A Longitudinal Analysis of Ability Grouping with College EFL Learners

Chiu-hui (Vivian) Wu
Wenzao Ursuline University of Languages, Kaohsiung, Taiwan

Chia-jung Tsai
Wenzao Ursuline University of Languages, Kaohsiung, Taiwan

Yi-Min Chiu
Wenzao Ursuline University of Languages, National Cheng Kung University, Taiwan

Abstract

Ability grouping, organizing classes homogeneously by L2 proficiency, has been commonly used in Taiwanese English as a Foreign Language (EFL) classes. This quasi-experimental (within-subjects design) study examined proficiency gains of 785 Taiwanese university students over three years enrolled in a general English (GE) program that employed ability grouping. The standardized test used for this study was the *College Students English Proficiency Test* (CSEPT). The results indicated students gained in English proficiency over time, from entry into the program to their last year of English instruction. Further post hoc analysis of the long-term proficiency changes showed that students with an observed A2 (CEFR) proficiency, upon entry, had more pronounced gains, over the three years, than their A1 and B1 counterparts. The study concluded that a leveled English curriculum maximized the learning experience for A2 level students and allowed them continuous proficiency gains. However, the fact that B1 level students did not show consistent progress is perhaps due to plateau effect when their test scores hit the graduation benchmark. As for the A1 students, their lack of achievement may be due to their low self-esteem. The pedagogical implication suggests the need to revisit the leveled (ability grouping) English curriculum for A1 and B1 level learners.

Keywords: ability grouping, general English curriculum, graduation benchmark, college English, language program design

Introduction

Over the past two decades, Taiwan has attempted to increase its international participation in the global market by prioritizing English language education through a national development plan (Chen & Hsieh, 2011). It is believed that increased English proficiency of Taiwanese citizens would give them greater opportunity to participate in international affairs (Chen, 2011). Thus, the status of English in Taiwan has shifted from being a foreign language to being a quasi-official language, which is illustrated by the fact that signs in English are used in many public places (Chen, 2011; Feng, 2012). In addition, to better prepare citizens and students for English proficiency—and hence internationalization—the Ministry of Education (MOE) in Taiwan has initiated and implemented policies for English curriculum reform. For instance, MOE recently made EFL courses compulsory starting at primary school (grades 3 to 6) rather than secondary school (grades 7-12), (Chen & Hsieh, 2011; Chern, 2002, 2010).

E-mail: joycecjtsai@hotmail.com, Tel.: 886-922452332, Address: Center for English Language Teaching, Wenzao Ursuline University of Languages, 900 Mintsu 1st R. Kaohsiung 807, Taiwan R.O.C.

Regarding tertiary or higher education, the MOE has set expectations for English curricula and language policies for some time (Hua & Beverton, 2013; Pan & Newfields, 2012). In the 1990s, the MOE began to move universities away from exclusively reading in English curricula to more comprehensive learning plans (e.g., emphasis on production). In 2003, the MOE further urged universities and colleges to set English graduation benchmarks and left the choice of how the benchmarks would be measured (e.g., standardized testing) to the universities (Pan & Newfields, 2012). In today's universities, GE programs are how universities prepare their students to reach the English proficiency benchmark prior to their graduation.

To better understand the GE programs implemented across Taiwan, Chern (2010) studied the programs at 60 universities, including public and private. Findings from the study revealed that 32 universities required students to take a one-year English course during their freshmen year, approximately four credit hours, and 20 universities required students to take English courses for two years with a total of four to six credit hours. Based on the findings from diverse GE programs implemented in Taiwan, Chern (2010) concluded that a systematic examination was needed to determine if the curriculum prepared students to meet English language proficiency benchmarks.

One widely utilized approach believed to be effective at the tertiary level of EFL education was ability grouping. When this strategy is implemented, students are placed in different levels of GE groups based on their English proficiency. Research in EFL contexts on the effects of ability grouping found positive results for college freshmen (e.g., Khazaeenezhad, Barati, & Jafarzade, 2012; Kulik, 1992; Liu, 2008). Kulik (1992) contended that it would be a mistake if schools abolished ability grouping. Yet, the controversy and debate regarding its effectiveness has continued.

This current study was conducted to investigate how students of different English proficiency upon entry would progress over three years of ability grouping instruction. Because of university policies, a control group and/or other groups receiving non-ability grouping instruction was not possible. This lack of an experimental design meant that the findings of the current study could not facilitate claim a direct casualty between ability grouping and proficiency gains. The rationale for this current study, instead, was grounded in a desire to understand how ability grouping could have influenced the proficiency gains of students of different proficiency levels upon entry. In other words, this study contributes to an ongoing dialogue while also being aimed at inspiring future research that could address its unavoidable design limitations.

Literature Review

Ability grouping in language education

Ability grouping refers to the practice of placing students in a classroom or small groups based on ability or achievement. This is usually done by assessment of ability with standardized tests (Kim, 2012). This teaching and program design strategy has been used in education, especially in primary and secondary schools, since the 20th Century (Slavin, 1987). The earliest reviews regarding ability grouping were found in the 1920s and early 1930s (Kulik, 1992).

Subsumed under ability grouping are two types: (1) within-class and, (2) between-class grouping (Ireson & Hallam, 2001). Within-class grouping, or mastery learning, is usually practiced in a class, and students of different perceived levels are assigned to groups for specific or adaptive instruction to accommodate their learning needs (Ireson & Hallam, 2001). Between-class grouping, by comparison, is a school-level practice that places students in different ability groups or tracks by class (Ireson & Hallam, 2001). The current study specifically looked at the between-class grouping model for ability grouping with respect to EFL students.

Between-class grouping model for ability grouping

Ability grouping has been widely adopted in pre-secondary and secondary English language education in several countries such as the UK (Hallam & Ireson, 2003; Ireson, Hallam, Hack, Clark, & Plewis, 2002), and the USA (Slavin, 1990), and Korea (Jung, 2000; Kim, 2012). Previous research, nevertheless, has yielded divergent results regarding the effect of ability grouping on English proficiency (L2) attainment.

It has been evident in some survey-based studies with teachers that ability grouping facilitated English language teaching and learning, yet it catered more to the needs of higher-level learners than to those of their lower-level counterparts (Hallam & Ireson, 2003; Ireson et al., 2002). Hallam and Ireson (2003), for example, researched secondary school teachers' attitudes toward and beliefs about ability grouping. Their sample was comprised of more than 1,500 teachers from 45 schools in the UK. They found overall agreement among the teachers that ability grouping ensured maximum learning outcomes for the most advanced students.

Additionally, Hallam and Ireson (2003) found strong agreement that ability grouping was beneficial, especially for teachers, because it was advantageous for student learning, while also making class management easier. Furthermore, when using ability grouping, it was possible for teachers and programs to better design curriculum to meet the needs of a variety of students. Regarding subject matter, ability grouping was seen as specifically beneficial when it came to the disciplines of mathematics and foreign languages.

Other studies (e.g., Kim, 2012; Slavin, 1990), in contrast, reported ability grouping to be only slightly beneficial or ineffective. According to Slavin's (1990) review on 29 (experimental, correlational, or case) studies of between-class ability grouping for junior high and high school students, no positive effects on student achievement were observed. This review of ability grouping, in the US, included students in various courses over a period of five years.

Kim (2012) found that there was not a positive attitude among students toward between-class ability grouping. A survey was administered as part of the study to 754 students from six different Korean middle schools (grades 7-9). Due to a variation in ability grouping practices among the schools participating in the study, Kim focused on only three comparable schools. Findings revealed that in two schools with three group levels (high, intermediate, low), higher-level students' responses to between-class ability grouping were slightly positive or neutral, while lower-level students were neutral or negative about its effectiveness. Moreover, in one school with two group levels, both high and low-level students reported a negative attitude toward between-class ability grouping.

While the findings of ability grouping research in secondary contexts across various disciplines have been divergent, studies done in post-secondary EFL settings tended to observe positive results (Wen, 2011). Khazaeenezhad, Barati, and Jafarzade (2012) conducted an experimental design inquiry using test and control groups to examine the effectiveness of ability grouping on college-level English language learners in Iran. The study investigated ability grouping (less-able, intermediate, and advanced groups) and various amounts of exposure to English (two, three, and four hours) in relation to academic gains in one semester. The study recruited 320 non-English major undergraduates and divided them into *different ability* groups and *non-ability* groups. Findings indicated that the students in the ability groups significantly outperformed their counterparts in the different ability groups as exemplified by their test scores. This clearly revealed the positive effects of ability grouping on the subjects' academic gains in GE training.

Ability grouping implemented in GE training in Taiwan

Specific to the Taiwanese EFL context, ability grouping has been a popular policy in secondary education and widely advocated and practiced by many universities and colleges (Chern, 2010; Feng & Chang, 2010; Lee & Su, 2009; Wen, 2011). Some studies have reported a positive impact on learning and a positive attitude from students regarding the effectiveness of ability grouping in the GE courses (Lee & Su, 2009; Liu, 2008; Wen, 2011).

For example, ability grouping was positively perceived by university instructors, as well as by students (Liu, 2008). In a survey, Liu investigated the perceptions of 582 freshmen and sophomores and 34 English teachers at university in central Taiwan. The focus of the survey was to measure the participants' attitudes toward ability grouping. The participants were divided into the following groups based on their scores from the *General English Proficiency Test* (GEPT) when they enrolled at the university: (a) basic, (b) intermediate, and (c) advanced. The GEPT is an English proficiency assessment designed by the *Language Training and Testing Center* (LTTC) of Taiwan to measure citizens in four skills of English proficiency. The four skills assessed were: (1) listening; (2) speaking; (3) reading; and (4) writing. From the learners' perspective, the results demonstrated that freshmen held positive attitudes toward ability grouping, particularly those with basic English proficiency. These students reported that working with students of similar ability reduced the pressure and anxiety of learning and enhanced their motivation. However, positive attitudes toward ability grouping weakened by the end of sophomore year.

In a semester-long project, Lee and Su (2009) studied 2,230 non-English-major students from a technical university in Taiwan with respect to ability grouping. The participants were leveled into three ability groups based on English proficiency: (1) beginning, (2) intermediate, and (3) higher intermediate. The aim of the project was to compare achievement scores before and after taking the one-semester, freshman English course. Results from the students' achievement tests indicated a significant difference between pre- and post-test scores, with the intermediate level students making the greatest progress. Yet, the short intervention time (one semester) and lack of detailed description on leveled instruction lessened the validity of the findings.

Likewise, the purpose of Wen's (2011) study was to examine the effects of ability grouping on technical university students' general English learning achievement in a year-long, two-semester, program. The subjects in this study consisted of 792 freshmen from three colleges at one university (business, engineering, and electronics and information) who were divided into three ability groups: (1) high achievers, (2) medium achievers, and (3) low achievers. Ability grouping was based on students' English scores from the *Joint College Entrance Exam* (JCEE), a regular (non-technical) university entrance exam. All the participants took one pretest (listening and reading) before the year began and two posttests (listening and reading) at the end of the GE course. The findings indicated that low-achieving students did not benefit from ability grouping, but students in the medium and high groups showed significant progress on listening and reading scores.

Some studies related to the two previously discussed research lines—English as a native language (ENL) (Slavin, 1990) studies and EFL studies (Kim, 2012; Trautwein, Koller, & Kammerer, 2002)—voiced concerns about the potential negative consequences caused by the implementation of between-class ability grouping. These concerns were raised because students who were less proficient in EFL were deprived of what could be better instruction because the teachers had lower expectations of them compared to their more proficient counterparts (Kim, 2012). Kim (2012) also reported that between-class ability grouping often had adverse effects because it widened the gap between high- and low-level learners. Kim (2012) concluded that the effectiveness of ability grouping was determined by how it was implemented (e.g., the number of group levels) and if it was supported by other school policies. In studying the effects of ability grouping on students grades 6 to 9 in EFL and math classes, Trautwein, Koller and Kammerer (2002) found that the between-class ability grouping enhanced lower ability students' academic involvement in class.

These concerns about ability grouping, however, have been addressed by researcher such as Khazaeenezhad et al. (2012) and Wen (2011). They suggested that the negative possible effects of ability grouping could be mitigated through careful planning and decision making where all the different agents in the teaching and learning process collaborated. Policy makers are tasked with the responsibility for how ability grouping is implemented, keeping in mind the different levels with respect to curriculum design, materials development. Teachers should be trained in how to deliver the instruction at the specific level to which they are assigned.

Taken together, two significant gaps exist in the above-discussed research. First, little empirical research to date has explored ability grouping and proficiency gains over time. Among the limited EFL studies done (e.g., Khazaeenezhad et al., 2012; Wen, 2011), most were done during a period of one semester or one academic year, thereby contributing limited information about the short-term effects and not exploring the long-term effects. Second, relevant studies have used different proficiency measurements where subjects varied by absolute proficiency level. Lee and Su's (2009) and Wen's (2011), for example, were different in relation to their tests and subject proficiency level. Comparing their findings would therefore be problematic. Perhaps a more widely used reference framework could provide researchers common reference points for students' proficiency levels. These issues indicate the need for (a) long-term studies with a systematic examination of how universities' GE curriculum prepares students to meet language proficiency requirements and (b) the need for using a common reference framework to ensure a consistent interpretation of students' proficiency levels.

To address these research gaps in the literature, the current study conducted a longitudinal study in Taiwan within the context of EFL programs. Specifically, it adopted a quasi-experimental within-subject design and interpreted students' proficiency levels based on the Common European Framework of Reference for Languages (CEFR – Council of Europe, 2011).

Research Questions

Having noted the aforementioned gaps in the research, this quasi-experimental study examined whether or not students enrolled in a longitudinal, ability-grouping GE curriculum improved over time at a university in Southern Taiwan (hereafter SU). The study proposed the following research questions:

1. How do the observed English (L2) proficiency scores of a group of Taiwanese EFL university students change over time where their GE curriculum was designed around ability-grouping principles?
2. How do the observed English (L2) proficiency scores of a group of Taiwanese EFL university students who were observed to have an Al, A2, or B1 CEFR level upon entry to the university change over time where their GE curriculum was designed around ability-grouping principles?

RQ2 was posed as consisting of three separate hypotheses and as a post hoc of RQ1.

Methodology

Subjects

This study used the three-year CSEPT test records from 785 students at SU. The subjects were first enrolled in the 2012 academic year and received a three-year-long intervention of leveled GE (ability grouping) instruction from Fall 2012 to Spring 2015. Because the study was designed as a longitudinal study, these 785 subjects took the pretest, the one-year posttest, the two-year posttest, and the three-year posttest. In other words, any subjects who did take the CSEPT these four times were excluded. The informed consent for each student was obtained prior to taking the pretest.

At the beginning of the subjects' freshman year at SU, they took the CSEPT pretest for placement purposes. Based on the CSEPT scores, all freshmen were grouped within Levels One to Eight. Table 1 illustrates the range of scores from Levels One (lowest) to Eight (highest), their equivalence on CEFR, and the number of subjects at each level. In addition, Level 7 subjects (N=60), equivalent to B2 level on CEFR, were excluded from the analysis for the second research question given that they had reached graduation benchmark and accounted for a small percent of the total sample. On the basis of CEFR, of the 725 subjects (excluding 60 B2 level subjects), 110 were observed to have an A1 level, 223 an A2 level, and 392 a B1 level upon entry to the university.

College Student English Proficiency Test

This study adopted the *College Student English Proficiency Test* (CSEPT) as the initial reference points for placing students into different levels/groups by proficiency, which were converted to CEFR later on for data analysis purpose. The CSEPT, designed by the *Language Training and Testing Center* (LTTC)[see Endnote 1] for higher education institutes in Taiwan, is an English proficiency test for EFL college students. The purpose of the test was to evaluate university students' English proficiency; primarily targeting students' receptive skills including listening, reading and grammar. The test fulfills the need of analyzing the outcomes of English language teaching and learning. The Primary Level CSEPT was made available in 1997 followed by the Secondary Level in 1998 (LTTC, 2007). Table 1 presents the measurements of the CSEPT and its equivalent, the *Common European Framework of Reference for Languages* (CEFR, Council of Europe, 2011) as illustrated by LTTC (n.d.). The primary level is the equivalent of CEFR B1, a level at which the test questions measure intermediate level of proficiency. The secondary level is the equivalent of CEFR B2 and measure the English proficiency of intermediate to advanced level learners. The test has been adopted by many technical schools and colleges and universities in Taiwan (Pan & Newfields, 2012). It is intended to measure language learners' receptive skills, such as listening and reading proficiency within the context of everyday and campus life.

Table 1
Level Groups and Their Proficiency at SU for Freshmen

		General English at SU		
CEFR	SU's GE Level	Student's CSEPT Score	Subjects (N)	Notes
A1	1	~119	110	Level 1 – Level 4:
A2	2	120~144	111	Extra 2 hours of remedial instruction;
	3	145~169	112	Self-access to learning resources
B1	4	170~200	119	
	5	201~219	166	
	6	220~239	107	
B2	7	240~259	60	Benchmark for non-English majors
	8	260~344	NA	Benchmark for English related majors
Total:			785	

Notes.
1. During the time of this study, a level nine course was not offered and was not implemented until 2016 for freshman.
2. English related majors include: English, Foreign Language Instruction and Translation and Interpreting.

Adopting CSEPT at SU

In 1997, SU was one of the universities to adopt CSEPT as a mandatory test for all students to measure their gains in English proficiency. The university adopted CSEPT for three reasons: first, to place students in leveled classes based on their proficiency; second, to document learners' language proficiency so that SU could constantly evaluate the effectiveness of its language curriculum; and third, as an English proficiency benchmark for students to fulfill as a partial graduation requirement. The secondary-level CSEPT test was used for all students at SU and administered to the subjects of this study. This CSEPT test has three sections: listening, grammar, and reading. First, in the listening test, students listen to and understand short conversations in addition to short speeches. The listening test includes a total of 30 questions. For the grammar test, students are required to complete sentences and short passages that consist of 50 questions. Finally, the reading test consists of 30 reading comprehension questions. The total time allowed to complete the CSEPT test is 90 minutes.

In this study, the CSEPT tests were officially administered by the LTTC at SU when the subjects attended a mandatory summer camp before their first semester. The actual CSEPT scores collected during the summer camp were considered the pretest scores (T1). Near the end of the first (T2), second (T3), and third (T4) years of the GE training program, the official CSEPT tests were administered by LTTC as posttests to measure the students' progress in English.

Southern University (SU) and its ability-grouping GE curriculum

Founded in 1966, SU is known for its foreign language pedagogy, with a vision that all students will demonstrate English proficiency to complement their knowledge in their respective majors, such as communication arts, digital content application, international business, international affairs, foreign language teaching, etc. SU believed that through foreign language learning, students would be able to understand global culture and expand their world views. Thus, each college student was required to take an adaptive three-year GE program before they graduated. The program was designed to ensure students' English language proficiency by the time they exited the program. For example, students at SU were eligible to become exempt from some credit hours as soon as they completed the highest level of English proficiency (level nine) or when they demonstrated high English proficiency (CSEPT test score over 345). For example, if a student's level of English proficiency was at eight when admitted to SU, he would be required to take two years of EFL to exit the program. In other words, he would only have to complete 16 credit hours. As for level upgrading, two rules applied. The first was to upgrade students to one level automatically after one year of GE training, regardless of their updated CSEPT scores. The other was that students could apply to be upgraded, to the appropriate level, based on their updated CSEPT scores.

The regular GE program design was conceptualized in a student proficiency-based teaching philosophy and embraced Communicative Language Teaching (CLT) as an approach. The program offered the integration of four skills- listening, speaking, reading, and writing- and leveled materials for each skill. The teaching delivery varied, including listening and speaking training, simulated dialogues, reading skills, and writing practices.

General English is a required subject for all students at SU. Beginning in 2013, SU required students to complete a total of 24 credits (approximately six courses with each class worth four credits) in English within the first three college years. However, this excludes an additional two non-credit hours of remedial instruction per week for students under Level Five during freshman and sophomore years. This consideration was based on the assumption that students' exposure to English would gradually increase their EFL proficiency. It is noteworthy that the total hours required by SU were exceptionally high compared to other universities in Taiwan. For example, Chern (2010) reported that a range of four to six credit hours (usually two to three courses) was a common requirement at many universities in Taiwan.

Although the program at SU requires students to take 24 credit hours, not all students receive the same amount of English training. The primary feature of the GE curriculum at SU is that it is adapted to students' level of English proficiency, measured by a recognized English proficiency test. Students whose CSEPT pretest is under 200 (equivalent to CEFR A1 and A2 levels) receive additional two hours of remedial instruction per week, whereas students whose entry level is Level Nine (equivalent to CEFR B2 level) were only required to take 1 year of GE instruction. This allowed those students to take advanced English or English as a Medium Instruction courses as electives. Table 1 illustrates the group levels and the entry levels of freshmen students at SU. Students whose levels were under five in their freshman and sophomore years (i.e., CSEPT test score below 200; the CSEPT test will be introduced later in the text) received an additional two hours per week, for remedial instruction.

The language curriculum for the subjects was tailored to meet their different needs. In addition to the remedial hours, SU provided each student self-access to language learning consultation and resources in the Language Diagnostic and Consulting Center (LDCC). In the LDCC, students can consult teachers about their learning styles and strategies, as well as practicing language with computer assisted leaning programs. In order to provide incentives for students to study English on their own time, the record of students' self-access learning progress was considered part of their overall course performance.

To ensure teaching quality and consistency, SU implemented a structured curriculum with the same textbooks being used by all teachers at each level as determined by the level coordinator in consultation with instructors. Exams were also created and administered in a similar fashion. For each level, teachers were expected to be consistent with their content materials and assessments. The CSEPT washback effect was minimal because SU did not tailor the curriculum to prepare students to take the CSEPT, as the school-based exams evaluated both the receptive and productive skills of language learners, including speaking and writing. Every semester, faculty meetings were held several times a semester for staff to discuss their teaching with other colleagues, including the authors of the study.

Data Collection and Analysis

The subjects took the official CSEPT administered by LTTC as a pretest for placement purposes at the beginning of their freshman year. After approximately one year of GE training, in May, at the end of their freshman year, the subjects took an alternate version of CSEPT for the one-year posttest. After two years of GE training, the subjects took another CSEPT for the two-year posttest. Finally, in a three-year English program, the subjects took the last official CSEPT, the three-year posttest. The subjects' pretest scores were regarded as their English proficiency before the intervention of GE classes at SU. The posttest scores were considered a measure of the subjects' progress in English after taking the GE classes.

To analyze the data that had a within-subjects design, one-way repeated measures ANOVAs were performed to investigate whether all the subjects' observed English proficiency scores changed significantly over time (RQ1). This same analysis was then done on three groups from the sample to answer RQ1: A1 upon entry, A2, B1. Main effect sizes (time and proficiency scores), for RQ1 and RQ2, were reported via partial-eta-squared (Lakens, 2013). Post hoc pairwise comparisons where employed to assess significance and effect size (via Cohen's d-average) of differences between two measurements, e.g, pretest and year one posttest. The magnitude thresholds for d-average are the same as with Cohen's *d* (see Cohen, 1988: .2-small - .5-medium - ,8-large) way repeated measures ANOVA were referred to as a within-subject ANOVA for the same group of subjects. Since RQ2 was framed as 3 independent hypotheses, α-level for statistical significance left at .05. For all ANOVAs, the Mauchly's tests were significant (p_s <.01), indicating that the sphericity assumption was violated, Greenhouse-Geisser corrections were therefore applied.

Results

First Research Question

Table 2 presents the descriptive statistics of subjects' proficiency scores over time. Via one-way repeated measures ANOVA testing, a significant association was observed among the four proficiency scores across time [*F (2.53, 1982)* =223.34, *p* <.01, partial-eta-squared=.22]. There were 5 significant (p_s < .01) observed pairwise post hoc comparisons: T4 > T3 (d-average=.1); T4 > T2 (=.12); T4 > T1 (=.42); T3 > T1 (=.33); T2 > T1 (=.3). T3 > T2 (p=.14; d-average=.02) was nonsignificant.

Table 2
Descriptive Statistics of the Scores over Time

	M	*SD*	*N*
Pretest	188.82	57.27	785
Freshman posttest	206.12	57.00	785
Sophomore posttest	207.49	57.10	785
Junior posttest	213.00	56.47	785

Second Research Question

The second research question was analogous to the first except for the creation of 3 independent samples based on observed CEFR proficiency level upon entry into the program: A1, A2, B1.

Table 3 presents the descriptive statistics of the A1 group's proficiency scores over time. Via one-way repeated measures ANOVA testing, a significant association was observed among the four proficiency scores across time [*F (2.19, 238.22)* =39.98, *p* <.01, partial-eta-squared=.27]. There were 5 significant (p_s < .01) observed post hoc comparisons: T4 > T2 (d-average=.32); T4 > T1 (=.92); T3 > T2 (=.22); T3 > T1 (=.84); T2 > T1 (=.64). T4 > T3 (p=.21; d-average=.12) was nonsignificant.

Table 3
Descriptive Statistics of the Scores: A1 Group

	M	*SD*	*N*
Pretest (T1)	103.08	26.22	110
Freshman posttest (T2)	120.73	30.95	110
Sophomore posttest (T3)	127.67	31.93	110
Junior posttest (T4)	131.56	35.69	110

Table 4 presents the descriptive statistics of A2 group's proficiency scores over time. Via one-way repeated measures ANOVA testing, a significant association was observed among the four proficiency scores across time [*F (2.58,* 572.49*)* =111.71, *p* <.01, partial-eta-squared=.34]. There were 6 significant (p_s < .01) observed post hoc comparisons: T4 > T3 (d-average=.25); T4 > T2 (=.42); T4 > T1 (=1.17); T3 > T2 (=.15); T3 > T1 (=.87); T2 > T1 (=.79).

Table 4
Descriptive Statistics of the Scores: A2 Group

	M	*SD*	*N*
Pretest (T1)	147.91	23.53	223
Freshman posttest (T2)	168.73	28.90	223
Sophomore posttest (T3)	173.43	34.76	223
Junior posttest (T4)	182.25	35.28	223

Table 5 presents the descriptive statistics of B1 group's proficiency scores. Via one-way repeated measures ANOVA testing, a significant association was observed among the four proficiency scores across time [*F (2.55, 995)* =76.32, *p* <.01, eta-partial-squared=.16]. There were 5 significant (p_s < .01) observed post hoc comparisons: T4 > T3 (d-average=.16); T4 > T2 (=.1); T4 > T1 (=.65); T3 > T1 (=.48); T2 > T1 (=.61). T2 > T3 was nonsignificant with a direction contrary to expectation.

Table 5
Descriptive Statistics of the Scores: B1 Group

	M	*SD*	*N*
Pretest (T1)	222.15	24.93	392
Freshman posttest (T2)	237.99	26.61	392
Sophomore posttest (T3)	235.86	32.75	392
Junior posttest (T4)	240.98	32.64	392

Discussion

This study first aimed at uncovering whether the students made progress over time after enrolled in the three-year GE training program with ability grouping between-class. The findings demonstrated significant gains in the subjects' CSEPT scores from the first year to the third year. Time and proficiency scores shared 22% of the variance, and post hoc comparisons revealed 5 significant differences where the later test score average was higher. These observations suggested that students, in the aggregate, had made somewhat continuous progress over time since their first-year enrollment in the three-year GE program designed based on ability grouping principles at SU. This finding was constrained and limited by the lack of a comparison with a control group or non-ability grouping treatment group.

Ability grouping supporting L2 proficiency attainment over time was also suggested by the existing literature in domestic (Wen, 2011) and international contexts (Khazaeenezhad et al., 2012). Significant gains were found in the majority of students who received long-term leveled GE instruction. Unlike Wen (2011) who focused on the effect of one-year leveled GE instruction, the current study demonstrated gains over three years. Therefore, this inquiry has contributed to the case for ability grouping in both Taiwanese and other contexts.

In spite of the proficiency gains, students' motivation may have lessened after the first year of GE instruction (e.g., T3 > T2 – nonsignificant; T4 > T3 – d-average=.09/very weak effect) due to the lack of integrated, as well as instrumental, motivation. Warden and Lin (2000) posited that Taiwanese students at a technical college had undergone this very phenomenon. As pointed out by Hua and Beverton (2013), GE courses in Taiwan were made compulsory to increase the nation's global competitiveness. However, if the courses did not relate to their major field of study throughout the program's duration, learning English might not have offered any specific value to the students. The subjects of this study might not have seen the value of their efforts and eventually lost interest in learning English.

L2 Proficiency Gains across Different L2 Proficiency Levels upon Entry

This study also proposed to investigate how groups at different CEFR levels upon entry varied in relation to proficiency gains. The A2 group as presented above had the strongest gains over time as evidenced by all 6 possible post hoc comparison being significant, in the expected direction and the highest observed partial-eta-squared. The A1 group also made gains, but they were not as pronounced as the A2 group. The B1 group had the weakest observed effect between time and proficiency scores with one post hoc comparison (T2 > T3) being in a direction contrary to expectation. This difference was nonsignificant however.

A similar result found in Kim's (2012) study was also observed regarding the subjects with an A1 entry level in this study. Kim (2012) found that mid- and lower-level students struggled with ability grouping and suggested that ability grouping alone would not lead to significant improvements in students' English proficiency. In order for ability grouping to create an environment where significant gains could be made, Kim claimed that it would be necessary to have a combination of curricula that corresponded to the students' learning styles, interests, and abilities.

Lastly, Level A1 and B1 students' attitudes when taking the CSEPT may have influenced the findings. A1 students' lack of achievement may be due to their low self-esteem as they probably knew they were the lowest group. On the other hand, perhaps a plateau effect occurred among B1 learners. As most of the students whose entry level was B1 passed the required graduation benchmark (240 for non-English majors – B2) in the first year of GE learning, these subjects may not have been taking the subsequent tests seriously, leading to underperformance.

Conclusion

Adopting a quasi-experimental (within-subject) design, this study contributed to the understanding of the observed longitudinal language gains of learners who received General English (GE) instruction designed around ability-grouping principles. EFL students with an A2 entry level experienced an ongoing progress when the stratified English curriculum was adaptive to offer remedial instruction and to challenge their current level by upgrading annually. Below are implications for EFL education and suggestions for future research.

Implications for EFL education

Ability grouping has been commonly used in EFL college settings including Taiwan, particularly when implementing GE courses. The implication is that policy makers need to re-conceptualize ability grouping as a way to increase language proficiency (Lee & Lin, 2013). The ultimate goal of ability grouping is not to widen the proficiency gap among language learners but to offer different kinds of scaffolding for different levels of students. Therefore, a well-designed leveled (ability grouping) curriculum helps learners to challenge their current levels. Rather than seeing students' diverse levels as a problem, teachers and policy makers can regard it as an opportunity to make the curriculum more adaptable for learners at all levels. As Kim (2012) noted, the effects of ability grouping can be enhanced or lessened depending on materials used, teaching hours, assessments, and resources provided by a university. Administrators need to consider how the leveled curriculum is implemented and adapted, and for what purpose. A leveled curriculum requires an integration of school-related learning resources such as remedial instruction and self-access learning into the curriculum. It also allows teacher collaboration within the same level to share their experiences. This allows the school to provide various accommodations for the needs of students with differing proficiency levels.

The current study also illustrated how A2 upon entry students, as a group, had progressed most since their enrollment in a long-term program designed based on ability grouping. Perhaps, the supplementary remedial intervention and the use of the school's self-access learning resources worked best for them. The effect of these types of resources can be positive for many students with a similar entry level. However, the language progress was least pronounced for B1 upon entry students, particularly those who have reached the English proficiency benchmark for graduation. These students may be more focused on their professional studies rather than concentrating on English language learning.

Implications for Future Research

As this study was quasi-experimental in design, future researchers may consider conducting studies involving a control group to compare gains via an experimental design. These studies could also include other treatment groups to compare other class organization strategies, e.g., heterogenous (in relation to proficiency) grouping. Like this study, future studies may continue to use CEFR to systematically report learners' levels of proficiency in their research designs. By doing so, researchers can make cross-study comparisons to show the effects of leveled GE instruction in various contexts. Finally, as the findings of this study suggest, A2 learners progress more than those of the other levels, future studies could further investigate the phenomenon of why A2 learners make smooth gains, whereas those of the other levels do not. Studies could also delve into the ceiling effect for B1 learners as implied by this study. Qualitative studies with interviews or observations could be conducted to explore learners' perceptions of effectiveness of ability grouping as a curricula and program design scheme.

Limitations of this study

Our design for the ability groups has several limitations. First, the CSEPT only evaluated students' listening, reading and grammar (usage norms) proficiency. In contrast, the course design and materials at SU were integrated with all four language proficiency skills (speaking, writing, listening, reading). Although the GE courses at SU placed a strong emphasis on the subjects' speaking and writing skills, students' production skills were not measured by the CSEPT. Second, the program's automatic progression, for level X to X + 1, regardless of CSEPT score could have influenced the observed findings. The final and, perhaps, most important limitation was the lack of a control group. The findings of this current study cannot be used to suggest direct causality.

Endnotes

1. The CSEPT's psychometrics have been assumed as credible for some time in the Taiwanese context given its long history of development. The CSEPT's governing body, the LTTC is partners with several international English proficiency testing groups such as Cambridge language assessment. Given these observations, the CSEPT's validity and reliability was accepted on its face. LTTC website: https://www.lttc.ntu.edu.tw/languagetesting.htm
2. Beginning in May 2016, all test takers who reached the English proficiency benchmark could be exempted from taking further CSEPT tests. However, this is not applicable to this study.

Acknowledgements

1. The Editorial Board wishes to thank: Dr. Daniël Lakens for his input on Cohen's d-average, and Dayna Jost for her role as a guest production editor for this paper.
2. The authors would like to express their sincere gratitude to the two anonymous reviewers for their constructive comments that helped to improve the quality of the manuscript. They also would like to thank Dr. Cathrine Beaunae for English language editing.
3. This study was one of an integrated research project funded by Wenzao Ursuline University of Languages, Kaohsiung, Republic of China. (Grant #: GERS 104002).

References

Chen, A. (2011). Parents' perspectives on the effects of the primary EFL education policy in Taiwan. *Current Issues in Language Planning*, *12*(2), 205-224.

Chen, I. W.-L., & Hsieh, J. J.-C. (2011). English language in Taiwan: An examination of its use in society and education in schools. In A. Feng (Ed.), *English language education across greater China* (pp. 70-94). Bristol: Multilingual Matters.

Chern, C. (2002). English language teaching in Taiwan today. *Asia-Pacific Journal of Education*, *22*(2), 97-105.

Chern, C. (2010). General English programs at universities in Taiwan: Curriculum design and implementations. *Chang Gung Journal of Humanities and Social Sciences, 3*(2), 253-274.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Feng, A. (2012). Spread of English across Greater China. *Journal of Multilingual and Multicultural Development*, *33*(4), 363-377. doi:10.1080/01434632.2012.661435.

Feng, M., & Chang, H. (2010). Effects of ability grouping and development of English graduation benchmarks in a national technological university in Taiwan. *Journal of Applied Foreign Languages*, *13*, 71-101.

Hallam, S., & Ireson, J. (2003). Secondary school teachers' attitudes towards and beliefs about ability grouping. *British Journal of Educational Psychology*, *73*(3), 343-356.

Hua, T.-L., & Beverton, S. (2013). General or vocational English courses for Taiwanese students in vocational high schools? Students' perceptions of their English courses and their relevance to their future career. *Education Research Policy Practice, 12*(2), 101-120.

Ireson, J., Hallam, S., Hack, S., Clark, H., & Plewis, I. (2002). Ability grouping in English secondary schools: Effects on attainment in English, mathematics and science. *Educational Research and Evaluation*, *8*(3), 299–318.

Ireson, J., & Hallam, S. (2001). *Ability grouping in education*. London: Paul Chapman Publishing House.

Jung, M. (2000). Ability grouping and equity of educational opportunity. *The Journal of Curriculum Studies*, *18*, 275–297.

Khazaeenezhad, B., Barati, H., & Jafarzade, M. (2012). Ability grouping as a way towards more academic success in teaching EFL – A case study of Iranian undergraduates. *English Language Teaching*, *5*(7), 81-89.

Kim, Y. (2012). Implementing ability grouping in EFL contexts: Perceptions of teachers and students. *Language Teaching Research*, *16*(3), 289-315.

Krashen, S. D. *(1985). The input hypothesis: Issues and implications*. New York: Longman.

Kulik, J. A. (1992). *An analysis of the research on ability grouping: Historical and contemporary perspectives.* Storrs, CT: National Research Center on the Gifted and Talented.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 1-12.

Language Training and Testing Center (2007). Research reports on College Students English Proficiency Test—Primary level. Retrieved on July 9th, 2017, from https://www.lttc.ntu.edu.tw/academics/csept/CSEPT%E7%AC%AC%E4%B8%80%E7%B4%9A%E5%B8%B8%E6%A8%A1%E5%BB%BA%E7%BD%AE%E7%A0%94%E7%A9%B6%E8%A8%88%E7%95%AB%E5%A0%B1%E5%91%8A.pdf.

Language Training and Testing Center (n.d.). CESPT and CEFR comparison chart. Retrieved on July 9th, 2017, from https://www.lttc.ntu.edu.tw/CEFRbyLTTC_tests.htm.

Lee, C., & Lin, C. (2013). A method of English ability-grouping teaching in a university of technology. *International Journal of Information and Education Technology*, *3*(2), 268-272.

Lee, C.-L., & Su, C. C. (2009). The effectiveness of English ability-grouping teaching: A case study of a university of technology in central Taiwan. *Journal of Humanities and Social Science of NHCUE, 2*(1), 233-253.

Liu, H. J. (2008). An analysis of the effects of ability grouping on student learning in university-wide English classes. *Feng Chia Journal of Humanities and Social Sciences*, *16*, 217-249.

Pan, Y., & Newfields, T. (2012). Tertiary EFL proficiency graduation requirements in Taiwan: A study of washback on learning. *Electronic Journal of Foreign Language Teaching*, *9*(1), 108-112.

Slavin, R. E. (1987). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research*, *57*(3), 293-336.

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research*, *60*(3), 471-499.

Trautwein, U., Koller, O., & Kammerer, E. (2002). The effects of within-class and between-class ability grouping on academic self-concepts, reported classroom participation, and perceived social integration. *Psychologie in Erziehung Und Unterricht, 49*(4), 273-286.

Warden, C. A., & Lin, H. J. (2000). Existence of integrative motivation in an Asian EFL setting. *Foreign Language Annals*, *33*(5), 535-545.

Wen, S.-M. (2011). An analysis of the implementing outcomes of ability grouping of Freshman English in a university of technology. *Journal of National Taichung University of Education*, *25*(2), 65-80.

About the Authors

Chiu-hui (Vivian) Wu is an associate professor in the Department of English and the Director of Center for English Language Teaching at the Wenzao Ursuline University of Languages, Kaohsiung, Taiwan. Her research interests include: intercultural education, English language education and qualitative research.

Chia-jung Tsai is an assistant professor in the Center for English Language Teaching at Wenzao Ursuline University of Languages, Kaohsiung, Taiwan. Her areas of research include learning strategies, vocabulary teaching, and materials design.

Yi-Min Chiu is currently an English lecturer at Wenzao Ursuline University of Languages and a doctoral candidate at Department of Foreign Languages and Literature, National Cheng Kung University, Taiwan. Her main research interests lie in peer response in L2/FL writing and English for Academic Purposes (EAP).