Article:
Wang, J, An, N and Wright, C orcid.org/0000-0003-3962-7903 (2018) Enhancing beginner learners' oral proficiency in a flipped Chinese foreign language classroom. Computer Assisted Language Learning, 31 (5-6). pp. 490-521. ISSN 0958-8221

https://doi.org/10.1080/09588221.2017.1417872

Enhancing Beginner Learners' Oral Proficiency in a Flipped Chinese Foreign Language Classroom

Jun Wang[a], Na An[b]* and Clare Wright[c]

[a] School of Humanities, Shanghai Jiao Tong University, RM 239 Lecture Hall No.1, No.1954 Huashan RD, Shanghai, China;

[b]School of Humanities, Shanghai Jiao Tong University, RM 327 Lecture Hall No.1, No.1954 Huashan RD, Shanghai, China;

[c]School of Languages, Cultures and Societies, RM B48 Michael Sadler Building, University of Leeds, Leeds, UK

Abstract: Flipped instruction has become a hot issue in foreign language teaching technology, a trend intensified by the emergence of Massive Open Online Courses (MOOCs). In this study, we tracked learners in a beginner-level Chinese Foreign Language classroom to see if flipped teaching based on a MOOC made a difference to their oral proficiency development and rate of progress, compared to a class-only baseline group, using the same syllabus over one semester. Language development, pre and post-intervention, was assessed by standard complexity, accuracy and fluency measures, alongside subjective teacher ratings. Learners' investment of time and perceptions of the new method were also investigated. Results showed that learners exposed to flipped instruction significantly ($p<.01$) outperformed the baseline group in oral proficiency in many measures at post tests, especially in speech fluency, though their advantage in complexity and accuracy was less evident. Rate of progress through the syllabus for the flipped group was also faster, requiring 25% less face-to-face time. Learners in the flipped group also demonstrated more (out of class) time investment in their learning and more positive attitudes toward the course, though these two factors did not significantly associate with the proficiency measures. These results support the implementation of flipped instruction in foreign language classrooms for both better and faster learner improvement; we explore how far psycho-social models of active learning might explain its methodological advantages.

Key words: flipped classroom; Chinese L2 oral proficiency; MOOC; foreign language teaching

Introduction

The past 5 years have seen a rapid and explosive development in MOOC (massive online open course) technology. Since its emergence as a popular mode of learning in 2012 (MacLeod, Haywood, & Woodgate, 2015), many research studies have addressed the advantages and insufficiencies of MOOCs (e.g. see Furneaux, Wright,& Wilding, 2015). Implementing pure MOOCs alone in higher education still remains a controversial issue, for both pedagogic and resourcing reasons.

However, the blossoming of MOOCs has helped, perhaps indirectly, revive another trend of educational methodology. Flipped classrooms - also known as inversed teaching or blended learning (King, 1993) - have thrived alongside MOOCs. A definition given by Lage, Platt, & Treglia (2000) can concisely explain its nature: "Inverting the classroom means that events that have traditionally taken place inside the classroom now take place outside the classroom and vice versa". Table 1 which

* Corresponding author. Email: anna@sjtu.edu.cn

is borrowed from Bishop, & Verleger (2013) demonstrates the typical (restricted) understanding of flipped classroom. By reversing the traditional learning environment - delivering instructional content (in the form of course videos, or more ideally, MOOCs) outside of the classroom while bringing other activities (including "homework") into the classroom–flipped classroom could be seen as ideally suited to work within a MOOC approach. Arguably, the combination of online instruction with traditional classrooms could overcome the problems of pure MOOCs, especially the rapid loss of learners through time and difficulties in providing instruction. Nevertheless, it should be noted that many flipped classrooms having been studied so far, including most of the studies mentioned hereinafter, used means other than MOOCs as the out-of-class content-delivery methods, therefore fitting a broader definition of flipped learning by Bishop, & Verleger (2013): "an educational technique that consists of two parts: interactive group learning activities inside the classroom, and direct computer-based individual instruction outside the classroom."

Table 1.Restricted definition of the flipped classroom. (borrowed from Bishop, & Verleger, 2013)

Style	Inside Class	Outside Class
Traditional	Lectures	Practice Exercises & Problem solving
Flipped	Practice Exercises & Problem solving	Video Lectures (MOOCs)

There have been many studies focused on the design, delivery, assessment and students' acceptance of flipped teaching over the past five years. Yarbro, Arfstrom, McKnight, & McKnight's summary (2014) of recent research on flipped learning showed how widely flipped learning has become embedded in most disciplines, including foreign languages, and widely used in higher education. Studies have found that flipped teaching can generate either better learning experiences/higher student satisfaction (Smit, Brabander, & Martens., 2014; Wilson, 2013; Strayer, 2012), or more learning investment (Hung, 2015) or better academic performance (Flumerfelt, & Green, 2013; Forsey, Low, & Glance 2013), or a combination of the above benefits(Tun, Sturek, & Basile, 2013; Hung, 2015), though the interactions between those aspects are not clear.

Foreign language classrooms have been included in this trend (see e.g. Basal, 2015; Mehring, 2016) with evidence that flipped classrooms can support the implementation of a communicative, student-centered learning in EFL settings. Hung (2015) found that flipped lessons helped the students attain better learning outcomes, develop better attitudes toward their learning experiences, and devote more effort within the learning process. Chen, Wu, & Marek (2016) exposed their subjects to learning English idioms by flipped learning, using the LINE smartphone app, compared to using conventional instruction. Results showed that flipped instruction not only significantly improved learners' idiomatic knowledge but also enhanced their motivation, making them more active in using idioms in class.

We argue that further research is required to broaden these initial findings of flipped learning benefits. Firstly, EFL classrooms, especially at college level, are limited in validity, since most college English L2 learners are already "advanced", so effects on progress in the steeper early stages of learning also need to be researched. We also should look beyond English-focused learning, considering the comparative lack of learning resources for any other language than English. If the flipped model does work, it will directly benefit all the learners of that language around the world, particularly for rapidly growing languages such as Mandarin Chinese (Wright, & Zhang, 2014). Since

the out-of-class (usually online) instructional material could be shared through the internet, the replication of the flipped method would not be difficult.

Secondly, previously studied flipped foreign language classrooms all used tools for general learning or online communication purposes, such as TED-ED (Hung 2015) or the Line app (Chen, Wu, & Marek 2016), therefore are all flipped classrooms by the broader definition cited above (Bishop, & Verleger, 2013). It would be better to create, a more narrowly defined "typical" flipped classroom using a specifically designed L2 MOOC, which is publically available on a major MOOC platforms, as the sole means of out-of-class instruction, since as we believe the "optimal effect" of flipped instruction might emerge with maximal course content delivered in the out-of-class phase.

Thirdly, previous studies (such as Hung, 2015) have tended to use limited indices of linguistic knowledge rather than assessing general proficiency and communicative effectiveness (e.g. through oral performance). We argue that longitudinal data, gathered through a semester-long interventionist study, measuring proficiency development through objective linguistic indicators, would improve the methodological validity of flipped research and give it wider relevance to language teaching research.

Finally, it remains unclear whether flipped teaching impacts more on the learners' actual academic performance, or their attitudes towards the course (motivation, investment of time and effort), or both. A well-designed study which maps the interaction between these factors will provide a better integrated picture of the mechanisms by which flipped teaching (might) enhance learners' performance in a foreign language classroom.

The Active Learning framework used in the present study

We now briefly discuss the key learning framework on which we base our assumptions about how flipped learning works in general and for language learning particular – Active Learning Theory.

Active learning has been prevalent in many disciplines since the 1980s, and fits well with assumptions about learner autonomy used in communicative language teaching approaches (Whong, 2013). Active learning, expecting students to learn by thinking about the things they are doing (Bonwell, & Eison, 1991), means students "approach course content through problem-solving exercises, informal small groups … and other activities—all of which require students to apply what they are learning" (Meyers, & Jones, 1993, p. 6). Active learning encompasses a wide variety of forms, including guided, self-directed learning and cooperative learning (Shen, & Xu, 2015), which could possibly explain the power of flipped method in foreign language education.

First of all, guided, self-directed learning could justify not using class time to deliver lectures. This refers to situations in which students make learning decisions with the guidance of an expert(Hout-Wolters, Simons, &Volet, 2000).Studies have shown that, in second language (L2) classrooms, students who undergo guided, self-directed learning have a positive attitude toward their ability to study, to take the initiative, and to play important roles in successfully learning the material (Cotterall, 1995). In addition, their active usage of cognitive, meta-cognitive, affective and social strategies is significantly associated with gains in language proficiency (Gan, 2004). In well-designed flipped instruction, the self-directed learning could take place in the out-of-class phase. With ready-made online learning tools, especially with a tailor-made MOOC, a learner could easily fit him/herself into the course content with a study plan that matches his/her learning style and habit well, given that the limitation of a real classroom is lifted. Additionally, we argue there is another advantage. Since self-directed learning can move the delivery of most of linguistic knowledge out of

the real classroom, in the face-to-face class, not only can students spend more time in cooperative activities as we will mention next, but also a lot of time can be saved for extra learning. This effect has been frequently reported (Prefume, 2015; McDonald, & Smith, 2013), which we believe is highly valuable at college level and which we have tried to verify again in this paper.

Second, cooperative learning may explain the effectiveness of the face-to-face teaching phase in a flipped model. The theory emerged in the1970s (Johnson, & Johnson, 1974). As a classroom learning approach, cooperative learning refers to organizing classroom activities in which students work in groups to complete tasks collectively based on the principles of positive interdependence, individual accountability, mutual interaction, and group decision making (Johnson, Johnson, & Stanne, 2000; Shaaban, & Ghaith, 2005). Studies have shown that, in the L2 context, cooperative learning promotes positive attitudes toward learning content and instructional experiences, greater motivation to achieve, and more supportive and caring relationships between peers and teachers (Gunderson, & Johnson, 1980). Studies of cooperative learning have shown that it results in overall improved target language skills (Bejarano,1987;McDonough, 2004; Seostek, 1994) and aids both low-achieving and high-proficiency students in acquiring language mechanics (Ghaith, &Yaghi, 1998; Shokouhi, & Alishaei, 2009). It is important to note that although the principles and practices of cooperative leaning are not novel to second language educators and learners, a flipped classroom can magnify its strength by enabling more class time to be used for cooperative activities while learners come into the classroom with most of required linguistic knowledge.

The two phases of a flipped classroom directed by active learning correspond well to Bloom (1964)'s taxonomy of educational objectives, ranging from lower to higher and simple to complex levels of cognitive thinking (i.e., knowledge, comprehension, application, analysis, synthesis, and evaluation). While the low level knowledge is left for students to acquire out of class, the time meeting with the instructor (in this case more a "guide" than a "bearer") and peers is optimally used for acquisition of high level knowledge and skills through complex cognitive activities and meaningful practices. As believed by many, the joint force of the two phases is strong.

Hence our study aimed to examine the effectiveness of MOOC-based flipped learning compared with traditional foreign language classrooms within the perspective of these theories of learning, with a novel perspective combining measures of linguistic development alongside affective factors, and tracking speed of progress.

A college level beginner's Mandarin Chinese class was redesigned and analyzed as the basis for the study. We focused on learners' development of oral proficiency as the main dependent variable for the following reasons: 1) Oral proficiency offers a broad perspective on language ability (Carroll, 1961; Anderson, 1982), allowing us here to draw more generalizable conclusions regarding flipped teaching effects on language development; 2) a beginner-level oral class is usually balanced in delivering both L2 knowledge and communicative skills, allowing us to evaluate how students used the flipped approach optimally to balance their time in developing different levels of knowledge or skill; 3) We wished to avoid confounding factors created by requiring students to learn characters, thus avoiding any additional literacy challenges in mastering the Chinese orthographic system (Everson, 2011) which is especially difficult for western learners (Ke, Wen, & Kotenbeutel, 2001). Interestingly, at the time of our study, the 4 major Mandarin Chinese MOOCs available on Courera.org and Edx.org, all abstain from the teaching of Chinese characters or set it as "optional".

In sum, our quasi-experimental study conducted at a Chinese university aimed to investigate

whether a flipped teaching method could have a positive impact on Chinese L2 learning outcomes at beginner level (measured as oral proficiency), alongside learners' investment in learning[1], and how these elements interacted. Our three research questions were:

RQ1: Can flipped teaching based on a MOOC enhance the development of learners' oral proficiency in a college level Chinese L2 beginner's class, measured in better linguistic performance and faster rate of progress, compared to a traditional non-flipped classroom?

RQ2: Are there differences in learners' perception of/attitude toward and investment in their learning comparing the two teaching methods (the flipped Chinese L2 class and the traditional class)?

RQ3: Are there interactions between teaching method, learners' attitudes/investment, and learning outcomes (oral proficiency)?

Method

Participants

The experimental group were 42 adult learners in a beginners' class of Mandarin Chinese from various non-sinosphere L1 backgrounds (i. e. Chinese characters in any form are not used in their L1 writing system). None of them were Chinese heritage learners. At the time of the study, they were first year international graduate students at a university in China taking various majors, and were required to take a Mandarin Chinese course for 1.5 hours per week through one semester (16 weeks) in partial fulfillment of the requirements for their academic degree. Almost all of them had arrived in China only about 1 week before the experiment started, and placement test showed that they had no previous exposure to Chinese, therefore the identity of "true beginner" could be guaranteed.[2] By the end of the experiment, 11 participants' data were discarded due to low attendance rate or failure to submit all required data (surveys/logs/quizzes – see below). 31 learners' data were thus used for final analysis.[3]

The baseline group were another group of 40 beginning learners of Chinese with the identical background as the experimental group; 10 participants were discarded for incomplete data and attendance, as for the group above, leaving 30 learners' data used for the final analysis. The first language background of learners in the two groups are unbiased and evenly distributed.[4]

1"Learners' investment" can have different meanings in different paradigms. Limited by research condition, in this study only learners' self-reported time investment out of class was measured. The same measurement was taken by another study on flipped instruction for foreign language (Hung, 2015). We assume this can at least make the two studies comparable.

2Unlike most college level EFL classes, CFL classes in China often includes lots of "true" zero beginners. For the current study, all 82 participants took a placement test together with 200+ other learners at the beginning of the semester. The placement test had a written part (at the difficulty level of HSK 1/very easy) and an oral part (also very basic). Learners were told not to guess any of the questions (since they were all in the multiple choice form) if they don't understand the questions in the written test. The 82 learners all received 0 percent in the written test and could produce nothing more than greeting words like "ni3hao3" (hello) in the oral test, and were therefore randomly placed in the two beginners' classes involved in this study. Other learners with higher proficiency were all placed in different (more advanced) classes.

3The 31 learners were from 21 countries and with 18 different first languages respectively, as shown in the following list in the format of "Nationality (first language/number of learners if multiple)": Pakistan (Urdu/4), Italy (Italian/3), US (English/3), France (French/2), Nepal (Nepali/2), Norway (Norwegian/2), Afghanistan (Dari), Bangladesh (Bengali), Canada (English), Chile (Spanish), Denmark (Danish), Ethiopia (Amharic), Hungary (Hungarian), Mongolia (Mongolian), Philippines (Filipino/English), Russia (Russian), Slovakia (Slovak), Spain (Spanish), Sweden (Swedish), Turkey (Turkish), UK (English)

4The 30 learners were from 20 countries and with 18 different first languages respectively, as shown in the following list in the format of "Nationality (first language/number of learners if multiple)":Pakistan (Urdu/5), France (French/3), Ethiopia (Amharic/2), Nepal (Nepali/2), Tunisia (Arabic/2), US (English/2), Afghanistan (Dari), Algeria (Arabic), Bangladesh (Bengali), Brazil (Portuguese), Czech (Czech), Ecuador (Spanish), Germany (German), India (Hindi), Israel (Hebrew), Mongolia (Mongolian), Netherlands (Dutch), Serbia (Serbian), Turkey (Turkish), UK (English)

Teaching method

In the two academic years previous to the time of study, the same course was taught in a traditional non-flipped way over two cycles (in which the presentation of new content was all delivered by the instructor in class, followed by interactive practice and homework, as elaborated below). There are 15 lessons (units) in the book covering about 30 grammar points and 400 words; a task-based teaching approach (or communicative method in a broader sense, see Littlewood's (2014) disambiguation of the two terms) was adopted by the author, so handling 15 real-life situations was also included in the course objectives. However, due to very limited class time (24 hours in total for each cycle), in each of the two previous cycles only 6 lessons could be delivered in the course, covering about 180 words, 18 grammar points and 6 real-life situations. As mentioned above, no Chinese character learning was required.

Before the semester in which we ran the current study, the content of the course book was remade as a MOOC and publicly released on www.coursera.org. The MOOC has 15 modules corresponding to the 15 lessons of the original paper book. Each module consists of: 1) new word lists; 2) course videos introducing the main content of that lesson (acting as the equivalent to an instructor's delivery in a face-to-face classroom; 3) auto-grading exercises, matched to the textbook exercises; 4) discussion forums in which learners can interact with the instructors and each other;[5] 5) text for downloading. The MOOC can be studied alone and has actually been followed by many subscribers around the world; in this current study, it formed the major out-of-class learning resource.

In the study, the baseline group were not told about the existence of the MOOC, and the course was taught in exactly the same way as in previous semesters, with some additional surveys and assessments required by the study. Baseline class procedures were as follows: 1) Learners started each module in the face-to-face class; 2) the teaching method was regular communicative teaching, with presentation of new content by the instructor and some practice and pair/group work involved, matching interactive activities in the experimental group; 3) each lesson lasted 180 minutes; 4) homework was completing written and oral exercises in the textbook; 5) a short in-class quiz at the start of the following lesson was used to assess grammar and vocabulary learned in the previous lesson or via homework; these quizzes accounted for a large portion of the student's final grade; 6) 6 lessons of the text book were covered by the end of the 16-week semester.

The experimental group was taught by the same instructor. Procedures were as follows: 1) learners were required to study the online module of the MOOC before the face-to-face class; 2) at the beginning of the face-to-face class (instead of after class for the baseline group), learners took a quiz based on module content, matching the quiz taken by the baseline group and accounting for the same percentage of their grade (for the purpose of ensuring the participants did study the online module in advance); 3) the following face-to-face class focused on the language practice and pair/group activities, as used in the baseline group; 4) the face-to-face class time lasted 135 minutes; 5) homework was completing the online module of the next lesson (i.e. step 1) for next lesson; 6) 9 units (lessons) of the text book were covered by the end of the 16-week semester.

[5]One may question the extra influence of interactions via online forums upon the result of the study. We argue that, compared to a pure MOOC, flipped learning provides opportunities of direct learner-learner and learner-instructor interaction in the face-to-face phase, which, in current technological conditions, are more efficient than interactions via online forums. In addition, the instructional setting in this study required learners to attend face-to-face classes right after they took the relevant online materials, which left no time for them to reflect on any online interactions, since the latter were not instant communication. Combining these two factors, we assume the potential interaction via forums had little marked influence upon amount of additional interactive exposure and would not therefore confound the results of the experiment. In post-hoc interviews, students noted the amount of this kind of online interactions had been negligible for them, and our tracking data suggested it was much less than when compared with those taking the MOOC alone.

A comparison of the teaching methods used for each group and their different timelines is shown by in the flowchart in figure 1 and in table 2 below; the assessments and data collection procedures are explained in more detail in the following section. The flowchart highlights the time saving and extra learning involved in the experimental flipped group; as noted, four weeks' extra class time (6 hours) was saved by this group, allowing three more lessons to be covered. Table 2 shows how time was used in the face-to-face class. The experimental group used less time in total to progress through the syllabus, while more class time was allocated for interactive groupwork as well as checking written exercises (which for the baseline group were homework). It should be noted that there was a "traditional" teaching session in the experimental class, though more briefly and with more practice than presentation while compared with the baseline class. As shown in table 2, for the experimental group, 35 mins were saved from explaining new words, 45 mins from grammar/sentence related instruction, then 20 mins added for group work and 15 mins added for in-class written exercise, resulted in a 45 mins shorter face-to-face session with more in-class practice time. (Like most Chinese universities, at the university where this study was conducted, all classes are delivered at 45-minute time segments, making it possible to use the saved 45 minutes from each lesson for the teaching of the next, which resulted in the delivery of 3 more lessons within an identical total time capacity for the experimental group.)

To rule out the possibility that the experimental group might go at a pace too fast for the learners, learners' perception of class pace was also investigated, as detailed later.

Placebo effect was also taken into consideration. Boot et. al.(2013) has indicated that an active control group (like the baseline group in the current study) may still be insufficient to rule out a placebo effect for the experimental group when the expectation for the treatment by the latter is significantly higher. However, in the current design, learners' attitude toward the teaching method and actual investment to learning are variables within the general framework (see RQ2), and their interaction with teaching method on learning outcomes was observed (see below). In short, the "placebo effect" for the experimental group (if defined as better attitude and greater learning effort) was accommodated into the design, and as we anticipate, would be welcomed by most instructors using flipped teaching if it does exist.

Figure 1. Teaching procedures (timeline) for both groups

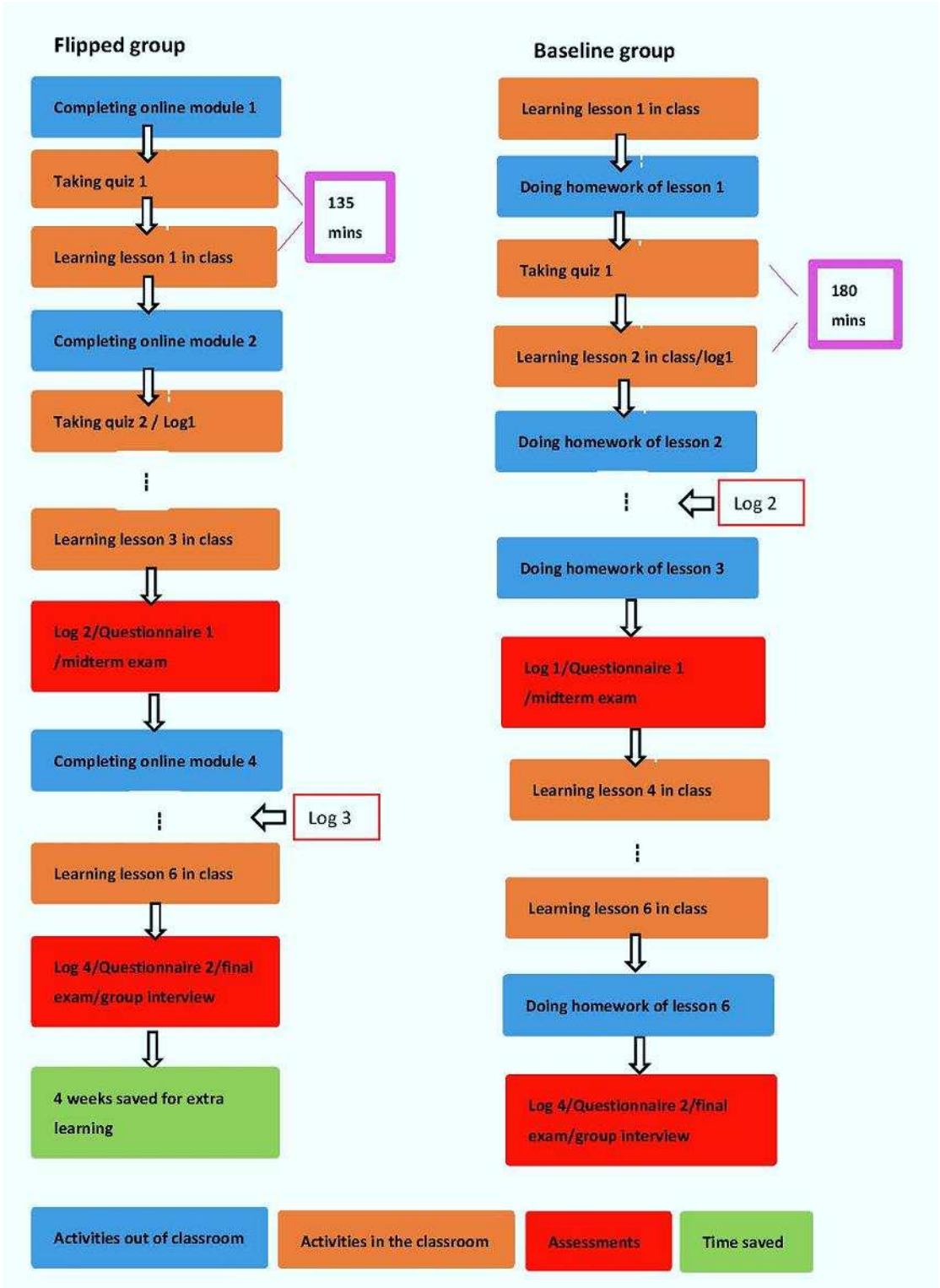

Flipped group
Completing online module 1
Taking quiz 1
Learning lesson 1 in class
135 mins
Completing online module 2
Taking quiz 2 / Log1
Learning lesson 3 in class
Log 2/Questionnaire 1 /midterm exam
Completing online module 4
Log 3
Learning lesson 6 in class
Log 4/Questionnaire 2/final exam/group interview
4 weeks saved for extra learning
Baseline group
Learning lesson 1 in class
Doing homework of lesson 1
Taking quiz 1
180 mins
Learning lesson 2 in class/log1
Doing homework of lesson 2
Log 2
Doing homework of lesson 3
Log 1/Questionnaire 1 /midterm exam
Learning lesson 4 in class
Learning lesson 6 in class
Doing homework of lesson 6
Log 4/Questionnaire 2/final exam/group interview
Activities out of classroom
Activities in the classroom
Assessments
Time saved

Table 2. Use of time in the face-to-face class per each unit (lesson)

The experiment group			The baseline group		
Time spent		Activities	Time spent		Activities
135' in total	15'	quiz (for new lesson)	**180' in total**	15'	quiz (for old lesson)
	10'	Students reading the conversation		35'	instructor explain the new words
	15'	drill practice of key sentences (syntactic points) led by instructor		20'	Students reading the conversation and answer questions
	15'	group practice of key sentences		30'	Instructor presenting the new key sentences
	45'	**group work such as situational (free) conversation, role play and language games**		20'	drill practice of key sentences (syntactic points) led by instructor
	15'	group representative reporting to the class		15'	group practice of key sentences
	20'	**completing (written) exercises in the textbook[6]** and a brief summary		**25'**	**group work such as situational (free) conversation, role play and language games**
				15'	group representative reporting to the class
				5'	a brief summary
After class	Flexible time	Learning the online module of next lesson (videos and auto-grading exercises)		Flexible time	Doing written and oral homework required by the book

Data collection: Assessments of oral proficiency

As shown by the red boxes in figure 1, there were two time points of collecting oral proficiency data from both groups. A mid-term exam was administered after completing lesson 3(Time 1). It consisted of two parts: a written test and an oral test, each accounting for 50% of the grade. The data generated by the former were kept for another study, while the latter is the main source of our analysis here. In the oral test, the learners were required to make a short presentation on a given topic (which had been covered in the previous classes, i.e. "Describing my activities on a weekday"). They had five minutes to prepare, and then the presentations were recorded with digital devices. The final exam was administered after completing lesson 6 (Time 2); again there were two parts identical to the mid-term exam, and the oral data were recorded for study (The topic is "Introducing myself and my family"). To ensure comparable data collection, the final exam occurred four weeks before the end of the semester for the experimental group, once the same amount of course content had been delivered to the two groups. After explaining the reason for this perceived early assessment, the learners all expressed understanding and acceptance.

[6]The auto-grading exercises are matched to the written exercises in the textbook, which means the flipped group had about 10 minutes in-class time to retake the auto-grading exercises for each lesson (see table 2). This part was designed to help the flipped learners consolidate the knowledge they absorbed in the online portion and also to eliminate the psychological effect that could be induced by "strange blanks in the book". It should be noted that the precondition of using only 10 minutes for doing these exercises by experimental group was that they were repetitive work for them. The baseline group, though spent less total out-of-class learning time (e.g. 82 vs 176 mins at T1, see table 6), should have spent much longer than 10 minutes to complete these exercises after class, since they were all "new" to them, and there was no other assignments than taking those exercises for them (watching course videos would have taken a lot of time for group 1).

For data analysis methods, there remains a debate within linguistic testing and teaching literature over the best way to test for evidence of development (e.g. Wright and Zhang in press). We therefore used a mixed-method model including subjective and objective ratings based on externally-derived measures to triangulate our data analyses, aiming for optimal external test validity (Cohen, Mannion, & Morrison, 2013). Since our institution's own assessment methodologies are not yet externally validated, we used high-stakes international oral proficiency scoring assessments to train our teachers in holistic subjective ratings, thus providing appropriate "cultural validity" for our learners and teachers as stakeholders interested in our outcomes (Cohen, Mannion, & Morrison, 2013: 194). However, subjective ratings may obscure the level of linguistic detail we were interested in, so we created a set of objective measures based within existing linguistic development frameworks (see below) to probe the data more specifically, as well as providing an opportunity to compare both subjective and objective outcomes to assure internal test validity (see also, e.g. Wright & Tavakoli, 2016).

Our objective measures investigated specific aspects of learner's oral performance at the two time points. Learners' speech recordings were collected from the two exams (61 samples in total across both classes at each time point), and were collated into 50-second lengths; this ensured reliability of analysis, since many of the individual clips were very short and disfluent.[7]

Data were analyzed using standard complexity, accuracy and fluency (CAF) measures (Skehan, 2009). Among the many ways of measuring CAF, we wished to identify indices that seemed most relevant to beginner-level oral Chinese. Some measures such as number of subordinate clauses as a measure of complexity or number of repairs to measure fluency (see e.g. Tavakoli, 2016), were excluded as we judged that these could hit floor effects or provide very limited indications of what our learners were able to achieve. We therefore identified the following 10 sub-indices as appropriate for our study for testing progress from Time 1 (mid/term/after completing three units) to Time 2 (after completing six units). For syntactic and lexical **Complexity**: words per AS-unit (the Analysis of Speech Unit), clauses per AS-unit, type-token ratio (Foster, Tonkyn, & Wigglesworth, 2000; Yuan,& Ellis, 2003). For Accuracy: syntactic errors per AS-unit, lexical errors per AS-unit, pronunciation (consonant/ vowel/tone) errors per AS-unit (Yuan,& Ellis, 2003, Robinson, 2007, Bygate, Swain, & Skehan, 2013). For **Fluency**: total syllables per 50 seconds, mean length of run (number of syllables between pauses), mean length of unfilled pauses, mean length of filled pauses (Towell, 1996;Tavakoli, 2016).

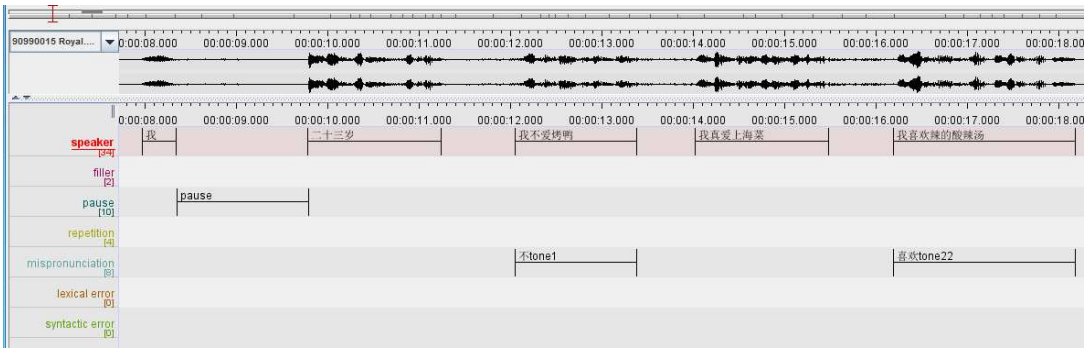
Using the software ELAN, we coded all the above indicators for each audio clip at Time 1 and Time 2. The coding was performed by two trained assistants; all coding judgments were cross-checked between the two assistants to reach consensus. Figure 2 below is an example of the coding. Each index was then transformed to a z-score for convenient comparison (both raw and z-scores will be presented below)[8]. The z-scores of the relevant sub-indices were also combined into a

[7]Typically fluency data are usually calculated either by per second or per minute rates, to ensure comparability. However, other papers justify other ratings, e.g. Du 2013 takes a section out of her recordings of speech samples to represent the point in the task when her speakers were speaking most effectively (i.e. not the early or finishing part of the task), and in another paper the authors present speech rate data rated over a 20-second sample (Wright and Zhang 2014, p73) as "a sufficiently clear length of run, not confounded by task process, which would be valid and reliable as a measure of speech rate for each participant across all four tasks at both times of assessment". For the current study, all oral data collected are longer than 50 seconds, but a few are shorter than 60. Therefore, to keep the authenticity of the data, we chose "per 50 seconds" instead of "per minute". Given that this study focused on between group comparison, we assume this will not affect the reliability of analysis. If researchers are to compare these data with data from other studies, we believe multiple our data up to minute-long will be an option.

[8]This method has not been used by any other studies in this field, however, it seemed to be the only way to form the 3 general objective

categorical average for each of the three general categories of complexity, accuracy and fluency, by using the following formulas: z-complexity = ("z-words per AS-unit" + "z-clauses per AS-unit" + "z-type-token ration")/3 ; z-accuracy = (- "z-syntactic errors per AS-unit"- "z-lexical errors per AS-unit" – "z-pronunciation errors per AS-unit")/3 ; z-fluency = ("z-total syllables per 50s" + "z-mean length of run between pauses" - "z-mean length of silent pauses" - "z-mean length of filled pauses")/4. Subtractions were used for negative indices of oral proficiency. General (objective) oral proficiency scores are shown as the mean value of the combined and standardized CAF scores.

Figure 2. An example of objective coding with ELAN



Subjective rating measures of oral proficiency were also employed to compare with and validate the objective data. Several well-known foreign language proficiency scales, including HSK, CEFR, ACTFL, TOEFL and IELTS were referred to and compared by the authors to form a criterion that fit the current study. Though HSKK (the oral part of HSK, i.e., Chinese Proficiency Test) seemed to be the best choice for its Chinese-specific nature, it was not selected for lacking the CAF related rubrics and comparatively shorter history of the current form (HSK official website, 2017). CEFR (Common European Framework of Reference for Languages) was rejected for the same reason. ACTFL (American Council on the Teaching of Foreign Languages) guidelines require continuous interaction between the tester and testee, which did not match the form of the test in the current design. Between the two remaining English-specific tests, TOEFL (Test of English as a Foreign Language) has rubrics that are more content related, leaving IELTS (International English Language Testing System) the only CAF-focused scale. Therefore, the band descriptors of the IELTS speaking test (IELTS website, 2016) were borrowed and simplified (see Appendix 1) to form a 4-category, 5-scale marking rubric. Two professional Chinese teachers who were naïve of the current study's aims were invited to mark all 61 recordings (randomized and double-blinded) using these descriptors to assess (1) fluency and coherence, (2) lexical resources, (3) grammatical range and accuracy, (4) pronunciation, using a 1-5 scale, which put the total score of the oral production within a 4-20 range. Inter-rater reliability is good (Interclass Correlation Coefficient was .777 at Time 1 and .898 at Time 2), therefore the mean values of scores given by the two raters for each sub-index were adopted for analysis.

indices of complexity, accuracy and fluency, and a general objective score for oral proficiency, which were comparable to subjective measures. We agree that the subjective rating is a more reliable indicator for general proficiency in this study, since it has been used by many studies as well as major foreign language oral proficiency tests, and was therefore used in the GLM analysis while answering RQ3. We believe for the purpose of comparing pedagogical effectiveness of different teaching methods, this way of generating a general objective score can triangulate the findings with subjective measures.

Assessment of learners' investment into learning and attitude toward/perception of the teaching methods

Since the new flipped method might cause fluctuation of learners' investment into the course (learning effort), which might in turn interact with the method itself to influence the learning outcome, we designed study logs to let the learners of both groups report their actual input into the course (given that objective MOOC usage reports from the MOOC platform could not feasibly be identified for our specified participants). Two slightly different sets of five questions were asked in the study logs for each group taking account of the different methods used in the two classes, although all items were kept comparable, following Hung (2015) (see appendix 2). Question 1 and 5 were about learning effort for time spent pre-class (for the experiment group) or post-class (for the baseline group), asked in the format of "how many minutes did you spend?" The mean of the sum of answer 1 and 5 was used to measure the "out-of-class study time" of learners; this score could be used to check for statistical interaction with each teaching method. Questions 2-4 were about the self-reported completion rates of different types of assignment, as detailed in the next section. The study logs were issued and collected at four evenly distributed time points throughout the semester (see figure 1). The data from log 1 and log 2 were averaged to reflect learning investment at midterm, and those from log 3 and log 4 were averaged as investment at the end of semester.

Learners' acceptance and perceptions of the teaching method they experienced were investigated in a separate survey with 12 5-scale Likert questions (see appendix 3) at both mid-term (T1) and semester end (T2).[9]Question 1 to 10 were adapted from Murray's (1983) study, tapping learners' perceptions toward specific aspects of each teaching method and their specific and general acceptance of the instruction. The last two questions in this part tapped learners' perception of course pace（item 11） and difficulty (item 12, as noted about potential differences arising from flipped learning at the end of "teaching method" section above). We recalculated the raw scores for these items (original reading -3) so that "0" was standardized as "just right"); the data could be used in further correlational analysis with other standardized scores. As above, the two versions for each group were slightly different to fit the context, although all the corresponding items were kept comparable. Two experts in the field of CFL were invited to review the questionnaire to ensure that the finalized version was a comprehensive assessment of the instruction from the learners' viewpoint and had good validity. Post-hoc Cronbach's Alpha of question 1 to 10 was .80 (T1) and .91 (T2) respectively, showing good reliability of the survey data, especially when learners became more familiar with their teaching methods. Qualitative data were also gathered to explore further insights from students over potential strengths (or weaknesses) of flipped teaching compared with traditional teaching, through two semi-structured group interviews after the final exam. Six representative learners (chosen by stratified sampling according to course grade) from each group were interviewed. Interview questions were listed in appendix 4.

The quantitative data collected by study logs and questionnaires were analyzed using statistical analyses. Group data were found to be normally distributed, so we proceeded using standard tests for between-group and within-group comparison, and associations with oral proficiency measures.

[9]We didn't assess learners' start-off motivation for the following reasons: as introduced in the "participants" section, all participants were beginner learners of Chinese and first year graduate students who arrived in China only about one week before the beginning of the study. Their placement into the two groups were randomized. We believe the randomization and similar background should have guaranteed their similar start-off motivation. In addition, a survey on motivation at the beginning of the instruction might raise participants' awareness of the experiment, hence skew the effect of the instruction (also see the last paragraph in "teaching method" section for the consideration of placebo effect).

Qualitative interview data are included in our discussion section to illustrate and provide context for the quantitative data.

Results

RQ1: Can flipped teaching based on a MOOC enhance the development of learners' oral proficiency in a college level Chinese L2 beginner's class, measured in better linguistic performance and faster rate of progress, compared to a traditional non-flipped classroom?

In terms of rate of progress, we showed in table 2above that on average the experimental group (group 1) spent 135 minutes for a unit (lesson) in the textbook, while the baseline group (group 2) spent 180 minutes. This difference enabled the former to learn 50% more content by the end of the semester. In terms of linguistic performance, we start with the subjective ratings across five IELTS-based measures (see table 3), which are intended here to provide a recognizable measure of development using holistic scoring methods drawn from international language testing paradigms, and thus fairly easily comparable to other studies tracking linguistic development using similar measures.

Independent sample t-tests at Time 1 (T1) showed a significant advantage for the experimental group in lexical resources (p=.000), a near-significant advantage in grammatical range and accuracy (p=.061), no significant difference in fluency and coherence (p=.126) and no significant difference in pronunciation (p=.234). At Time 2 (T2), a significant advantage for the experimental group remained in lexical resources (p=.000) and grammar (p=.000); fluency and coherence were now also significantly different (p=.000), while we found no difference in pronunciation (p=.382). The total oral scores for the experiment group were significantly better than the baseline group at both T1 (p=.011) and T2 (p=.000). These outcomes reflect the increased level of mean between-group difference by T2 for all indices except for pronunciation, and suggest a widespread advantage for the experimental flipped group (other than pronunciation).

Table 3. Independent sample T-tests for subjective oral proficiency indices between groups

Subjective oral proficiency indices	Group	Value at T1					Value at T2				
		Mean	SD	T	p	Mean between-group difference	Mean	SD	t	p	Mean between-group difference
Fluency and coherence	1	3.81	.61	1.556	.126	.21	4.02	.47	3.685	.000*	.43
	2	3.60	.40				3.58	.44			
Lexical resources	1	4.19	.54	3.798	.000*	.46	4.35	.39	5.243	.000*	.60
	2	3.73	.39				3.75	.50			
Grammatical range and accuracy	1	3.92	.56	1.906.	.061	.24	4.08	.34	4.754	.000*	.48
	2	3.68	.38				3.60	.44			
Pronunciation	1	3.92	.43	1.203	.234	.12	3.81	.48	.881	.382	.11
	2	3.80	.34				3.70	.47			

Total subjective oral proficiency score	1	15.84	1.85	2.651	.011*	1.02	16.26	1.40	4.349	.000*	1.62
	2	14.82	1.08				14.63	1.52			

**Note: significance: p<0.05 group1=experiment group group2=baseline group*

T1=values after completing lesson 3 T2=values after completing lesson 6

We hoped we could find parallel developments within the objective measures drawn from the task-based SLA literature within the complexity, accuracy, fluency framework (CAF). However, the objective measures revealed a different, more complicated picture, both in individual indices (Table 4) and in combined categorical scores (Table 5). As shown by t-test results in Table 4, the experimental group demonstrated an advantage in speech complexity and fluency at both times, but no advantage in accuracy. The experimental group significantly outperformed the baseline group in words per AS-unit (p=0.008), total syllables per 50 seconds (p=.000) and mean syllables between pauses (p=.000) at T1; they also had a near-significant advantage in clauses per AS-unit (p=.058) and (shorter) mean length of silent pauses (p=.084) at T1. The remaining indices were all insignificant. At T2, the advantage for words per AS-unit disappeared (p=.776), also for mean length of silent pauses (p=.707). Higher scores on total syllables per 50 seconds showed a trend toward significance (p=.095). Mean syllables between pauses remained highly significant (p=.000), and clauses per AS-unit remained near-significance (p=.066). Lexical complexity (type-toke ratio), approached significance (p=0.056).

While, in broad terms, the findings differed between subjective and objective measures, closer inspection in fact found more similarities, providing some internal validity between our approaches. Although the advantage for the experimental group fluctuated between indices over time, the baseline group outperformed the experimental group in none of the 10 objective indices, which was in line with subjective ratings. We also noted that total AS output and speaking time (not specifically analysed here), also favoured the experimental group; the total number of AS-units produced by the flipped group were significantly larger than the baseline group at both times (p=.010 and p=.000), and the time duration of the (whole) speech improved, though only as a trend towards significance (p=.131 to p=.070). As expected, these two indices were correlated to each other (r=.608 p=.000 at T1, r=.458 p=.000 at T2), and they might explain the gap between subjective and objective measurements, as we will discuss later.

Table 4. Independent sample T-test for objective oral proficiency indices (raw scores) between groups

Objective oral proficiency indices	Group	Value at T1					Value at T2				
		Mean	SD	T	p	Mean difference	Mean	SD	t	p	Mean difference
Words per AS-unit	1	4.54	.44	2.803	.008*	.58	6.50	1.39	.286	.776	.09
	2	3.96	1.05				6.41	1.06			
Clauses per AS-unit	1	1.06	.075	1.932	.058	.04	1.34	.30	1.889	.066	.11
	2	1.02	.095				1.22	.14			
Type-token ratio	1	.61	.055	.662	.511	.01	.49	.05	-1.953	.056	-.03
	2	.60	.075				.52	.06			

Syntactic errors per AS-unit	1	.11	.15	-.241	.810	-.01	.14	.14	-.767	.446	-.03
	2	.12	.18				.17	.12			
Lexical errors per AS-unit	1	.04	.07	-.394	.695	-.01	.05	.07	.516	.608	.01
	2	.05	.11				.04	.06			
Pronunciation errors per AS-unit	1	.55	.25	.664	.509	.05	.82	.54	-1.503	.138	-.23
	2	.50	.30				1.05	.64			
Total syllables of 50 seconds	1	48.61	11.47	5.810	.000*	14.55	66.51	14.50	1.694	.095	6.02
	2	34.07	7.64				60.50	13.17			
Mean syllables between pauses	1	3.55	1.06	5.348	.000*	1.18	4.01	.97	6.469	.000*	1.37
	2	2.37	.61				2.63	.67			
Mean length of silent pauses	1	1.35	.67	-1.756	.084	-.30	1.13	.45	-.378	.707	-.04
	2	1.65	.66				1.16	.27			
Mean length of filled pauses	1	.25	.24	-1.631	.108	-.10	.31	.35	-1.549	.128	-.12
	2	.36	.24				.43	.21			
Total number of AS-units	1	8.74	2.25	2.681	.010*	1.44	14.68	4.17	3.814	.000*	3.28
	2	7.30	1.93				11.40	2.31			
Time duration of speech	1	63.16	26.50	1.530	.131	10.49	125.58	52.16	1.845	.070	21.08
	2	52.67	27.08				104.50	35.11			

**Note: significance: p<0.05 group1=experiment group group2=baseline group*

To display the CAF data more clearly, and allow for association analysis with the subjective ratings as well as the learner study log data, the 10 objective indices were converted to standardized values and combined into categorical mean scores for Complexity, Accuracy and Fluency, as well as a mean score for general oral proficiency combining all three categories (see table 5).

These findings, summarizing individual indicators in table 4, confirm that the flipped group's advantage remained only on Fluency measures. This does not wholly match the subjective findings in Table 3, which showed that the advantage remained across almost all measures at T2. We suggest an explanation for this in our discussion. However, correlation coefficientsbetween general objective scores and total subjective scores are .422 (p=.000) and .520 (p=.000) at both times, showing medium correlation of the grading.

Table 5. Independent sample T-tests for combined oral proficiency indices (z-scores) between groups

Standardized objective oral proficiency indices (z-score)	Group	Value at T1					Value at T2				
		Mean	SD	t	p	Mean difference	Mean	SD	t	p	Mean difference
Complexity	1	.22	.45	2.579	.012*	.45	.01	.83	.107	.916	.02
	2	-.23	.85				-.01	.43			

Accuracy	1	.00	.51	-.014	.989	.00	.07	.58	.990	.326	.15
	2	.00	.81				-.08	.60			
Fluency	1	.39	.57	5.842	.000*	.79	.27	.78	3.129	.003*	.55
	2	-.40	.49				-.28	.56			
General objective oral proficiency z-score (mean of the above)	1	.20	.32	4.529	.000*	.41	.12	.36	2.521	.014*	.24
	2	-.21	.39				-.12	.38			

**Note: significance: p<0.05 group1=experiment group group2=baseline group*

Summing up the above findings, given the fact that general objective scores and total subjective scores both significantly favoured the flipped group, it was safe to say that by using flipped teaching in a beginner-level Chinese L2 classroom, learners were able to save 25% percent of class time (4 weeks out of 16), and were able to achieve better oral proficiency. A significant advantage was found in all sub-measurements (except for pronunciation) and at both times using subjective grading, as for objective measures, though on surface some of the 10 sub-indices were insignificant, the aggregated objective measure did prove that the flipped group had advantage in oral fluency and complexity. The advantage seemed less clear at Time 2 using objective grading.

RQ2: Are there differences in learners' perception of/attitude toward and investment in their learning comparing the two teaching methods (the flipped Chinese L2 class and the traditional class)?

We first looked at learning investment by both groups. Table 6showed that the experimental group spent significantly more time out of classroom (p=.011) at T1. The significance disappeared at T2, but it might be caused by a very large SD (141.31)[10]. The mean difference at T2 between groups was still very large (41.26 minutes). Overall, the flipped method led to an average 176.35 minutes study time out of class at T1 and 151.18 minutes at T2 per lesson, which are more than the in-class study time (135 minutes), while the baseline group only spent about half of their in-class study time out of class. The experimental group also demonstrated higher completion rate of online/in-book exercises (p=.007 and .033) and pre-class assignments (T2 only, p=.022). For both groups, most of the investment indices dropped at T2 compared to T1, but paired sample t-tests within each group showed that only the fall in pre-class assignment completion rates for the experimental group was significant (p=0.028). This might be a normal trend that could take place in any academic courses.

Table 6. Independent sample T-tests for learning investment between groups

Learning investment	Group	Value at T1					Value at T2				
		Mean	SD	t	p	Mean difference	Mean	SD	t	p	Mean difference
Time spent out of classroom	1	176.35	190.38	2.612	.011*	93.91	151.18	143.31	1.518	.138	41.26

[10]We believe the large SDs means that different learners show very different learning styles and pace in self-study when given the freedom of arranging their own learning, noted in the qualitative comments in post-hoc interviews. This is the major purpose of flipped instruction, because being able to cater to different learning style /aptitude is its advantage over traditional teaching. We assume the better learning outcomes of the experimental group were partially the result of the above situation. It also reminded us that in the traditional classroom, many learners are forced to follow a pace that does not fit their learning style, and hence may fail to achieve optimal learning outcomes.

per lesson (minutes)	2	82.43	50.89				109.92	47.97			
Completion rate of pre/post-class assignment (%)	1	87.26	17.49	.825	.413	4.34	83.55	18.00	2.350	.022*	11.33
	2	82.92	23.31				72.22	19.65			
Times of watching the videos/reading the texts	1	2.31	3.44	-.548	.586	-.87	1.89	1.17	-.935	.353	-.95
	2	3.18	8.11				2.83	5.51			
Completion rate of online/in-book exercises (%)	1	82.10	25.79	2.799	.007*	23.60	78.15	25.62	2.185	.033*	12.90
	2	58.50	38.58				65.25	20.04			

**Note: significance: p<0.05 group1=experiment group group2=baseline group*

As we introduced a new teaching method with rich forms, we assume that learners' attitudes toward its various perspectives should be different, therefore in the questionnaires (appendix 3), we asked nine questions regarding the various components of the course, and hope they can help us to understand the learners' acceptance of them(on a 1-5 scale). Question 10 asked about overall satisfaction of the learning experience, which was correlated with the mean value of the answers to question 1 to 9 (r=.339, p=.008 at T1, and r=.771, p=.000 at T2). However, in the current study there is no room for a thorough analysis of all these components (but they could be used in the future studies), and we don't believe the mean value of the 9 questions is a better indicator than an overall rating of the course, given that the weights of each component in forming the overall satisfaction is not clear. Therefore we used the overall rating (question 10) as the measure for learners' attitude due to its comprehensiveness. Independent sample t-tests revealed that at both times the experimental group showed greater acceptance of the method they received than the baseline group, though at T1 the difference was not significant (p=.077 and p=.049, see table 6). Paired sample t-tests showed no decline in acceptance from T1 to T2 for both groups (p=.586 and p=.601). Even for the baseline group, the lowest mean attitude score was at 3.87, meaning all learners had positive attitudes toward the teaching method, though the flipped method was obviously better accepted.

The survey also measured perceptions of course pace and content difficulty, standardized as z-scores (Table 7). At T1the experimental group perceived the course pace as slower (mean difference =- .93, p=.001), and course content much easier (mean difference =-1.02, p=.000) than did the baseline group. However, at T2, this gap disappeared (p=.116 for pace and p=.739 for content difficulty); the experimental group now felt the course pace was faster than the baseline group (mean difference=.35). Paired sample t-tests revealed that for the experimental group, both scores were significantly larger at T2 than T1 (p.006 for pace and p=.002 for difficulty, meaning faster and more difficult). For the baseline group, the perception of speed was significantly slower at T2 (p=.044 mean=.03 at T2, meaning learners felt the pace was just right) and the perception of difficulty was not significantly different (p=.130). The shift in perception of pace and difficulty among the experimental group is assumed to relate to less class time per lesson. Given the simple content early in the course, such as "greetings", "self-introduction", which could be easily learned out-of-class, in-class time-saving on content compared to practice magnified the advantage of flipped teaching. As course content and skills became more complex, less class time could become a challenge to learners in the flipped group, creating a weakened perception of any advantage, even though both subjective and

objective data showed that those learners did achieve both time-saving effect and better oral proficiency. In other words, it seems although time saving and better learning outcomes can be achieved at the same time, the trade-off between the two factors are highly possible and can affect learner perceptions of time-efficient progress. This effect will be further interpreted in the "discussion" section.

Table 7. Independent sample T-test for learners' perception of the course between groups

Learners' perception of the course	Group	Value at T1					Value at T2				
		Mean	SD	t	p	Mean difference	Mean	SD	t	p	Mean difference
Learners' acceptance of (attitude toward) course	1	4.42	.62	1.808	.077	.42	4.32	.79	2.011	.049*	.46
	2	4.00	1.11				3.87	.97			
Learners' perception of course pace (speed)	1	-.19	.95	-3.392	.001*	-.93	.39	.76	1.594	.116	.35
	2	.73	1.17				.03	.96			
Learners' perception of course difficulty	1	-.29	.74	-3.924	.000*	-1.02	.16	.69	-.335	.739	-0.7
	2	.73	1.23				.23	.97			

**Note: significance: p<0.05 group1=experiment group group2=baseline group*

RQ3: Are there interactions between teaching method, learners' attitudes/investment, and learning outcomes (oral proficiency)?

As explained at the end of our method section, the total subjective scores were adopted as the indicator of general oral proficiency of learners. Bi-variate correlation analysis found no significant correlation between time investment and oral proficiency at T1 (r=.024 p=.854) and T2 (r=.082 p=.531), nor were correlations found between learners' attitude and oral proficiency at T1 (r=.087 p=.503) and T2 (r=-.030 p=.816).

Total subjective scores were then brought into a generalized linear model as the dependent variable, with standardized investment and attitude scores as covariates. Teaching method was a fixed factor. Our 3-factor model showed a significant main effect of teaching method (B=1.098 p=0.007) and no significant main effect of time investment (B=-.147 p=.471) and attitude (B=.039 p=.844) on learners' oral proficiency at T1. After adding interaction between method and time investment, and interaction between method and attitude as potential predictors, a 5-factor model only showed a significant main effect of teaching method (B=1.332 p=.004), and no significant main effect of the other 4 factors on learning outcomes(see table 7). Similarly, at T2, in the 3-factor model, only teaching method was significant (B=1.771 p=.000). In the five-factor model (see table 8), again only teaching method showed significant main effect on learning outcomes. It seemed that the flipped teaching method alone contributed to better oral performance of the experiment group, while the role of learners' positive attitudes toward the teaching method, or their effort measured by time investment out of class, were less important.

Table 8. Generalized linear model analysis for main effects on oral proficiency

Factors may affect oral proficiency (Independent variables)	Dependent variables					
	General subjective oral proficiency at T1			General subjective oral proficiency at T2		
	B	SE	p	B	SE	p
Teaching method	1.332	.47	.004*	1.750	.40	.000*
Time investment (standardized)	-.894	.81	.238	.044	.64	.992
Learners' attitude (standardized)	.111	.23	.949	-.355	.27	.218
Method*Time investment	.806	.83	.333	-.095	.67	.888
Method*Learners' attitude	-.193	.46	.677	.219	.40	.583

**Note: significance: p<0.05*

Discussion

This mixed-methods empirical study seems to confirm the strength of using flipped learning in beginner-level foreign language classrooms, demonstrating clear evidence from faster and better development of oral proficiency development in L2 Mandarin among MOOC-supported students, compared to traditional non-flipped teaching. However, this conclusion masks some inconsistencies between the objective and subjective measures for proficiency development used here, which we now address. At both T1 and T2, the objective measures were correlated with subjective measures in general terms. Digging deeper, inconsistencies specifically existed in two aspects: grammatical accuracy (in which the subjective measures favored the experimental group, while the objective measures showed no significant difference) and the proficiency gap between the two groups tracked over time (the subjective measures showed a wider gap by T2, the objective measures showed a narrower gap at T2). Currently, comparative studies of holistic (subjective) versus objective oral proficiency measurement remain inconclusive - some researchers questioning the validity of holistic testing (e.g. Clark & Clifford 1988, Tarone 1987) - but to date no robust operationalization for using valid objective indicators to assess general L2 oral proficiency has been agreed, particularly within the CAF framework (see author 3 and collaborator). Halleck (1992, 1995) suggests that holistic measurements can reliably predict objectively-measured syntactic maturity, though less consistently at lower levels of proficiency where communicative factors may be more significant for even professionally trained raters for determining proficiency rating. In this study, subjective raters repeatedly used band descriptors relating to communication (i.e. "comprehension" and "no misunderstanding") as criteria for grammatical accuracy (appendix 1). We also noted that oral fluency, total speaking time and output of learners in the experimental group were significantly greater than the baseline group; this might have enhanced a tendency for the two raters to use communicative factors as proxies for syntactic proficiency. If so, this can explain the first inconsistency. The second inconsistency is to do with the changing gap in proficiency between both groups over time. This could be explained as argued above, that subjective grading of syntactic accuracy was influenced by an even greater fluency/communicative gap at T2; added to this, we found when answering RQ2 some evidence of a trade-off effect between time-saving and learning outcomes, which is perhaps reflected in the higher SDs in the subjective ratings at T2. Fortunately, for the purpose of this study it is not necessary to decide whether the subjective or objective measurement is a better instrument to assess learners' oral proficiency, given the fact that they were correlated and the

baseline group were not superior in any of the subjective or objective sub-indices.

We therefore feel confident in claiming that the flipped teaching method can enhance second language learners' oral proficiency while saving time spent in face-to-face classes. Learners exposed to this method also demonstrate more positive attitudes toward teaching method and more time spent on learning out of classroom, though these two factors (usually advised by instructors) were not direct causes of better learning outcomes here. In addition, the decreasing advantage in perception of course pace and difficulty for the experimental group reminds us of two possibilities: 1) as mentioned earlier, though flipped teaching could lead to both less time-use and better learning outcomes, if both were set as pedagogical target, a trade-off effect might take place, meaning the instructor and learners would need to choose their stance within a continuum between the two ends.In the current study, this possibility can be triangulated by the fact that advantage in objective oral proficiency (especially in complexity) also decreased while the perception of course pace increased for experimental group at T2; 2) considering there was no direct correlation between perceived course pace/difficulty and learning outcomes, it could also be a simpler situation that the perceived "freshness" of flipped teaching by learners decreased over time. This could be triangulated by the fact that though course content became richer and more challenging over time, experimental group spent less out-of-class learning time at T2 than T1, while the baseline group spent more time as expected, though both differences were insignificant (see table 6).If this is the case, then how to maintain learners' motivation while implementing the new method would become a critical task for the instructors. However, only when more studies on flipped teaching with a strict control of time spending become available can we find evidences for one or both of the two possibilities.

Another minor finding of this study which needs explanation is that flipped teaching may have different power over different aspects of second language oral proficiency. The objective measures showed that flipped teaching may benefit oral fluency most, with less influence on complexity and least on accuracy. In subjective ratings, the fact that learners exposed to flipped teaching enjoyed no advantage in the acquisition of L2 pronunciation could be evidence for the limited effectiveness of flipped teaching on specific linguistic components required for accuracy e.g. where phonology matters. We note that the teaching methods for both groups leaned toward a content/communication approach by design. As a result, flipped instruction magnified the advantages of better communicative performance as well as the disadvantages of limiting development of accuracy (in pronunciation especially) (see Richards & Rodgers, 2014). In fact, similar phenomena have been identified in immersion-style classrooms, right back to the communicative classes in Canada in the 1970s and 80s, which led to the huge Focus on Form/Focus on FormS debate.(See, e.g. Laufer, 2006), One important finding during that debate was that bilingual-style communicative learning doesn't necessarily trigger accuracy. We assume that the face-to-face instruction for the flipped group held similar features, as the learners have more flexibility and less control from the instructor, which resulted in better learner fluency and no advantage in accuracy. Another possible cause of this phenomenon could be a trade-off effect between accuracy and fluency. In Skehan (2009)'s "limited-capacity cognition model", he argued that induced by different task types, accuracy is traded off against fluency, despite, despite other claims that C, A and F, can operate in parallel (Robinson's 2007 Cognition Hypothesis) – see Awwad, Tavakoli, &Wright, 2017 for details.) In future research, the effects of implementing a flipped instruction focusing on accuracy should be tested.

We also wished to investigate the learning mechanisms by which flipped teaching method could be argued to be effective, by turning to active and collaborative learning theories. Group interview data showed that the learners in both groups viewed the instructor and teaching methods well, but learners in the flipped classroom perceived that the online part (MOOC) particularly enabled self-paced learning. We noted a high SD among the experimental group in the time spent on our-of-class activities (Table 6); we suggest this can be explained by the individual differences entailed by independent self-study in that different learners may choose to study at different rates when arranging their own learning, which was noted in the qualitative data. Learners commented "it worked well", "it was the most interesting part", "it allows me to put in the amount of work I want to put, it didn't restrict me by having to wait for other people", "in that class (with a lot of students) the teacher-student time is so limited, so if we could have a teacher online, and we go back to that again, actually it could be the same thing as a 1-on-1 session."

Learners' comments on the face-to-face classroom (also collected by the group interview) were similar for both groups, since a certain amount (though less) of interactive teaching was kept for the baseline group. Learners commented "because there is not enough chance for us to really talk to the local people, the professor letting us talk in situ is really the way to learn", "it boosted our confidence", "I like the fact that it was very relaxed, the mood and everything is very nice", "it was like a comfort zone because it is very difficult to interact with people in the street, they don't understand us even if we speak Chinese, so our classroom is like a comfort zone to practice our speaking." But learners in the baseline group also pointed out the insufficiency of interactive practice: "One thing I found is that lots of guys in our class feel that time is limited, and you (the instructor) cannot focus on everyone."

On the other hand, only learners in the flipped classroom mentioned the teacher-guided practice characteristic of the class: "it was kind of revision work, it was like we have already memorized the words, and the teacher was there to help us revise them, and people always like to tell something so confidently".

The above comments fit well with the claims of active learning theory, i.e. it allows the learners to acquire low-level content knowledge outside of classroom and make full use of face-to-face class time to develop the high-level analysis and interaction skills under the teacher's guidance. However, we also see the value of the collaborative approach, embedded in Socio-Cultural Theory, where language is seen as an outcome of social communicative activity (Thorne, & Lantolf, 2007). In this framework, grammar as well as vocabulary are both learnable by the individual through mediated (communicative) language experience, not just between teacher and learner, but also peer-to-peer dialogue (Swain et al., 2002). As we found, a flipped L2 classroom can provide learners with more time and occasions for meaningful communication, within which more scaffolding from the teacher, and also between peers could be expected to take place. In turn, more productive learning can occur in the Zone of Proximal Development (Vygotsky, 1978, p85). The "comfort zone" metaphor made by the learners reminds us of this feature particularly.

We therefore regard MOOCs and the flipped classroom approach developed on a MOOC-base as important new tools for mediating language learning in this technological age. Since the current and previous studies have proven their effectiveness, at least in certain aspects of language teaching/learning, their time saving characteristic, and their capability of reaching many people while catering to the learning habits of different individuals, it seems we have no reason not to apply,

promote and develop these tools with a careful but positive attitude.

Conclusions and limitations

The current study found that flipped instruction enhanced adult learners' L2 oral proficiency while saving face-to-face class time, with useful implications how to optimize instructional goals, balancing individual learner needs with efficiencies of class-based interaction. This result is also interesting within the relatively understudied area of L2 Mandarin, and even at beginner level, particularly for checking the validity of current measures of proficiency, which have often been based on L2 English. We therefore believe the findings could also be applied to foreign language classrooms of other languages, particularly if designed for college-age learners, who can be assumed to have developed the self-management strategies required by flipped instruction. We see that research on flipped instruction in secondary education is emerging, which will be useful in extending our understanding of the socio-cultural learning mechanisms we assume are involved in successful flipped learning.

There were some limits in this research: firstly, its focus on beginner learners, though clearly justified, means we require further exploration of the potential of using flipped methods on oral proficiency at higher levels, as does the developmental trajectory. Secondly, other language skills adopted in the classroom, such as reading, writing and listening must be related to oral skills in some way, but were not investigated in this study. Finally specific linguistic elements of vocabulary, explicit/implicit syntactic knowledge implicated in our measures of complexity, accuracy and fluency were not measured in detail. Full understanding of using flipped instruction for foreign languages cannot be acquired unless these issues are uncovered. Despite these limitations, the findings of this study support the further implementation of flipped teaching in foreign language classrooms.

References

Anderson, J. R. (1988). Acquisition of cognitive skill. Readings in Cognitive Science, 89(4), 362-380.

Awwad, A., Tavakoli, P., & Wright, C. (2017). "i think that's what he's doing": effects of intentional reasoning on second language (l2) speech performance. System, 67, 158-169.

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. The Modern Language Journal, 70(4), 380-390.

Basal, A. (2015). The implementation of a flipped classroom in foreign language teaching. Turkish Online Journal of Distance Education, 16(4), 28-37.

Bejarano, Y. (1987). A cooperative small-group methodology in the language classroom. TESOL Quarterly, 21(3), 483-504.

Bishop, J. L., & Verleger, M. A. (2013). The flipped classroom: A survey of the research. ASEE National Conference Proceedings, Atlanta, GA, 30(9), 1-18.

Bloom, B. S. (1964). Taxonomy Of Educational Objectives (Vol. 2). New York: Longmans, Green.

Bonwell, C. C., & Eison, J. A. (1991). Active learning: Creating excitement in the classroom. ASHE-ERIC Higher Education Reports. ERIC Clearinghouse on Higher Education,

Boot, W. R., Simons, D. J., Stothart, C., & Stutts, C. (2013). The pervasive problem with placebos in psychology: why active control groups are not sufficient to rule out placebo effects. Perspectives on Psychological Science A Journal of the Association for Psychological Science, 8(4), 445.

Bygate, M., Swain, M., & Skehan, P. (2013). Researching pedagogic tasks: Second language learning, teaching, and testing. Routledge.

Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. Testing. Washington, DC: Center for Applied Linguistics.

Chen Hsieh, J. S., Wu, W. C. V., & Marek, M. W. (2016). Using the flipped classroom to enhance EFL learning. Computer Assisted Language Learning, 1-25.

Clark, J. L., & Clifford, R. T. (1988). The FSI/ILR/ACTFL proficiency scales and testing techniques. Studies in Second Language Acquisition, 10(02), 129-147.

Cohen, L., Manion, L.,& Morrison, K. (2013). Research Methods in Education, 7th Edition.

Cotterall, S. (1995). Readiness for autonomy: Investigating learner beliefs. System, 23(2), 195-205.

Everson, M. E. (2011). Best practices in teaching logographic and non-Roman writing systems to L2 learners. Annual Review of Applied Linguistics,31, 249-274.

Flumerfelt, S.,& Green, G. (2013). Using lean in the flipped classroom for at risk students. Educational Technology & Society, 16(1), 356-366.

Forsey, M., Low, M., & Glance, D. (2013). Flipping the sociology classroom: towards a practice of online pedagogy. The Australian Sociological Association, 49(4), 471-485.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. Applied Linguistics, 21(3), 354-375.

Furneaux, C., Wright, C, & Wilding, E. (2015) First experiences of MOOCS in the UK; the case of the University of Reading's "Beginners' Guide to writing in English for university study". In J.R Corbeil, M.E. Corbeil, and B.H Khan (Eds.) The MOOC Case Book: Case Studies in MOOC Design, Development and Implementation. New York: Linus Publications. pp337-348. ISBN 13: 978-1-60797-561-8

Gan, Z. (2004). Attitudes and strategies as predictors of self- directed language learning in an EFL context. International Journal of Applied Linguistics, 14(3), 389-411.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. Statistics in Medicine, 27(15), 2865-2873.

Ghaith, G. M., & Yaghi, H. M. (1998). Effect of cooperative learning on the acquisition of second language rules and mechanics. System, 26(2), 223-234.

Gunderson, B., & Johnson, D. (1980). Building positive attitudes by using cooperative learning groups. Foreign Language Annals, 13(1), 39-46.

Halleck, G. B. (1992). The oral proficiency interview: Discrete point test or a measure of communicative language ability? Foreign Language Annals,25(3), 227-231.

Halleck, G. B. (1995). Assessing oral proficiency: A comparison of holistic and objective measures. The Modern Language Journal, 79(2), 223-234.

HSK official website (2017) Introduction to HSKK. Retrieved on 11 Aug. 2017 from www.chinesetest.cn.

Hung, H. T. (2015). Flipping the classroom for English language learners to foster active learning. Computer Assisted Language Learning, 28(1), 81-96.

IELTS website (2016) IELTS Speaking Band Descriptor. Retrieved on 12 Dec 2016 from www.ielts.org

Johnson, D. W., & Johnson, R. T. (1987). Learning together and alone: Cooperative, competitive, and individualistic learning . Prentice-Hall, Inc.

Johnson, D. W., Johnson, R. T., & Stanne, M. B. (2000). Cooperative learning methods: A meta-analysis. Minneapolis: University of Minnesota.

Ke, C., Wen, X., & Kotenbuetel, C. (2001). Report on the 2000 CLTA articulation project. Journal-Chinese Language Teachers Association, 36(3), 25-60.

King, A. (1993). From sage on the stage to guide on the side. College teaching, 41(1), 30-35.

Lage, M. J., Platt, G. J., & Treglia, M. (2000). Inverting the classroom: A gateway to creating an inclusive learning

environment. The Journal of Economic Education, 31(1), 30-43.

Lantolf, J. P., & Thorne, S. L. (2006). Sociocultural theory and the genesis of L2 development.

Lantolf, J. P. (2011). The sociocultural approach to second language acquisition. Alternative approaches to second language acquisition, 24-47.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. Language Testing, 16(1), 33-51.

Laufer, B. (2006). Comparing focus on form and focus on forms in second-language vocabulary learning. Canadian Modern Language Review, 63(1), 149-166.

Littlewood. (2014). Communication-oriented language teaching: where are we now? where do we go from here? (i). Language Teaching, 47(8), 32-36.

MacLeod, H., Haywood, J., Woodgate, A. (2015) Emerging patterns in MOOCs: Learners, course designs and directions, TechTrends59(1, 1), 56-63

McDonald, K., & Smith, C. M. (2013). The flipped classroom for professional development: part I. Benefits and strategies. The Journal of Continuing Education in Nursing, 44(10), 437-438.

McDonough, K. (2004). Learner-learner interaction during pair and small group activities in a Thai EFL context. System, 32(2), 207-224.

Mehring, J. (2016). Present Research on the Flipped Classroom and Potential Tools for the EFL Classroom. Computers in the Schools, 33(1), 1-10.

Meyers, C., & Jones, T. B. (1993). Promoting Active Learning. Strategies for the College Classroom. San Francisco, CA: Jossey-Bass.

Michael, J. (2006). Where9s the evidence that active learning works?.Advances in physiology education, 30(4), 159-167.

Murray, H. G. (1983). Low-inference classroom teaching behaviors and student ratings of college teaching effectiveness. Journal of Educational Psychology, 75(1), 138.

Ohta, A. S. (2000). Rethinking interaction in SLA: Developmentally appropriate assistance in the zone of proximal development and the acquisition of L2 grammar. Sociocultural theory and second language learning, 4, 51-78.

Ohta, A. S. (2001). Second language acquisition processes in the classroom: Learning Japanese. Routledge.

Piaget, J. (1968). Six psychological studies. (A. Tenzer Trans.). London: University of London Press.

Prefume, Y. E. (2015). Exploring a flipped classroom approach in a Japanese language classroom: a mixed methods study (Doctoral dissertation).

Prince, M. (2004). Does active learning work? A review of the research. Journal of engineering education, 93(3), 223-231.

Richards, J. C., & Rodgers, T. S. (2014). Approaches and methods in language teaching. Cambridge: Cambridge University Press.

Robinson, P. (2007). Re-thinking-for-speaking and L2 task demands: The Cognition Hypothesis, task classification, and sequencing. In Plenary address at the Second International Conference on Task-based Language Teaching, Hawaii.

Salomon, G. (1983). The differential investment of mental effort in learning from different sources. Educational Psychologist, 18(1), 42-50.

Salomon, G. (1984). Television is" easy" and print is" tough": The differential investment of mental effort in learning as a function of perceptions and attributions. Journal of educational psychology, 76(4), 647.

Shaaban, K., &Ghaith, G. (2005). The theoretical relevance and efficacy of using cooperative learning in the ESL/EFL classroom. TESL Reporter, 38(2), 14-28.

Shen, H. H., & Xu, W. (2015). Active Learning: Qualitative inquiries into vocabulary instruction in Chinese L2 classrooms. Foreign Language Annals,48(1), 82-99.

Shohamy, E. (1988). A proposed framework for testing the oral language of second/foreign language learners. Studies in Second Language Acquisition,10(02), 165-179.

Shokouhi, H., & Alishaei, Z. (2009). Proficiency and collaborative learning. Indian journal of applied linguistics, 35(2), 129-141.

Skehan, P. (2009). Models of speaking and the assessment of second language proficiency. In A. Benati (Ed.) Issues in Second Language Proficiency. London: Continuum. pp 203-215.

Smit, K., Brabander, C. J., & Martens, R. L. (2014). Student-centered and teacher centered learning environment in pre-vocational secondary education: psychological needs, and motivation. Scandinavian Journal of Educational Research,58(6), 695-712.

Strayer, J. F. (2012). How learning in an inverted classroom influences cooperation, innovation and task orientation. Learning Environment Research, 15, 171-193.

Swain, M. (2006). Languaging, agency and collaboration in advanced second language proficiency. In H. Byrnes (Ed.) Advanced language learning: The contribution of Halliday and Vygotsky. London: Continuum. pp 95-108.

Swain, M., Brooks, L.,& Tocalli-Beller, A. (2002). Peer-peer dialogue as a means of second language learning. Annual Review of Applied Linguistics,22, 171-185.

Szostek, C. (1994). Assessing the effects of cooperative learning in an honors foreign language classroom. Foreign Language Annals, 27(2), 252-261.

Tarone, E. (1987). Methodologies for studying variability in second language acquisition. Second Language Acquisition in Context, 35-46.

Thorne, S. L., & Lantolf, J. P. (2007). A linguistics of communicative activity. Disinventing and Reconstituting Languages, 62, 170-195.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. Applied Linguistics, 17(1), 84-119.

Tune, J., Sturek, M., & Basile, D. (2013). Flipped classroom model improves graduate student performance in cardiovascular, respiratory, and renal physiology. Advances in Physiology Education, 37(4), 316-320.

Vardi, M. Y. (2012). Will MOOCs destroy academia? Communication ACM,55(11), 5.

Vygotsky, L. (1978). Interaction between learning and development. In M. Gauvain & M. Cole (Eds.) Readings on the Development of Children. New York: Scientific American Books. pp 34-41.

Vygotsky, L. S. (1987). The collected works of LS Vygotsky: Vol. 1, Problems of general psychology. R.W.Rieber & A.S. Carton (Eds.) (N. Minick, Trans). New York: Springer.

Whong, M. (2013). A linguistic perspective on communicative language teaching. Language Learning Journal14(3),115-128.

Wilson, S. G. (2013). The flipped class: a method to address the challenges of an undergraduate statistics course. Teaching of Psychology, 40(3), 193-199.

Wright, C., & Tavakoli, P. (2016). New directions and developments in defining, analyzing and measuring l2 speech fluency. International Review of Applied Linguistics in Language Teaching, 54(2), 73-77.

Wright, C., & Zhang C. (2014). Examining the effects of Study Abroad on L2 Chinese development among UK university learners. Newcastle & Northumbria Working Papers in Linguistics 20: 67-83

Yarbro, J., Arfstrom, K. M., McKnight, K., & McKnight, P. (2014). Extension of a review of flipped learning. Retrieved from

http://flippedlearning.org/cms/lib07/VA01923112/Centricity/Domain/41/Extension%20of%20Flipped%20Learning%20Lit%20Review%20June%202014.pdf.

Yuan, F., & Ellis, R. (2003). The effects of pre- task planning and on- Line planning on fluency, complexity and accuracy in L2 monologic oral production. Applied Linguistics, 24(1), 1-27.

Appendix 1 Band descriptors for subjective measuring of oral speech

Band	**Fluency and coherence**	**Lexical resource**	**Grammatical range and accuracy**	**Pronunciation**
5	•is willing to speak at length, though may lose coherence at times due to occasional repetition, self-correction or hesitation •uses a range of connectives and discourse markers but not always appropriately	•has a wide enough vocabulary to discuss topics at length and make meaning clear in spite of inappropriacies	•uses a mix of simple and complex structures, but with limited flexibility •may make frequent mistakes with complex structures though these rarely cause comprehension problems	•uses a range of pronunciation features with mixed control •shows some effective use of features but this is not sustained •can generally be understood throughout, though mispronunciation of individual words or sounds reduces clarity at times
4	•usually maintains flow of speech but uses repetition, self -correction and/or slow speech to keep going •may over-use certain connectives and discourse markers •produces simple speech fluently, but more complex communication causes fluency problems	•manages to talk about the topics but uses vocabulary with limited flexibility	•produces basic sentence forms with reasonable accuracy •uses a limited range of more complex structures, but these usually contain errors and may cause some comprehension problems	•shows all the positive features of Band 3 and some, but not all, of the positive features of Band 5
3	•cannot speak without noticeable pauses and may speak slowly, with frequent repetition and self-correction •links basic sentences but with repetitious use of simple connectives and some breakdowns in coherence	•can only convey basic meaning on the topics and makes frequent errors in word choice	•produces basic sentence forms and some correct simple sentences but subordinate structures are rare •errors are frequent and may lead to misunderstanding	•uses a limited range of pronunciation features •attempts to control features but lapses are frequent •mispronunciations are frequent and cause some difficulty for the listener
2	•speaks with long pauses	•uses simple vocabulary	•attempts basic sentence	•shows some of the

	•has limited ability to link simple sentences •is frequently unable to convey basic message	to convey personal information •has insufficient vocabulary for the topics	forms but with limited success •makes numerous errors	features of Band 1 and some, but not all, of the positive features of Band 3
1	•pauses lengthily before most words •little communication possible	•only produces isolated words or memorized utterances	•cannot produce basic sentence forms	•Speech is often unintelligible

Appendix 2 Study logs*

Study LogA1 (for experiment group)

Date: Name:

Please answer the questions briefly according to your situation in recent weeks. Note **this is not part of your grade**, honest answers will do nothing but help us to improve our teaching in the future, therefore answer them as **accurately** (some of the questions need estimation) **and honestly** as you can. You only need to give **numbers or percentages** for each question.

(1) How many minutes did you study for each lesson (including videos, online exercises and extra time you spend to prepare for the quiz, which is what we call pre-class assignment) before the class?
(2) Did you always complete your pre-class assignments? What is the estimated percentage of you completing them (in %)?
(3) How many times did you watch the videos that go with the lesson by average? (If you didn't watch, put 0.)
(4) Did you always complete the online exercise? Put the estimated percentage of you completing them (in %)?
(5) How many extra minutes did you spend to prepare for the quiz before each lesson. (If you spend no more time other than watching the videos and taking the online exercises, put 0.)

Study Log B1 (for baseline group)

Date: Name:

Please answer the questions briefly according to your situation in recent weeks. Note **this is not part of your grade**, honest answers will do nothing but help us to improve our teaching in the future, therefore answer them as **accurately** (some of the questions need estimation) **and honestly** as you can. You only need to give **numbers or percentages** for each question.

(1) How many minutes by average did you use for studying Chinese **after completing each lesson** (including written assignment, oral assignment, time you spend to prepare for the quiz and extra time you spend to study Chinese by yourself in any forms)?

*There are two different versions for the two groups issued for four times each. They are marked as A1, B1-A4, B4 respectively. Study logsother than A1 and B1 are omitted here.

(2) Did you always complete your post-class assignments? What is the estimated percentage of you completing them (in %)?

(3) How many times did you read/review the texts of each lesson **after class** by average? (If you didn't, put 0.)

(4) Did you always complete the written exercises in the book? Put the estimated percentage of you completing them (in %)?

(5) How many extra minutes did you spend to prepare for the quiz (arranged in next class) after each lesson. (If you spend no more time other than completing the written and oral assignments, put 0.)

Appendix 3 Questionnaires*

Questionnaire A1 (for experiment group)

Name: Date:

This questionnaire asks about your satisfaction with the current teaching/leaning approach. There is no right/wrong criteria for the answers. It is important that you answer each question as honestly as you can, so as to help future learners.

The numbers alongside each number stand for the following response.

1—this item is never or only rarely true of me

2—this item is sometimes true of me

3—this item is true of me about half the time

4—this item is frequently true of me

5—this item is always or almost always true of me

Do not spend a long time on each item: your first reaction is probably the best one. Please answer each item. Do not worry about projecting a good image. Your answers are CONFIDENTIAL.

	Circle your answer
1. I am satisfied with the way/order that the course content is arranged/delivered.	1 2 3 4 5
2. I am satisfied with the way that quizzes are designed and delivered.	1 2 3 4 5
3. I am satisfied with the way that course videos are used for instruction. **("the way that the instructor delivers/presents the content of the course" here for the baseline group).**	1 2 3 4 5
4. I am satisfied with the way that online exercises are used for instruction. **("the way that post-class assignments are used for instruction" here for the baseline group).**	1 2 3 4 5
5. I am satisfied with the way that group/pair discussion are used for instruction.	1 2 3 4 5
6. I am satisfied with the clarity of the course (everything is explained clearly).	1 2 3 4 5
7. I am satisfied with the way/degree of interaction of the course.	1 2 3 4 5
8. I am satisfied with the task orientation of the course (always focus on the right topic).	1 2 3 4 5
9. I am satisfied with the organization of the course. (always feel guided)	1 2 3 4 5

*There are two different versions for the two groups issued for twice each. They are marked as A1, B1, A2 and B2 respectively. Questionnaire A2, B1 and B2 are omitted here.

10. Overall, I am satisfied with the learning experience in this course.	1 2 3 4 5
Finally, 2 questions that should be answered in a special format:	
11. My feeling about the pace of the course. **(put 1 if too slow, 5 if too fast, 3 if just right)**	1 2 3 4 5
12. My feeling about the difficulty of the course. **(put 1 if too easy, 5 if too hard, 3 if just right)**	1 2 3 4 5

Appendix 4 Group interview questions

(1) What do you like and dislike about this learning experience?

(2) Do you perceive anything special about the teaching method of this course?

(3) Do you like the out-of-class learning approach used in this course (interviewer should go specific in this according to different groups), why?

(4) Do you like the instruction and interactive activities in the face-to-face classroom (interviewer should go specific to help the interviewee understand what are referred to)? Why?

(5) Do you think you were fully motivated in learning this course? Why?