -6.24, p = .000) than BC. Within FC, students placed more postings in student-led discussions (M = 1.10, SD = 0.56) than teacher-led topic discussions (M = 0.83, SD = 0.33), but the difference is not significant (Z = -2.00, p = .45). They placed significantly more postings (Z = -4.35, p = .00) in teacher-led topic discussions than in teacher-led exercise discussion (M = 0.23, SD = 0.26). They placed the least postings in student exemplar work (M = 0.05, SD = 0.10), which was significantly less than those in teacher-led exercise discussions (Z = -3.63, p = .00) Students in BC posted significantly more in teacher-led exercise discussion (U = 104, Z = -4.40, p = .000, r = .60), teacher-led topic discussion (U = 151, Z = -3.59, p = .000, r = .49) and student exemplar work (U = 229, Z = -2.48, p = .013) than FC. They placed the most postings in teacher-led topic discussions (M = 1.23, SD = 0.45), which was more than those in teacher-led exercise discussions (M = 0.98, SD = 0.72), but the difference is not significant (Z = -2.37, p = .02) at the significant level p = .008). They placed significantly fewer postings in student-led discussions (M = 0.38, SD = 0.39) than in teacher-led exercise discussions (Z = -4.27, p = .00). They placed the least postings in student exemplar work (M = 0.36, SD = 0.48), but the difference was not significant from student-led discussions (Z = -0.36, p = .72).

Teacher feedback	Both classes reported enhanced language knowledge, motivated reflection and higher-order thinking skills. Students demonstrated the least autonomy. Though some students said they reviewed their exercises for better effects, most just sat and listened. Seldom did students respond to teachers' questions. FC students held that teacher feedback was indispensable because it was authoritative and reliable. Teacher feedback solved their problems beyond them to handle, extended video content and forum discussion, guided thinking and motivated reflection. Some students felt it was irrelevant to their problems. BC students preferred to solve their puzzles via face-to-face teacher feedback. They felt that it refreshed their memory of knowledge content, made them more aware of their mistakes and helped them produce better language forms.	Data are not available.
Individual exercises	Both classes were comfortable with individual exercises due to the moderate workload and students' long-established habit of working independently. The exercises motivated students to learn texts carefully, but many were performance-oriented. Many complained about the unfriendly LMS interface. Students had shared problems in correcting errors by themselves and doing paraphrases. Most students made preparations to get higher marks. Many ignored teacher feedback. FC students held individual exercises as the most efficient and effective learning activity, more effective than listening to lectures. They reported enhanced impressions and improved language knowledge and skills. They were adapted quickly. They reviewed texts to complete the task within the time limit, were highly concentrated when doing the exercises, and exploited group discussions to improve their work. BC students had more negative feelings. They encountered more difficulties in doing the exercises, but feedback and	There was no significant statistical difference in students' essay writing in either completion rate ($U = 337.50$, $Z = -0.945$, $p = .13$) or revision attempts ($U = 247.00$, $Z = -1.84$, $p = .066$) between FC and BC. Nor was there in essay grades ($t = -0.45$, $df = 40.98$, $p = .656$, $d = .12$) FC significantly outperformed BC in individual exercise results ($t = 3.62$, $df = 51$, $p = .001$, $d = .96$). FC students' individual exercise grades were correlated with their final examination grades ($p = .004$, $r = .560$) and predicted their final examination grades with a moderate effect size ($\beta = .42$, $t = 2.53$, $p = .019$). FC students' essay attempts were positively associated with their essay grades ($p = .021$, $r = .458$) and their individual exercise grades ($p = .045$, $r = .405$) BC significantly outperformed FC in the completion rate of individual work ($U = 280.00$, $Z = -2.46$, $p = .014$). Within BC, students had a statistically significant

	support were not timely. Some were anxious to catch the deadline. They disapproved of the exercises or the digital form of assignments. There were distinct differences in performances. Some always had late submissions.	higher completion rate of individual work than quizzes (Z = -2.86, *p* = .004). BC students' individual exercise grades are significantly correlated with their final examination grades (*p* = .001, r = .583) and predicted their final examination grade with a strong effect size (β = .51, t = 3.40, *p* = .002).
Group discussion	FC students had divergent affective and behavioural engagements in group discussions, but they all emphasized teacher support. Many perceived that group discussions extended previous activities and made learning go deeper and broader. Learning became effective through interaction with peers and the teacher. They solved problems, expanded knowledge, improved understanding, approached problems appropriately and developed higher-order thinking skills. They demonstrated fair autonomy and managed discussion better as the semester went on. On the other hand, some students felt group discussions were less effective and difficult to follow. Some students were too shy to express their ideas. Some perceived no value in interactions.	Data are not available
interactive lecture	BC students were consistently positive about the lecture, describing it as substantive, systematic, experiential and easy to adapt. Many believed that they learned better in lectures than from videos. Some admitted that the lecture was of moderate effect, and they were less engaged in interaction, but they felt comfortable in lectures because they were habituated. They felt anxious and nervous only when answering questions. Some students lacked language skills to follow lectures in English or note-taking skills. Many reported that they learned well knowledge transmitted to them, and they managed to understand lectures in English in two weeks. Many of them prepared before class and used other resources to facilitate learning. Still, some students failed to stay focused in lectures.	Data are not available
Group work	<u>There were more differences among individual students than between classes.</u> Both classes had divergent affective reactions to group work. Many students held that group work did not pay off the time and effort. They suggested reducing workload. Many felt it challenging to	There was no significant difference between FC and BC in students' completion rate of the group work projects (U = 336.00, Z = -1.06, *p* = .29, r = .15), students' attempts of between-group peer assessment (U = 318.50, Z = -0.56, *p* = .57, r = .08) or students' attempts of within-group peer

	interact effectively between group members and hoped to work in smaller groups. Students preferred to work on more interactive and creative tasks. They felt that output tasks were more challenging. They disliked peer assessment the most. They reported increased generic knowledge and enhanced learning skills but less subject knowledge. They developed effective online communication skills and higher-order thinking skills. Students prioritized convenience when joining groups, and their performance was influenced by their language proficiency. Most of them did not interact effectively and efficiently. They settled minor conflicts by negotiation or compromise. Many of them assessed peers' work indiscrimately.	assessment (U = 251.00, Z = -1.77, *p* = .08, r = .24).
	FC students were fond of interactive tasks requiring good language competence and higher-order thinking skills, e.g. Debate. They were more accustomed to using LMS. They reported changes in attitude. Student autonomy was diverse. Some learned to keep working journals. However, some students did not make full use of classroom time.	Within FC, there was no significant difference (t[24] = -1.58, *p*[2-tailed] = .13, d = .32) between students' attempts of between-group peer assessment (M = 9.96, SD = 5.04) and those of within-group assessment (M = 11.48, SD = 4.51). Peer assessment attempts in FC were positively correlated with placement test grades (*p* = .019, r = .467). The mean score for group work projects in FC (M = 86.53, SD = 2.47) was statistically significantly lower (t = -2.56, df = 51, *p* = .014) than that in BC (M = 88.08, SD = 1.93, *d* = .70).
	BC students favoured creative tasks, e.g. role play. They felt that the debate was too challenging. They encountered more obstacles in interaction but regulated their tasks better than FC students. They did their projects online but could not guarantee participation from all group members.	BC did significantly more (Z = -2.57, *p*[2-tailed] = .01) within-group peer assessment (M = 13.43, SD = 5.12) than between-group peer assessment (M = 9.14, SD = 7.71).

*There was no significant difference (t = -0.35, df = 51, *p* = .726, d = .10) between FC (M = 71, SD = 9.25) and BC (M = 71.88, SD = 8.82) in the final examination grades. In both classes, student grades in the placement test (β_{FC} = .40, t = 2.37, *p* = .027; β_{BC} = .37, t = 2.49, *p* = .02) and individual exercises (β_{FC} = .42, t = 2.53, *p* = .019; β_{BC} = .51, t = 3.40, *p* =.002) can predict their grades in the final examination with at least a moderate effect size(R^2_{FC} = .40, F[2,22] = 9.12, *p* = .001; R^2_{BC} = .428, F[2,25] = 11.12, p < .001).

List of abbreviations

BC Blended Class

COVID Corona Virus Disease

EAP English for Academic Purposes

EFL English as a Foreign Language

EGP English for General Purpose

EMI English as Media of Instruction

ESP English for Specific Purposes

FC Flipped Class

FG Focus Group

GDP Gross Domestic Product

HE Higher Education

HEI Higher Education Institution

IEC Integrated English Course

L2 Second language

LMS Learning Management System

MoE Minister of Education

MOOC Massive Open Online Course

PI Principal Investigator

PPT Power Point

Pre-FG Pre Focus Group

Post-FG Post Focus Group

RQ Research Question

SHUPL Shanghai University of Political Science and Law

SLA Second Language Acquisition

SPOC Small Private Online Course

References

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

Adams, A. E. M., Garcia, J., & Traustadóttir, T. (2016). A quasi experiment to determine the effectiveness of a “partially flipped” versus “fully Flipped” undergraduate class in Genetics and Evolution. *CBE—Life Sciences Education, 15*, 1-9.

Akçayıra, 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

Al-Zahrani, A. M. (2015). From passive to active: The impact of the flipped classroom through social learning platforms on higher education students’ creative thinking. *British Journal of Educational Technology, 46*(6), 1133-1148. https://doi.org/10.1111/bjet.12353

Alderman, L., Towers, S., & Bannah, S. (2012). Student feedback systems in higher education: a focused literature review and environmental scan. *Quality in Higher Education, 18*(3), 261-280. https://doi.org/10.1080/13538322.2012.730714

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., CruiksHank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom’s Taxonomy of Educational Objectives*. New York: Longman.

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics, 34*(5), 502-508. https://doi.org/10.1111/opo.12131

Bamber, V. (2011a). Evaluative practice and outcomes: issues at the self-evaluative level. In M. Saunders, R. Bamber, & P. Trowler (Eds.), *Reconceptualising Evaluation in Higher Education: The Practice Turn* (pp. 193-199). Maidenhead: McGraw-Hill Education. https://ebookcentral.proquest.com/lib/lancaster/detail.action?docID=714184

Bamber, V. (2011b). Self-evaluative practice: diversity and power. In M. Saunders, R. Bamber, & P. Trowler (Eds.), *Reconceptualising Evaluation in Higher Education: The Practice Turn* (pp. 165-170). Maidenhead: McGraw-Hill Education. https://ebookcentral.proquest.com/lib/lancaster/detail.action?docID=714184

Barbour, R. S. (2013). Analyzing Focus Groups. In U. Flick (Ed.), *The Sage Handbook of Qualitative Data Analysis* (pp. 313-326). Thousand Oaks, CA: SAGE

Barratt-Pugh, L., Zhao, F., Zhang, Z., & Wang, S. (2018). Exploring current Chinese higher education pedagogic tensions through an activity theory lens. *Higher Education, 77*(5), 831-852. https://doi.org/10.1007/s10734-018-0304-8

Barton, G., & Ryan, M. (2013). Multimodal approaches to reflective teaching and assessment in higher education. *Higher Education Research & Development, 33*(3), 409-424. https://doi.org/10.1080/07294360.2013.841650

Bazeley, P. (2013). *Qualitative Data Analysis: Practical strategies* (1st ed.). London: SAGE.

Bazeley, P. (2021). *Qualitative Data Analysis: Practical Strategies* (2 ed.). London: SAGE Publications.

Bell, J., & Waters, S. (2014). *Doing Your Research Project: A Guide For First-Time Researchers*. Milton Keynes: McGraw-Hill Education. http://ebookcentral.proquest.com/lib/lancaster/detail.action?docID=1910218

BERA. (2011). *Ethical Guidelines for Educational Research*. London: British Educational Research Association.

Bergmann, J. (2017). *Solving the Homework Problem By Flipping the Learning*. Alexandria, Va: ASCD.

Bergmann, J., & Sams, A. (2014). *Flipped Learning: Gateway to Student Engagement*. Eugene, Oregon: International Society for Technology in Education.

Betihavas, V., Bridgman, H., Kornhaber, R., & Cross, M. (2016). The evidence for 'flipping out': A systematic review of the flipped classroom in nursing education. *Nurse Education Today, 38*, 15-21. https://doi.org/10.1016/j.nedt.2015.12.010

Biesenthal, C. (2014). Pragmatism. In D. Coghlan, & Brydon-Miller, M. (Ed.), *The sage encyclopedia of action research* (pp. 647-650). London: Sage Publications.

Bishop, J. L., & Verleger, M. A. (2013). *The flipped classroom: A survey of the research.* The 120th American Society of Engineering Education Annual Conference & Exposition, Atlanta, GA.

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in Higher Education, 38*(6), 698-712. https://doi.org/10.1080/02602938.2012.691462

Braun, E., & Leidner, B. (2009). Academic course evaluation. *European Psychologist, 14*(4), 297-306. https://doi.org/10.1027/1016-9040.14.4.297

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

Carbone, A., Ross, B., Phelan, L., Lindsay, K., Drew, S., Stoney, S., & Cottman, C. (2014). Course evaluation matters: improving students' learning experiences with a peer-assisted teaching programme. *Assessment & Evaluation in Higher Education, 40*(2), 165-180. https://doi.org/10.1080/02602938.2014.895894

Chang, C., & Lin, H.-C. K. (2019). Classroom interaction and learning anxiety in the IRS-integrated flipped language classrooms. *The Asia-Pacific Education Researcher, 28*(3), 193-201. https://doi.org/10.1007/s40299-018-0426-x

Chang, M.-M., & Hung, H.-t. (2019). Effects of technology-enhanced language learning on second language acquisition: A meta-analysis. *Educational Technology & Society, 22*(4), 1-17. https://www.jstor.org/stable/10.2307/26910181

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

Chen, S., Ouyang, F., & Jiao, P. (2021). Promoting student engagement in online collaborative writing through a student-facing social learning analytics tool. *Journal of Computer Assisted Learning, 38*(1), 192-208. https://doi.org/https://doi.org/10.1111/jcal.12604

Chen, Y., Wang, Y., Kinshuk, & Chen, N.-S. (2014). Is FLIP enough? Or should we use the FLIPPED model instead? *Computers & Education, 79*, 16-27. https://doi.org/10.1016/j.compedu.2014.07.004

Cheng, J., & Wei, L. (2021). Individual agency and changing language education policy in China: Reactions to the new 'Guidelines on College English Teaching'. *Current Issues in Language Planning, 22*(1-2), 117-134. https://doi.org/10.1080/14664208.2019.1700055

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

Cheng, S.-C., Hwang, G.-J., & Lai, C.-L. (2022). Critical research advancements of flipped learning: A review of the top 100 highly cited papers. *Interactive Learning Environments, 30*(9), 1751-1767. https://doi.org/10.1080/10494820.2020.1765395

Chuang, H.-H., Weng, C.-Y., & Chen, C.-H. (2018). Which students benefit most from a flipped classroom approach to language learning? *British Journal of Educational Technology, 49*(1), 56-68. https://doi.org/10.1111/bjet.12530

Cohen, J. (1988). *Statistical Power Analysis for Behavioral Science*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, L., Manion, L., & Morrison, K. (2011). *Research Methods in Education*. London: Taylor & Francis Group. http://ebookcentral.proquest.com/lib/lancaster/detail.action?docID=1144438

Collins, K. M. T., & Onwuegbuzie, A. J. (2013). Establishing interpretive consistency when mixing approaches: Role of sampling designs in evaluations. In D. M. Mertens & S. Hesse-Biber (Eds.), *Mixed methods*

and credibility of evidence in evaluation: New Directions for Evaluation, (Vol. 138, pp. 85-95). Somerset: John Wiley & Sons, Incorporated. https://ebookcentral.proquest.com/lib/lancaster/detail.action?docID=1216101

Cousin, G. (2008). Focus group research. In G. Cousin (Ed.), *Researching Learning in Higher Education: An Introduction to Contemporary Methods and Approaches* (pp. 51-69). London: Taylor & Francis Group. https://ebookcentral.proquest.com/lib/lancaster/detail.action?docID=380854.

Creswell, J. W. (2013). *Educational Research: Pearson New International Edition: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*. Harlow: Pearson Education UK. http://ebookcentral.proquest.com/lib/lancaster/detail.action?docID=5137595

Creswell, J. W. (2014). *A Concise Introduction to Mixed Methods Research*. Thousand Oaks, CA: Sage Publications.

Creswell, J. W., & Plano Clark, V. L. (2017). *Designing And Conducting Mixed Methods Research* (3rd ed.). Los Angeles, CA: SAGE.

de Jong, T. (2009). Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional Science, 38*(2), 105-134. https://doi.org/10.1007/s11251-009-9110-0

Denner, J., Marsh, E., & Campe, S. (2016). Approaches to reviewing research in education. In D. Wyse, N. Selwyn, E. Smith, & L. E. Suter (Eds.), *The BERA/SAGE Handbook of Educational Research* (pp. 143-164). London: SAGE Publications.

Edström, K. (2008). Doing course evaluation as if learning matters most. *Higher Education Research & Development, 27*(2), 95-106. https://doi.org/10.1080/07294360701805234

Ellis, R. (2003). *Task-based Language Learning and Teaching*. Oxford: Oxford University Press.

Ellis, R. (2005). *The Study of Second Language Acquisition*. Shanghai: Shanghai Foreign Language Education Press.

Ellis, R. A., Han, F., & Pardo, A. (2017). Improving Learning Analytics – Combining Observational and Self-Report Data on Student Learning *Educational Technology & Society, 20*(3), 158-169. https://www.jstor.org/stable/10.2307/26196127

Elmaadaway, M. A. N. (2018). The effects of a flipped classroom approach on class engagement and skill performance in a Blackboard course. *British Journal of Educational Technology, 49*(3), 479-491. https://doi.org/10.1111/bjet.12553

Erikson, M., Erikson, M. G., & Punzi, E. (2016). Student responses to a reflexive course evaluation. *Reflective Practice, 17*(6), 663-675. https://doi.org/10.1080/14623943.2016.1206877

Feng, X. Y., Wang, R. X., & Wu, Y. J. (2018). A literature review on blended learning: Based on analytical framework of blended learning *Journal of Distance Education, 3*, 13-24. https://doi.org/10.15881/j.cnki.cn33-1304/g4.2018.03.002

Fisher, R., & Miller, D. (2008). Responding to student expectations: A partnership approach to course evaluation. *Assessment & Evaluation in Higher Education, 33*(2), 191-202. https://doi.org/10.1080/02602930701292514

Fletcher, R. B., Meyer, L. H., Anderson, H., Johnston, P., & Rees, M. (2011). Faculty and Students Conceptions of Assessment in Higher Education.

Higher Education, 64(1), 119-133. https://doi.org/10.1007/s10734-011-9484-1

Freeman, R., & Dobbins, K. (2013). Are we serious about enhancing courses? Using the principles of assessment for learning to enhance course evaluation. *Assessment & Evaluation in Higher Education, 38*(2), 142-151. https://doi.org/10.1080/02602938.2011.611589

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? : An experiment with data saturation and variability. *Field methods, 18*(1), 59-82. https://doi.org/10.1177/1525822X05279903

Hall, J. N. (2013). Pragmatism, evidence, and mixed methods evaluation. In D. M. Mertens & S. Hesse-Biber (Eds.), *Mixed Methods and Credibility of Evidence in Evaluation: New Directions for Evaluation* (Vol. 138, pp. 15-26). Somerset: John Wiley & Sons, Incorporated. http://ebookcentral.proquest.com/lib/lancaster/detail.action?docID=1216101

Hao, Y. W. (2016). Exploring undergraduates' perspectives and flipped learning readiness in their flipped classrooms. *Computers in Human Behavior, 59*, 82-92.

Harkavy, I., & Puckett, J. (2014). Dewey, John. In D. Coghlan & M. Brydon-Miller (Eds.), *The SAGE Encyclopedia of Action Research* (pp. 252-256). London: Sage Publications. http://ebookcentral.proquest.com/lib/lancaster/detail.action?docID=1712670

Henderson, M., Selwyn, N., & Aston, R. (2015). What works and why? Student perceptions of 'useful' digital technology in university teaching and learning. *Studies in Higher Education, 42*(8), 1567-1579. https://doi.org/10.1080/03075079.2015.1007946

Hennink, M. M., & Leavy, P. (2014). *Focus Group Discussions*. Cary: Oxford University Press, Incorporated. http://ebookcentral.proquest.com/lib/lancaster/detail.action?docID=1630622

Hew, K. F., Bai, S., Dawson, P., & Lo, C. K. (2021). Meta-analyses of flipped classroom studies: A review of methodology. *Educational Research Review, 33*, 100393. https://doi.org/10.1016/j.edurev.2021.100393

Holland, E. P. (2019). Making sense of module feedback: accounting for individual behaviours in student evaluations of teaching. *Assessment & Evaluation in Higher Education, 44*(6), 961-972. https://doi.org/10.1080/02602938.2018.1556777

Holton, E. F., III. (1996). The flawed four-level evaluation model. *Human Resource Development Quarterly, 7*(1), 5-21. http://proquest.umi.com/login/athens?url=https://www.proquest.com/scholarly-journals/flawed-four-level-evaluation-model/docview/234903996/se-2?accountid=11979

Hughes, G. (2011). Towards a personal best: a case for introducing ipsative assessment in higher education. *Studies in Higher Education, 36*(3), 353-367. https://doi.org/10.1080/03075079.2010.486859

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

Hwang, G.-J., & Tsai, C.-C. (2011). Research trends in mobile and ubiquitous learning: A review of publications in selected journals from 2001 to 2010. *British Journal of Educational Technology, 42*(4), E65.

Jensen, J. L., Kummer, T. A., & d M Godoy, P. D. (2015). Improvements from a flipped classroom may simply be the fruits of active learning. *CBE Life Sci Educ, 14*(1), 1-12. https://doi.org/10.1187/cbe.14-08-0129

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26.

Johnson, R. B., Onwuegbuzie, A. J., Waal, C., Stefurak, T., & Hildebrand, D. (2016). Unpacking pragmatism for mixed methods research. In D. Wyse, Selwyn, N., Smith, E., & Suter, L. E. (Ed.), *The BERA/SAGE Handbook of Educational Research* (pp. 300-320). London: Sage publications.

Kahu, E. R. (2013). Framing student engagement in higher education. *Studies in Higher Education, 38*(5), 758-773. https://doi.org/10.1080/03075079.2011.598505

Kahu, E. R., & Nelson, K. (2018). Student engagement in the educational interface: Understanding the Mechanism of student success. *Higher Education Research & Development, 37*(1), 58-71. https://doi.org/10.1080/07294360.2017.1344197

Kember, D., Leung, D. Y. P., & McNaught, C. (2008). A workshop activity to demonstrate that approaches to learning are influenced by the teaching and learning environment. *Active Learning in Higher Education, 9*(1), 43-56. https://doi.org/10.1177/1469787407086745

Kim, J.-e., Park, H., Jang, M., & Nam, H. (2017). Exploring flipped classroom effects on second language learners' cognitive processing. *Foreign Language Annals, 50*(2), 260-284. https://doi.org/10.1111/flan.12260

Kirkpatrick, D. (1996). Great ideas revisited. (evaluation of employee training programs) (Celebrating 50 Years). *Training & Development, 50*(1), 54-59.

Kirkpatrick, D., & Kirkpatrick, J. (2005). *Transferring Learning to Behavior: Using the Four Levels to Improve Performance*. Oakland: Berrett-Koehler Publishers, Incorporated. http://ebookcentral.proquest.com/lib/lancaster/detail.action?docID=335397

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating Training Programs: The Four Levels*. Oakland: Berrett-Koehler Publishers, Incorporated. http://ebookcentral.proquest.com/lib/lancaster/detail.action?docID=335391

Kirschner, P. A., & van Merriënboer, J. J. G. (2013). Do learners really know best? Urban legends in education. *Educational Psychologist, 48*(3), 169-183. https://doi.org/10.1080/00461520.2013.804395

Kormos, J., & Trebits, A. (2012). The Role of Task Complexity, Modality, and Aptitude in Narrative Task Performance. *Language Learning, 62*(2), 439-472. https://doi.org/10.1111/j.1467-9922.2012.00695.x

Krause, K. L., & Coates, H. (2008). Students' engagement in first‐year university. *Assessment & Evaluation in Higher Education, 33*(5), 493-505. https://doi.org/10.1080/02602930701698892

Kuh, G. D. (2009). The national survey of student engagement: Conceptual and empirical foundations. *New Directions for Institutional Research, 141*, 5-20. https://doi.org/10.1002/ir.283

Kuh, G. D., Pace, C. R., & Vesper, N. (1997). The development of process indicators to estimate student gains associated with good practices in undergraduate education. *Research in Higher Education, 38*(4), 435-454.

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

Lai, H.-M. (2021). Understanding what determines university students' behavioral engagement in a group-based flipped learning context. *Computers & Education, 173*, 104290. https://doi.org/10.1016/j.compedu.2021.104290

Lai, H.-M., Hsieh, P.-J., Uden, L., & Yang, C.-H. (2021). A multilevel investigation of factors influencing university students' behavioral engagement in flipped classrooms. *Computers & Education, 175*, 104318. https://doi.org/10.1016/j.compedu.2021.104318

Lantolf, J. P. (2009). Second language learning as a mediated process. *Language Teaching, 33*(2), 79-96. https://doi.org/10.1017/s0261444800015329

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

Li, J., & De Luca, R. (2012). Review of assessment feedback. *Studies in Higher Education, 39*(2), 378-393. https://doi.org/10.1080/03075079.2012.709494

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/10.2307/26388398

Lin, C.-J., & Mubarok, H. (2021). Learning analytics for investigating the mind map-guided AI chatbot approach in an EFL flipped speaking classroom.

Educational Technology & Society, 24(4), 16-35. https://www.jstor.org/stable/10.2307/48629242

Liu, C. Y., Li, D., Zhang, B. R., & Hu, X. L. (2019). Teaching effectiveness of SPOC flipped classroom in college: A systematic review and meta-analysis. *Open Education Research, 25*(1), 82-91+36. https://doi.org/10.13966/j.cnki.kfjyyj.2019.01.009

Liu, Y. C., & Li, J. C. (2020). *Educational Statistics Yearbook of China 2020*. Beijing: China Statistic Press.

Lombardini, C., Lakkalab, M., & Muukkonenc, H. (2018). The impact of the flipped classroom in a principles of microeconomics course: evidence from a quasi-experiment with two flipped classroom designs. *International Review of Economics Education* (29), 14-28. https://doi.org/10.1016/j.iree.2018.01.003

Mabry, L. (2003). In living color: Qualitative methods in educational evaluation. In T. Kellaghan, D. L. Stufflebeam, & L. A. Wingate (Eds.), *Springer International Handbooks of Education: International Handbook of Educational Evaluation*. Springer Science+Business Media. http://ezproxy.lancs.ac.uk/login?url=https://search.credoreference.com/content/entry/spredev/1evaluation_theory_and_metatheory/0? institutionId=3497

Macfadyen, L. P., & Dawson, S. (2012). Numbers are not enough. Why e-learning analytics failed to inform an institutional strategic plan. *Educational Technology & Society, 15*(3), 149-163. https://www.jstor.org.stable/10.2307/jeductechsoci.15.4.149?seq=1&cid=pdf-reference#references_tab_contents

Martinez, F., Taut, S., & Schaaf, K. (2016). Classroom observation for evaluating and improving teaching: An international perspective. *Studies in Educational Evaluation, 49*, 15-29. https://doi.org/10.1016/j.stueduc.2016.03.002

Maxwell, J. A. (2006). Literature reviews of, and for, educational research: A commentary on Boote and Beile's "Scholars before researchers". *Educational Research, 35*(9), 28-31.

McNally, B., Chipperfield, J., Dorsett, P., Fabbro, L. D., Frommolt, V., Goetz, S., Lewohl, J., Molineux, M., Pearson, A., Reddan, G., Roiko, A., & Rung, A. (2017). Flipped classroom experiences: student preferences and flip strategy in a higher education context. *High Educ, 73*, 281-298. https://doi.org/10.1007/s10734-016-0014-z

Mertens, D. M., & Hesse-Biber, S. (2013). Mixed Methods and Credibility of Evidence in Evaluation. In D. M. Mertens & S. Hesse-Biber (Eds.), *Mixed Methods and Credibility of Evidence in Evaluation: New Directions for Evaluation* (Vol. 138, pp. 5-14). Somerset: John Wiley & Sons, Incorporated. http://ebookcentral.proquest.com/lib/lancaster/detail.action?docID=1216101

MoE. (2001). *Guideline for Improving the Quality of Undergraduate Teaching*. Beijing: Ministry of Education of the People's Republic of China.

MoE. (2010). *Outline of China's National Plan for Medium and Long-term Education Reform and Development (2010-2020)*. Beijing: Ministry of Education of the People's Republic of China.

Müller-Hartmann, A., & Ditfurth, M. S.-v. (2012). Research on the use of technology in task-based language learning. In M. Thomas & H. Reinders (Eds.), *Task-Based Lanugage Learning and Teaching with Technology* (pp. 17-40). London: Continuum International Publishing Group.

Munir, M. T., Baroutian, S., Young, B. R., & Carter, S. (2018). Flipped classroom with cooperative learning as a cornerstone. *Education for Chemical Engineers*(23), 25-33. https://doi.org/10.1016/j.ece.2018.05.001

Nguyen, T., & Foster, K. A. (2018). Research note—Multiple time point course evaluation and student learning outcomes in an MSW course. *Journal of Social Work Education, 54*(4), 715-723. https://doi.org/10.1080/10437797.2018.1474151

Nichols, S. L., & Dawson, H. S. (2012). Assessment as a context for student engagement. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of Research on Student Engagement* (pp. 457-477). New York, NY: Springer. https://doi.org/10.1007/978-1-4614-2018-7_22

Nicol, D., Thomson, A., & Breslin, C. (2013). Rethinking feedback practices in higher education: a peer review perspective. *Assessment & Evaluation in Higher Education, 39*(1), 102-122. https://doi.org/10.1080/02602938.2013.795518

Nicolaou, M., & Atkinson, M. (2019). Do student and survey characteristics affect the quality of UK undergraduate medical education course evaluation? A systematic review of the literature *Studies in Educational Evaluation, 62*, 92-103. https://doi.org/10.1016/j.stueduc.2019.04.011

Nielsen, T., & Kreiner, S. (2017). Course evaluation for the purpose of development: What can learning styles contribute? *Studies in Educational Evaluation, 54*, 58-70. https://doi.org/10.1016/j.stueduc.2016.10.004

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International journal of Qualitative Methods, 16*, 1-13. https://doi.org/10.1177/1609406917733847

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

Papamitsiou, Z., & Economides, A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence *Educational Technology & Society, 17*(4), 49-64. https://www.jstor.org/stable/10.2307/jeductechsoci.17.4.49

Pardo, A. (2014). Designing learning analytics experiences. In J. A. Larusson & B. White (Eds.), *Learning Analytics: From Research to Practice*. New York: Springer New York. https://doi.org/10.1007/978-1-4614-3305-7_2

Park, J. Y. (2014). Course evaluation: reconfigurations for learning with learning management systems. *Higher Education Research & Development, 33*(5), 992-1006. https://doi.org/10.1080/07294360.2014.890564

Patton, M. Q. (1998). Discovering process use. *Evaluation, 4*(2), 225-233.

Patton, M. Q. (2012). *Essentials of Utilization-Focused Evaluation.* Thousand Oaks, CA: Sage Publications.

Paull, M., Whitsed, C., & Girardi, A. (2016). Applying the Kirkpatrick model: Evaluating an Interaction for Learning Framework curriculum intervention. *Issues in Educational Research, 26*(3), 490-507.

Praslova, L. (2010). Adaptation of Kirkpatrick's four level model of training criteria to assessment of learning outcomes and program evaluation in Higher Education. *Educational Assessment, Evaluation and Accountability, 22*(3), 215-225. https://doi.org/10.1007/s11092-010-9098-7

Reeve, J. (2013). How students create motivationally supportive learning environments for themselves: The concept of agentic engagement. *Journal of Educational Psychology, 105*(3), 579-595. https://doi.org/10.1037/a0032690

Richardson, J. T. E. (2005). Students' approaches to learning and teachers' approaches to teaching in higher education. *Educational Psychology, 25*(6), 673-680. https://doi.org/10.1080/01443410500344720

Rosenstein, B. (2014). Evaluation. In D. Coghlan & M. Brydon-Miller (Eds.), *The sage encyclopedia of action research* (pp. 314-318). London: Sage Publications.

Saunders, M. (2011). Setting the scene: the four domains of evaluative practice in higher education. In M. Saunders, R. Bamber, & P. Trowler (Eds.), *Reconceptualising Evaluation in Higher Education: The Practice Turn* (pp. 1-17). Maidenhead: McGraw-Hill Education. https://ebookcentral.proquest.com/lib/lancaster/detail.action?docID=714184

Saunders, M. (2012). The use and usability of evaluation outputs: A social practice approach. *Evaluation, 18*(4), 421-436. https://doi.org/10.1177/1356389012459113

Saunders, M., Trowler, P., & Bamber, V. (2011). The practice turn: reconceptualizing evaluation in higher education. In M. Saunders, R. Bamber, & P. Trowler (Eds.), *Reconceptualising Evaluation in Higher Education: The Practice Turn* (pp. 203-226). Maidenhead: McGraw-Hill Education. https://ebookcentral.proquest.com/lib/lancaster/detail.action?docID=714184

Scriven, M. (1996). Types of evaluation and types of evaluator. *Evaluation Practice, 17*(2), 151-161.

Scriven, M. (2003). Evaluation theory and metatheory. In T. Kellaghan, D. L. Stufflebeam, & L. A. Wingate (Eds.), *Springer International Handbooks of Education: International Handbook of Educational Evaluation*. Springer Science+Business Media. http://ezproxy.lancs.ac.uk/login?url=https://search.credoreference.com/content/entry/spredev/1evaluation_theory_and_metatheory/0? institutionId=3497

Selwyn, N. (2016). Digital downsides: exploring university students' negative engagements with digital technology. *Teaching in Higher Education, 21*(8), 1006-1021. https://doi.org/10.1080/13562517.2016.1213229

Sozer, E. M., Zeybekoglu, Z., & Kaya, M. (2019). Using mid-semester course evaluation as a feedback tool for improving learning and teaching in higher education. *Assessment & Evaluation in Higher Education, 44*(7), 1003-1016. https://doi.org/10.1080/02602938.2018.1564810

Spolsky, B. (2000). *Conditions for Second Language Learning*. Shanghai: Shanghai Foreign Language Education Press.

Steyn, C., Davies, C., & Sambo, A. (2019). Eliciting student feedback for course development: the application of a qualitative course evaluation tool among business research students. *Assessment & Evaluation in Higher Education, 44*(1), 11-24. https://doi.org/10.1080/02602938.2018.1466266

Stöhr, C., Demazière, C., & Adawi, T. (2020). The polarizing effect of the online flipped classroom. *Computers & Education, 147*, 103789. https://doi.org/10.1016/j.compedu.2019.103789

Strayer, J. F. (2012). How learning in an inverted classroom influences cooperation, innovation and task orientation. *Learning Environments Research, 15*(2), 171-193. https://doi.org/10.1007/s10984-012-9108-4

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

Sun, J. J.-M., Hu, P., & Ng, S. H. (2016). Impact of English on education reforms in China: with reference to the learn-English movement, the

internationalisation of universities and the English language requirement in college entrance examinations. *Journal of Multilingual and Multicultural Development, 38*(3), 192-205. https://doi.org/10.1080/01434632.2015.1134551

Taras, V., Caprar, D. V., Rottig, D., Sarala, R. M., Zakaria, N., Zhao, F., ., & Huang, V. Z. (2013). A Global Classroom? Evaluating the Effectiveness of Global Virtual Collaboration as a Teaching Tool in Management Education. *Academy of Management Learning & Education, 12*(3), 414-435. https://www.jstor.org/stable/43696576

Tight, M. (2012). *EBOOK: Researching Higher Education*. Maidenhead: McGraw-Hill Education. http://ebookcentral.proquest.com/lib/lancaster/detail.action?docID=932633

Trowler, P. (2015). Change theory and changing practices: Enhancing student engagement in universities. http://www.research.lancs.ac.uk/portal/files/91990419/FINAL_Trowler_Student_Engagement_in_Europe.pdf

Trowler, P. (2016). *Doing Doctoral Research into Higher Education…..and getting it right*. CreateSpace Independent Publishing Platform: http://www.amazon.co.uk/Doctoral-Research-Higher-Education-getting/dp/1523779551/.

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

Vaismoradi, M., Turunen, H., & Bondas, T. (2013). Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. *Nursing and Health Science, 15*(3), 398-405. https://doi.org/10.1111/nhs.12048

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

Wilson, C., Sims, S., Dyer, J., & Handley, F. (2022). Identifying opportunities and gaps in current evaluation frameworks – the knowns and unknowns in determining effective student engagement activity. *Assessment & Evaluation in Higher Education, 47*(6), 843-856. https://doi.org/10.1080/02602938.2021.1969536

Wu, W.-C. V., Chen Hsieh, J. S., & C., Y. J. (2017). Creating an online learning community in a flipped classroom to enhance EFL learners' oral proficiency. *Educational Technology & Society, 20*(2), 142-157.

Xerri, M. J., Radford, K., & Shacklock, K. (2017). Student engagement in academic activities: a social support perspective. *Higher Education, 75*(4), 589-605. https://doi.org/10.1007/s10734-017-0162-9

Yilmaz, K. (2013). Comparison of quantitative and qualitative research traditions: epistemological, theoretical, and methodological differences. *European Journal of Education, 48*(2), 311-325. https://www.jstor.org/stable/26357806

Yu, S., Zhang, Y., Zheng, Y., Yuan, K., & Zhang, L. (2018). Understanding student engagement with peer feedback on master's theses: a Macau study. *Assessment & Evaluation in Higher Education, 44*(1), 50-65. https://doi.org/10.1080/02602938.2018.1467879

Zhang, H., & Bournot-Trites, M. (2021). The long-term washback effects of the National Matriculation English Test on college English learning in China: Tertiary student perspectives. *Studies in Educational Evaluation, 68*, 100977. https://doi.org/10.1016/j.stueduc.2021.100977

Zhang, J. H., Zhagn, Y. X., Zou, Q., & Huang, S. (2018). What Learning Analysis Tells Us: Group Behavior Analysis and Individual Learning Diagnosis based on Long-Tern and Large Scale Data. *Educational Technology & Society, 21*(2), 245-258. https://www.jstor.org/stable/10.2307/26388404

Zhoc, K. C. H., Webster, B. J., King, R. B., Li, J. C. H., & Chung, T. S. H. (2018). Higher Education Student Engagement Scale (HESES): Development and Psychometric Evidence. *Research in Higher Education, 60*(2), 219-244. https://doi.org/10.1007/s11162-018-9510-6