The effect of IELTS test preparation and repeated test taking on Chinese candidates' IELTS results, general proficiency, and their subsequent academic attainment

Ruolin Hu

PhD

University of York

October 2018

Abstract

To achieve academic success in English speaking higher education, English proficiency is key (Trenkic & Warmington, 2018). However, IELTS – a standardised test of English proficiency frequently used as a university entry requirement – has been reported an inconsistent predictor (e.g. Feast, 2002; Dooey & Oliver, 2002; Woodrow, 2006). Here, two quasi-experiments were conducted to explored potential reasons behind this inconsistency.

Using a pretest/intervention/posttest design, quasi-experiment 1 investigated whether IELTS test-preparation programmes can boost IELTS scores beyond test takers' actual levels of proficiency (N=89). In the intervention group, a significant boost in IELTS scores from pre-test to post-test was found but there was no significant improvement in general proficiency, measured through another standardised proficiency test, a vocabulary test, and a processing accuracy/speed test. In the control group, no difference in results on either IELTS or the other three measures.

Quasi-experiment 2 first examined whether repeated test-taking may boost IELTS scores beyond test-takers' actual proficiency levels and then explored the predictive validity of IELTS on academic attainment among 153 Chinese students at a UK university. Results indicated that repeated test-taking inflated IELTS scores beyond the level of proficiency, but only marginally. IELTS was significantly correlated with academic grades for students from both linguistically more and less disciplines. Moreover, IELTS was found to be a good predictor for grades for the linguistically more demanding.

In short, IELTS scores can be boosted beyond one's actual proficiency by attending dedicated test-preparation courses and to a lesser extent by taking the test repeatedly. Hence students admitted on the premises of a certain IELTS may in fact be of lower proficiency. This discrepancy can impact their academic achievement. This study offers insights as to why prior research on the relationship between IELTS scores and academic attainment yielded inconsistent findings and considers theoretical and practical implications.

Table of Content

***Abstract*......2**

***Table of Content*......3**

***List of Tables*......7**

***List of Figures*......10**

***Acknowledgement*......12**

***Author's declaration*......13**

***Chapter 1 Introduction*......14**

***Chapter 2 Literature Review*......19**

2.1 International students in English-speaking higher education institutions .19

2.1.1 How are international students performing academically?......20

2.1.2 International Chinese students at English speaking HEIs......22

2.2 Factors affecting academic performance of international students and the role of language proficiency......24

2.2.1 Sociocultural and psychological challenges......24

2.2.2 Differences in learning environment and learning style......26

2.2.3 Language as the main obstacle......29

2.3 Understanding L2 proficiency......32

2.3.1 Earlier frameworks(1960s-1980s): approaching proficiency as a general construct.32

2.3.2 Recent studies (1990s-2010s): approaching proficiency from four language skills ..33

2.3.3 Acknowledged indicators of L2 proficiency......36

2.4 Standardized English proficiency tests......38

2.4.1 ELTS and IELTS......38

2.4.2 TOEFL......39

2.5 Test reliability and validity......41

2.5.1 Reliability......42

2.5.2 Validity......43

2.5.3 Content validity......43

2.5.4 Construct validity and relevant threats......45

2.5.5 Criterion-related validity: concurrent validity and predictive validity......46

2.5.6 TOEFL as a predictor for academic attainment......48

2.5.7 IELTS and academic attainment 60
2.5.7 Alternative measures of proficiency and international students' academic attainment 68

2.6 Test preparation and relevant concerns 73

2.6.2 Test washback and test preparation 78
2.6.3 Preparation courses for proficiency tests and effects on test scores 81

2.7 Repeated test taking and test performance 92

2.8 Context of this present study 93

2.8.1 High-stake English proficiency tests and Chinese candidates 93
2.8.2 Test preparation for English proficiency tests in China 95
2.8.3 Chinese candidates' repeated test taking behaviour and relevant concerns 97
2.8.4 The present study 98

Chapter 3 Quasi-experiment 1: the effects of test preparation on Chinese candidates' IELTS test results and their general proficiency 99

3.1 Methodology and methods 99

3.1.1 Research Design 99
3.1.2 Participants 100
3.1.3 The Intervention 102
3.1.4 Research Instruments 103
3.1.5 Reliability 114
3.1.6 Ethics 116
3.1.7 Procedure 117
3.1.8 Scoring 118
3.1.9 Hypotheses 121
3.1.10 Analyses 123

3.2 Results and Analysis 124

3.2.1 To what extent does IELTS preparation course improve candidates' IELTS scores (overall and by skill)? 124
3.2.2 Exploratory analysis: do candidates of lower proficiency achieve more IELTS score gains (overall and by skill) than candidates of higher proficiency through attending IELTS test preparation? 135
3.2.4 To what extent does IELTS preparation course improve candidates' general proficiency, measured by linguistic measures they were not prepared for? 137

3.3 Discussion 146

3.3.1 To what extent does dedicated IELTS preparation course boost IELTS scores? 147
3.3.2 To what extent does dedicated IELTS preparation course improve candidates' general proficiency? 152

3.4 Conclusion 153

Chapter 4 Quasi-experiment 2: the effects of repeated test taking and test preparation on Chinese candidates' IELTS scores, general proficiency, and their academic attainment at a UK university *155*

4.1 Methodology and methods **155**

4.1.1 Research Design 155

4.1.2 Participants 156

4.1.3 Research Instrument 158

4.1.4 Ethics 168

4.1.5 Procedure 169

4.1.6 Scoring 170

4.1.7 Research questions and Hypothesis 172

4.1.8 Analysis 174

4.2 Results and Analysis **175**

4.2.1 Descriptive statistics for all linguistic measures 175

4.2.2 Effect of test preparation on Chinese candidates' IELTS scores (overall and by skill) and their general proficiency 177

4.2.3 To what extent does repeated IELTS taking affect Chinese candidates' IELTS scores and their general proficiency? 179

4.2.4 To what extent does IELTS predict international Chinese students' academic attainment at a UK university? 184

4.2.5 Have test preparation and repeated test taking eroded the validity of IELTS as a predictor for academic attainment? 195

4.2.6 Explorative analysis: IELTS repeater's profile 199

4.2.7 Explorative analysis: Does the predictive validity of IELTS on academic attainment change over time? 202

4.2.8 Explorative analysis: to what extent does general proficiency (indicated by the composite score of DET and C-test scores) predict international Chinese students' academic attainment? 203

4.3 Discussion 204

4.3.1 To what extent does repeated IELTS taking affect Chinese candidates' IELTS scores and their general proficiency? 205

4.3.2 To what extent does IELTS predict international Chinese students' academic attainment at a UK university? 207

4.4 Conclusion **209**

Chapter 5 General Discussion *210*

5.1 Test preparation, repeated test taking and construct validity **210**

5.2 Test preparation, repeated test taking and candidates' proficiency **214**

5.3 Test preparation, repeated test taking and academic attainment **215**

5.4 Theoretical and practical implications 216

Chapter 6 Conclusion 219

Appendix 1 Information page and consent form for Quasi-experiment 1 (Intervention group) 221

Appendix 2 Information page and consent form for Quasi-experiment 1 (Control group) 223

Appendix 3 Questionnaire for Quasi-experiment 1.....................................225

Appendix 4 Information page and consent form for Quasi-experiment 2 228

Appendix 5 Questionnaire for Quasi-experiment 2.....................................231

Appendix 6 C-test for Quasi-experiment 2 233

Reference...235

List of Tables

Table 3.1.1 Correlation coefficients calculated using all participants' T1 performance on all linguistic measures used in quasi-experiment 1 (N=89) 115

Table 3.1.2 Test-retest reliability calculated using participants' T1 and T2 performance on all IELTS modules (N=44) 115

Table 3.1.3 Test-retest reliability calculated using participants' T1 and T2 performance on both OOPT sections (N=44) 115

Table 3.1.4 Results from reliability analysis for both versions of Spot-the-word and Speed-of-comprehension at T1 and T2 (N=89) 116

Table 3.2.1 The intervention and control participants' T1 and T2 IELTS overall scores and between-group differences 124

Table 3.2.2 The intervention and control participants' T1 and T2 IELTS listening scores and between-group differences 127

Table 3.2.3 The intervention and control participants' T1 and T2 IELTS reading scores and between-group differences 128

Table 3.2.4 The intervention and control participants' T1 and T2 IELTS writing scores and between-group differences 130

Table 3.2.5 The intervention and control participants' T1 and T2 IELTS speaking scores and between-group differences 132

Table 3.2.6 Comparison of The Intervention and Control participants' group means on IELTS (overall and by section), taken at T1 and T2 (N=89).... 135

Table 3.2.7 IELTS score gains (overall and by skill) achieved by intervention participants' grouped according to participants' initial IELTS scores (N=45) 136

Table 3.2.8 The intervention and control participants' T1 and T2 OOPT overall scores and between-group differences 137

Table 3.2.9 The intervention and control participants' T1 and T2 vocabulary and sentence processing accuracy scores and between-group differences .. 141

Table 3.2.10 Comparison of the Intervention and Control participants' group means on OOPT (overall and by section), vocabulary test, reading comprehension accuracysentence processing accuracy test, taken at T1 and T2 (N=89).......... 144

Table 3.2.11 The intervention and control participants' T1 and T2 time taken(measured in seconds).. 144

Table 3.2.12 Comparison of the Intervention and Control participants' group means processing speed test, taken at T1 and T2 (N=89)...................... 146

Table 4.1.1 Interpretation of DET Scores (Duolingo English Test, 2018) 171

Table 4.2.1 Descriptive statistics of participants' IELTS scores (overall and by skill) and their performance on Duolingo and C-test................................ 175

Table 4.2.2 Number of IELTS attempts by participants (N=153) 177

Table 4.2.3 Comparison of English proficiency scores between participants who attended IELTS-preparation programmes and those who did not (N=153) 178

Table 4.2.4 Participants' general proficiency (composite score) and their final IELTS scores in relation to their number of IELTS attempts (N=153)...... 179

Table 4.2.5 Regression model using final IELTS overall score, attendance of IELTS-preparation programmes and number of test attempts to predict the English proficiency of participants on entry to university (N=153) 183

Table 4.2.6 Participants' weighted average academic grades classification IELTS (N=153).. 184

Table 4.2.7 Regression model using IELTS overall to predict the academic grades for all participant, controlling for non-verbal intelligence and working memory (N=153)... 185

Table 4.2.8 Weighted average academic grades and final IELTS overall of participants from less and more linguistically demanding disciplines (N=153) 188

Table 4.2.9 Regression model using IELTS overall and disciplinary linguistic demand to predict the academic grades for participants, controlling for non-verbal intelligence and working memory (N=153)................................... 189

Table 4.2.10 Regression model using IELTS overall to predict the academic grades for participants of linguistically more demanding disciplines, controlling for non-verbal intelligence and working memory (N=80)........ 192

Table 4.2.11 Regression model using IELTS overall to predict the academic grades for participants of linguistically less demanding disciplines, controlling for non-verbal intelligence and working memory (N=73) 192

Table 4.2.12 Participants’ weighted average academic grades, IELTS in relation to their attendance at test preparation courses 195

Table 4.2.13 Participants’ weighted average academic grades, IELTS in relation to number of IELTS attempts made 195

Table 4.2.14 Regression model using final IELTS overall, test preparation attendance and number of attempts to predict the academic grades for all participants, controlling for non-verbal intelligence and working memory (N=153) 197

Table 4.2.15 Regression model using final IELTS overall, test preparation attendance and number of attempts to predict the academic grades for participants from linguistically more demanding disciplines, controlling for non-verbal intelligence and working memory (N=80) 198

Table 4.2.16 Regression model using final IELTS overall, test preparation attendance and number of attempts to predict the academic grades for participants from linguistically less demanding disciplines, controlling for non-verbal intelligence and working memory (N=73) 199

Table 4.2.17 Summary of participants’ admission minimum IELTS requirement (overall and componential), number of attempts made (N=153) 200

Table 4.2.18 Summary of participants’ age and length of English learning (N=153) 201

Table 4.2.19 Regression model using general proficiency to predict the academic grades for all participant, controlling for non-verbal intelligence and working memory (N=153) 204

List of Figures

Figure 3.1.1 Example for Task 1 in Use of English, OOPT: Testing knowledge of grammatical forms at the A2 level by Purpura (2009, p. 17) 109

Figure 3.1.2 Example for Task 2 in Use of English, OOPT: Testing knowledge of implied meaning at the A2 level by Purpura (2009, p. 18)....................... 109

Figure 3.1.3 Example for Task 3 in Use of English, OOPT: Testing knowledge of grammatical form and meaning across a passage at the A2 level by Purpura (2009, p. 19).. 110

Figure 3.1.4 Example of OOPT listening tasks: Understanding literal meaning at B1 level by Purpura (2009, p. 20) 111

Figure 3.1.5 Testing procedure for participant ID001 of quasi-experiment 1.. 118

Figure 3.2.1 The intervention and control participants' T1 and T2 IELTS overall score distribution and both groups' mean scores at two time points (N=89) 125

Figure 3.2.2 The intervention and control participants' T1 and T2 IELTS overall score distribution, and both groups' mean scores at two time points (N=89) 127

Figure 3.2.3 The intervention and control participants' T1 and T2 IELTS reading score distribution, and both groups' mean scores at two time points (N=89) 129

Figure 3.2.4 The intervention and control participants' T1 and T2 IELTS writing score distribution, and both groups' mean scores at two time points (N=89) 131

Figure 3.2.5 The intervention and control participants' T1 and T2 IELTS speaking score distribution and both groups' mean scores at two time points (N=89) 132

Figure 3.2.6 The intervention participants' average IELTS score change from T1 to T2 (overall and by skill) (N=45) (Error bars ± 1 standard error)........... 133

Figure 3.2.7 Control participants' average IELTS score change from T1 to T2 (overall and by skill) (N=44) (Error bars ± 1 standard error).................... 134

Figure 3.2.10 The intervention and control participants' T1 and T2 OOPT overall scores (N=89) (Error bars ± 1 standard error) 138

Figure 3.2.11 The intervention and control participants' T1 and T2 OOPT use of English scores (N=89) (Error bars ± 1 standard error) 139

Figure 3.2.12 The intervention and control participants' T1 and T2 OOPT listening scores (N=89) (Error bars ± 1 standard error) 140

Figure 3.2.13 The intervention and control participants' T1 and T2 Vocabulary test scores (N=89) (Error bars ± 1 standard error) 142

Figure 3.2.14 The intervention and control participants' T1 and T2 sentence processing accuracy scores (N=89) (Error bars ± 1 standard error) 143

Figure 3.2.15 The intervention and control participants' T1 and T2 time taken measured in seconds (N=89) (Error bars ± 1 standard error).................. 145

Figure 4.1.1 DET Vocabulary Task by Brentel & Settles (2017)..................... 160

Figure 4.1.2 DET Completion Task by Ye (2014).. 161

Figure 4.1.3 DET Listening Task by Ye (2014).. 162

Figure 4.1.4 DET Speaking Task by Ye (2014) 163

Figure 4.1.5 Example WASI Matrix Reasoning Item by (Wechsler, 2011) 167

Figure 4.2.1 Relationship between number of IELTS attempts and general proficiency at difference final IELTS overall level (N=149) 181

Figure 4.2.2 Percentage of participants achieving fail, pass, merit, distinction grades by their final IELTS overall (N=148).. 186

Figure 4.2.3 Visualisation of the relationship between IELTS and academic grades by group (N=153)... 191

Figure 4.2.4 Percentage of linguistically more demanding participants achieving fail, pass, merit, distinction grades by their final IELTS overall (N=79).... 193

Figure 4.2.5 Percentage of linguistically less demanding participants achieving fail, pass, merit, distinction grades by their final IELTS overall (N=69).... 194

Acknowledgement

This thesis is dedicated to my parents and my partner William, for their unconditional love and support. I am also grateful to Stephanie Shan, Kaiqi Hang, Jingjing Huang, Jia Li, Haoruo Zhang and all my friends for their unreserved encouragement and undivided faith in me, even in the darkest of times.

My sincere gratitude goes to my supervisor, to whom I am forever indebted, for her rigorous feedback and continuous guidance.

Author’s declaration

I hereby declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other university. All sources are acknowledged as references.

Chapter 1 Introduction

For students whose first language is not English, English proficiency is crucial for the achievement of success in English speaking higher education (Trenkic & Warmington, 2018).

With the rapid internationalisation of higher education, the number of international students enrolled in English-speaking higher education institutions has increased substantially (HESA, 2018). UK, as one of the most popular destinations for international students, has witnessed a huge growth of international student population, in particular, the Chinese international student. With this increase in quantity, the academic attainment of international Chinese students has caught the attention of many researchers (e.g. He & Banham, 2009; Morrison, Merrick, Higgs & Le Métais, 2005; Paton, 2007).

In comparison with home students, previous research suggests that international students achieve less academic success and are at higher risks of failing their academic study (e.g. Morrison et.al., 2005; Paton, 2007). Although there are many factors contributing to their lower academic attainment, English language proficiency has been frequently argued as the key. International Chinese students have reported problems when it comes to following lectures (Du-Babcock, 2002), participating in classroom activities (Sun & Cheng, 1999), comprehending extensive academic texts (Trenkic & Warmington, 2018) and writing academic assignments (Angelova & Riatzatseva, 1999).

Given the important role proficiency plays and the hindering impact lack of proficiency has on students' academic attainment, it is common for the receiving universities to set a language requirement, often indexed by a certain score on a language proficiency test. For English-speaking higher education institutions, the two most commonly acknowledged tests are IELTS (the International English Language Testing System) and TOEFL (Test of English as a Foreign Language). As there is often a lack of rationale justifying the setting of language proficiency requirement, this threshold can be interpreted in various ways. On the one hand, one could say that it is set at a point where students who meet this requirement can benefit from studying at this institution at the level they desire, although their academic attainment may still be constrained by their language proficiency. On the other, it is also possible that this threshold is set to ensure that universities are only accepting international students whose academic attainment is no longer constrained by their language proficiency; this is to say that the threshold is there to warrant international students have acquired sufficient language skills to achieve the optimal academic attainment. Although either can be justified, this thesis stipulates that the latter justification for setting a language requirement is "correct" taking into account the high stakes of studying abroad for the majority of international students.

Given this variation in interpretation, the relationship between language proficiency and academic attainment also differs. Were the former argument to be the rationale for threshold setting, then a linear relationship between proficiency (indexed by test scores) and academic attainment would be expected whereas were the latter to be the case, this relationship would cease to exist. Taken together, this study hypothesises that the relationship between attainment and proficiency is only linear at lower levels and plateau out at a certain cut-off point. This cut-off point, according to IELTS Guide for for educational institutions (2015), is likely to be a minimal of overall 7.5 for students engaged in linguistically more demanding disciplines and 7 for students engaged in linguistically less demanding disciplines. Therefore, this thesis stipulates that an IELTS overall of 7.5 (or the equivalent scores in other tests) is the "correct" level of language requirement for students from linguistically more demanding disciplines and 7 is "correct" for those linguistically less demanding disciplines. Meanwhile, it should also be highlighted that this "correct" cut-off score may not necessarily be in line with the minimal language requirement set by the receiving institutions considering the lack of explicit rationale as discussed above.

On the note of the predictive validity of proficiency test scores, a review of relevant literature shows mixed results. On the one hand, some researchers reported insignificant correlation between proficiency and academic attainment (e.g. Dooey & Oliver, 2002; Kerstjens & Nery, 2000); thus claims, such as academic attainment at higher education level was more closely related with non-linguistic factors such as differences in learning styles, are made. On the other hand, many studies have found evidence showing language proficiency is indeed a hinderance obstructing international students' academic attainment and in many cases, scores on proficiency tests are good predictors for subsequent academic attainment (e.g. Daller & Xue, 2009; Daller & Phelan, 2013; Feast, 2002; Trenkic & Warmington, 2018).

Concerning the inconsistency in predictive research findings, the following questions are raised: why is language proficiency predictive in some cases but not in others? Could it be that in some cases the institution set the requirement correctly while others did not? Could it be related to the way the data were analysed? Or could it be because in some cases the validity of the proficiency measure has been eroded by other factors that are rarely explored in previous literature?

In the present study, an attempt to answer these questions and shed light on the inconsistent predictive power of IELTS was made through examining the effects of test preparation and repeated test taking on candidates' scores and general proficiency.

Test preparation, also known as coaching, refers to practices and procedures specifically undertaken to improve scores, either through improving the skills and abilities measured by the test (i.e. the construct), or by improving the skills for taking the test (i.e. testwiseness), or both (Cohen, 2014; Cronbach, 1971; Messick, 1981). The increasing popularity of test preparation, especially for proficiency tests such as IELTS, is closely related with test impact and washback, i.e. influence tests assert on teaching and learning (e.g. Alderson & Wall, 1993; Baker, 1991). Abundant literature has been dedicated to the description and discussion of how high stake tests such as IELTS alter the way teachers teach and learners learn in the process of test preparation (e.g. Erafni, 2012; Green, 2006). Most of the research in this strand adopted a qualitative and/or observational approach (e.g. Mickan & Motteram, 2008), and the findings are largely based on self-report data elicited from a small research sample.

In comparison, fewer efforts have been made to look at how test preparation affected scores on proficiency tests (e.g Bagheri & Karami 2014; Gan, 2009; Green, 2007; Issitt, 2008, Xie, 2013). Among studies that can be found, findings are limited by 1) small sample size, 2) incomprehensive research design (e.g. lack of control group to set the baseline, or lack of pre-test to account for candidates' pre-existing proficiency), 3) limited research scope (i.e. only looking at one specific componential module of the test). The fact that most of these preparation studies were situated in English as a native language context (e.g. Green, 2007) further complicates the interpretation of findings, because the reported scores gains might simply be the results of frequent English exposure, not dedicated test preparation.

Setting these limitations aside, based on the review of test preparation literature, the following consensuses can be reached: a) high stake proficiency tests such as IELTS do affect teaching and learning, and b) test preparation are fairly effective in terms of inducing score gains. What previous research fail to examine is how such score gains are achieved through test preparation. In other words, are score gains a reflection of improved English proficiency? or simply, increased testwiseness? Answering these questions has important implications on the predictive validity of IELTS, and the interpretation and use of IELTS scores in academic contexts.

In addition to test preparation, it has been noted that many candidates repeat test taking until they arrive at the desired scores (e.g. Ma, 2014; Ma & Cheng, 2015; Zhang, 2008), which begs the question: can scores be boosted simply through the act of sheer repetition? If so, what implication can this have on the reliability and validity of the test? Likewise, answers to these two questions cannot be found from existing literature.

To bridge these aforementioned gaps in existing literature, the present study made an empirical attempt to examine whether test preparation and repeated test taking could boost candidates' scores onto a level that is beyond their general proficiency and the implication this had on test construct validity and predictive validity in the context of IELTS. This study also set out to re-examine the role language proficiency played in international students' academic attainment at a UK university, taking into account the effects of test preparation and repeated test taking. With these aims, the following five research questions were put forward:

- To what extent does IELTS preparation course improve candidates' IELTS scores (overall and by skill)?
- To what extent does IELTS preparation course improve candidates' general proficiency?
- To what extent does repeated IELTS taking affect Chinese candidates' IELTS scores and their general proficiency?
- To what extent does IELTS predict international Chinese students' academic attainment at a UK university?
- Have test preparation and repeated test taking eroded the validity of IELTS as a predictor for academic attainment?

To answer these questions, two quasi-experiments were designed and conducted, one focusing primarily on the effects of test preparation, and the other on the combined effects of repeated test taking and test preparation. Together, this study intended to enrich the literature on the effects of test preparation and repeated test taking and clarify the role of proficiency in the achievement of academic success.

This thesis includes six chapters. Following this Introduction chapter is Literature Review, where previous research on international students' academic attainment, role of language proficiency, standardised proficiency tests and test validity, test preparation and repeated test taking are reviewed and discussed. Also in this chapter, research gaps are identified and research contexts are presented. Chapter 3 includes the methodology, results and discussion from the first quasi-experiment, which looked at the effects of test preparation while Chapter 4 presents the methodology, results and discussion from the second quasi-experiment, which examined the combined effects of test preparation and repeated test taking and predictive validity of IELTS. Chapter 5 is General Discussion where findings from both quasi-experiments are looked at together and linked back to the literature reviewed in Chapter 2. The last chapter summarises the key findings from the present study, puts forward the limitation and recommendation for future research, and finally concludes the contribution of this thesis.

Chapter 2 Literature Review

This chapter reviews literature relevant to the present study, starting with an overall description of academic attainment of international students enrolled in English speaking higher education institutions (2.1), setting the scene for the present study. The chapter then moves on to discuss problems this population faces and how these problems are accelerated by the lack of sufficient English language proficiency (2.2), followed by a brief discussion of the theoretical framework and relevant empirical findings underpinning the concept of second/foreign language proficiency (2.3). Following this, popular measures developed to provide an accurate indication of English proficiency are presented and discussed (2.4), focusing primarily on two standardised English tests that are widely accepted by English-speaking higher education institutions: IELTS (the International English Language Testing System) and TOEFL (Test of English as a Foreign Language). The validity and reliability of both tests are subsequently presented and debated, drawing reference from existing IELTS and TOEFL research studies (2.5). Given the role IELTS and TOEFL plays in English speaking higher education institutions, this chapter narrows the discussion of validity to focus mainly on the predictive validity of IELTS and TOEFL and probes deeper into the relationship between language proficiency as measured by standardised language tests and international students' academic attainment through critically reviewing relevant predictive validity literature (2.5.5-2.5.6). To account for the inconsistency in previous research findings regarding the predictive role of proficiency measured by proficiency tests, two factors are proposed, namely: test preparation and repeated test taking. Following this, research on test preparation and repeated test taking is reviewed (2.6-2.7), putting forward limitations among existing research this present study aimed to overcome and the gap this study set out to bridge. This chapter ends with a section explaining the context of the present study and the research questions proposed to be answered (2.8).

2.1 International students in English-speaking higher education institutions

Travelling overseas for academic purposes has become an overwhelming global trend. UK, as one of the most popular destinations in the world for international students, has witnessed a rapid increase in the total number of enrolled non-UK higher education (hereafter, HE) students, which more than doubled from 185,630 in 2001/02 to 442,375 in 2016/17, accounting for approximately one fifth of all students enrolled in all UK HE programmes and over half of all students enrolled in full-time postgraduate programmes 2016/17 (Higher Education Statistics Agency (hereafter HESA[1]), 2018; Universities UK international, 2017).

In the last decade, the most dramatic growth has been in the number of students from China, increasing from 25,000 in 2006/7 to 66,000 in 2016/7. Since 2012/13, each year, the number of students from China has far exceeded the number of students from all EU countries combined (Migration Advisory Committee, 2018). Furthermore, HESA data (2018) show that one in five non-UK domiciled students comes from China.

In the UK, this growth of the international student population, non-EU-domiciled student in particular, brings along not only huge economic benefits but also employment opportunities. Higher Education Policy Institute and Kaplan International Pathways (2018) reported that for the total economic impact per international student on UK economy, £68,000 was associated with every EU-domiciled student in the 2015/16 cohort, and £95,000 was associated with each non-EU-domiciled student of the same cohort. Oxford Economics (2014) reported that, in 2013, international students generated an estimated £713 million (from fees paid) gross value added (GVA), £123 million (from living expenditure) GVA and also £53 million (from visitor spending) GVA. The total GVA was about £890 million, estimated to support almost 23,000 jobs, and generate tax revenues of £385 million. There is no doubt that the increase in the international student population has its benefits, but at the same time, with this rapid growth come a series of concerns.

2.1.1 How are international students performing academically?

Although the overall population of international students enrolled in English speaking higher education institutions (hereafter HEIs) keeps growing, literature documenting this group's academic performance in HE institutions and literature on this topic is insufficient. The studies that are there (e.g. He & Banham, 2009; Morrison et.al., 2005; Paton, 2007) suggest that, in general, international students do not perform academically as well as home students. Using time series data, He and Banham (2009)

[1] HESA is the official agency for the collection, analysis and dissemination of quantitative information about higher education (HE) in the UK (https://www.hesa.ac.uk/) (HESA, 2018)

looked at the academic attainment of international students and home students enrolled in a Canadian HEI over a six-year time period (2002/03 to 2007/08). Results indicate that over the six years of study, international students continuously achieved a lower average grade than Canadian home students, although the between-group gap was narrowed over time. In other cases, even when international students manage to outperform home students at the beginning of their academic study, they often end their study being outperformed. For example, Zhao, Kuh and Carini's (2005) comparative study on the effective educational practice engagement of international students (N=2780) and home students (N=67072) in the U.S. showed that international students did begin their study with better engagement with academic work and made greater personal and social development than American home students, measured through self-reports. However, this competitive edge was lost over time and international students were eventually surpassed by the home students in active and collaborative learning.

Similar patterns have been observed in Australia. Through probing international students' pass/fail rate at an Australian university in 2003, Paton (2007) found that international students (N=1926) had fewer high achievers, as indicated by percentages of high distinctions grades(3.2%) and distinctions grades (12.4%), when compared to Australian home students (7.5% of whom achieved high distinctions and 18.8% achieved distinctions, N=1831). Not only were international students less likely to become academic high achievers, their percentage of failing their study was also higher; 14.1% of whom failed their study while only 10.8% of home students did so.

Similar findings were reported in the UK, where the present study is situated in, by Morrison et al. (2005) who scrutinised the class of degree obtained by undergraduate students of different nationality between 1995 and 2000 using centrally collected HESA data. Initial analysis showed that, in line with He & Banham (2009), Smith and Eccles (1993) and Makepeace and Baxter (1990), international students in the UK also achieved fewer first or upper-second-class honours, which are commonly considered as good degrees and prerequisite for securing good employment upon graduate, than UK home students. Further analysis revealed that students from European Union (EU), Asia (including China, the biggest population of non-EU international students studying in the UK), Africa and the Middle East performed less well than UK home students (Morrison et al. 2005).

Taken together, findings from studies conducted in Canada, Australia, and UK, showed that international students, especially those from non-western backgrounds, were achieving less academically than local home students in HEIs where English is spoken as the language of instruction, This, naturally, begs the question of why international students are not performing academically as well as home student? What could have contributed to their lower academic attainment? These were the two questions this present study aimed to shed some light on.

2.1.2 International Chinese students[2] at English speaking HEIs

As the largest international student population, Chinese students' struggle to achieve academic excellence has also been frequently documented in existing literature. Also using HESA data, Iannelli and Huang (2013) looked at the academic attainment of undergraduate and postgraduate Chinese students in the UK between the time period of 1998/99 and 2008/09. Despite the growth in the number of international Chinese students obtaining their first degrees from UK HEIs, the likelihood of these students gaining a good degree (i.e. 1st and 2:1) was about a third of that of UK home students in 2001 and this discrepancy continued to widen until 2009. Moreover, the likelihood of international Chinese students being awarded with first-class degrees was constantly lower when compared to UK home students, EU students, non-EU students and even other Asian students during the observed time period. Meanwhile, as the percentage of lower-second-class degree dropped from 50% to 43%, the percentage of international Chinese students graduating with third-class degree increased accordingly, from 14% to 21%.

Iannelli and Huang's findings are backed up by a smaller scale longitudinal study conducted by Crawford and Wang (2014) in a UK university, which compared the difference in academic attainment between international Chinese students enrolled in a UK university (N=52) and UK home students (N=60) (enrolment year 2006/07 and 2007/08). Analyses of overall academic grades showed that 80% of UK home students graduated with a first or upper-second-class degree while only 43% of Chinese students were able to achieve similar results; additionally, the percentage of international

[2] In the present study, the term *international Chinese student* refers to those who grew up in mainland China, received their education in mainland China with Mandarin Chinese as their first language and later travelled to a different country for academic purposes.

Chinese students unable to make progress in their second or their year of study (15%) almost doubled that of UK home students (8%). Interestingly, these international Chinese students outperformed their UK counterparts significantly and achieved a higher yearly average mark in their first year of study, but over time, this pattern reversed. International Chinese students underperformed UK students in the second year, achieving a lower yearly mark average and this gap continued to widen in the third year. Many factors could have contributed to this loss of competitive edge. For example, Crawford and Wang (2014, 2015) claimed that first year university learning was mostly associated with surface learning, which most international Chinese students were skilled at. As they progressed further into their study, learning in the later years of university education demanded "deep and strategic learning approaches which Chinese students fail to develop" (Crawford & Wang, 2014 p.917). However, although many cross-culture studies (e.g. Holmes, 2004; Ma, 2014; Spencer-Oatey & Xiong, 2006) have noted international Chinese students' lack of deep and strategic learning approaches, others (e.g. Chalmers & Volet, 2014; Cooper, 2004; Kember, 1996) have argued that Chinese students are, in fact, equally skilled and their low academic attainment is more likely to be affected by other factors, the most obvious being international Chinese students' proficiency in the language of instruction, i.e. English. It is possible that international Chinese students were able to perform better at the beginning of their study because the academic linguistic demand was comparatively low; for example, they only needed to read a small body of literature or write short essays. However, as they progressed further into their study, the amount of reading and the demand to produce written work of higher complexity increased. If these linguistic demands were above the level the students' existing language capability, academic attainment could be compromised. In fact, international Chinese students' lack of sufficient English skills has been discussed in a number of research studies (e.g. Mori, 2000), which are discussed at length in a later section of this chapter (2.2.3).

On the other side of the Pacific, international Chinese students' academic attainment in English HEIs also caught the attention of American researchers. For example, Ma's study (2014) showed that a high proportion of international Chinese first-year students (N=175) enrolled at an American university was low achieving or at risk for academic failure. Although international Chinese students significantly attempted and earned more credit hours than their American counterparts, their cumulative first-year Grade Point Average (hereafter, GPA) were lower than their American counterparts. In addition, comparison between first-year international Chinese students' and other first-year international students' attempted credit hours, earned credit hours, cumulative GPAs and persistence rate[3] showed that the latter outperformed the former in all of the four measures examined in Ma's study. To explore reasons underpinning such discrepancies, interviews were conducted with the international Chinese students (N=26). Interview results revealed that international Chinese students' low attainment could be attributed to insufficient preparation for studying abroad, learning skills and motivation and more importantly, their lack of sufficient English skills. This lends support to the afore-stated hypothesis that language plays a key role in the academic attainment for this particular population.

Given the growing number of international students enrolled in English speaking HEIs, it is of significance to look into the factors that may contribute to their low academic achievement. Compared to the large amount of research on the factors affecting home students' academic achievement (e.g. Kim, Newton, Downey & Benton, 2010; Kuh, Cruce, Shoup, Kinzie & Gonyea, 2008; Kuo, Hagie & Miller, 2004; McKenzie & Schweitzer, 2001), there is considerably less literature on determinants for international students' academic performance (e.g. Andrade, 2005; He & Banham, 2009; Pelletier, 2003). Although many variables have been put forward and many models have been built to account for this difference in academic attainment between home and international students, due to the great variation in demographics, previous education background and etc., it is not realistic to expect that a single model with certain predictors could apply to all groups of international students with different characteristics. Hence the following sections only concentrate on factors/determinants that are frequently debated among the existing literature.

2.2 Factors affecting academic performance of international students and the role of language proficiency

2.2.1 Sociocultural and psychological challenges

Academic adjustment issues are almost unavoidable for both home and international students, especially during the transitional period at the beginning of their study. Given that most international students have travelled a great distance to study in a foreign

[3] Here persistence rate refers to whether a participant manages to progress and become reenrolled from 1st year to 2nd year

country (e.g. from the East to the West), it may be more difficult for them to receive sufficient and timely support from their family and friends than home students (Andrade, 2006; Zhou, Jindal-Snape, Topping & Todman, 2008). Russell, Rosenthal & Thomson (2010) found that 41% of international students (N=900) in Australia experienced substantial levels of stress, often as a result of homesickness, cultural shocks, or perceived discrimination. This high level of stress may also be language related; for example, lack of sufficient proficiency may hinder international students' friendship building with local home students or other international students, further adding to their feeling of homesick. Likewise, the perceived cultural differences and discrimination could also be related to misunderstanding during communication if international students failed to grasp the meaning of colloquial expressions and slang because of their insufficient language proficiency.

In particular, using survey data, Ward & Masgoret (2004) reported international Chinese students (N=2659) to be not as engaged when interacting with their New Zealand peers both in academic contexts and social occasions, compared to other international students. As stated earlier, because of differences in culture and more importantly, lack of confidence in their English communication skills, most Chinese students did not develop friendship with New Zealand home students. Consequently, they were more likely to feel culturally excluded in New Zealand classrooms and discriminated against by host nationals (Ward & Masgoret, 2004). Similar findings were reported in the US; through interviews, Shu (2008) found that although the international Chinese students (N=6) viewed their overall study experience in a foreign country as "meaningful and worthwhile" they did experience "homesickness, friendlessness, and lack of sense of belonging" (p. 76), most of which could be related to their language skills.

Although research indicates developing new friendships with home students may compensate this aforementioned lack of social support, international students often find this challenging due to lack of opportunity and/or preference for friendships with students of their own nationalities (Hawthrone, Minas & Singh, 2004). Of course, the choice of establishing friendship is a personal and cultural decision, but one cannot deny that language proficiency also has a role to play, as establishing friendship with people from other nationalities often demands both parties to be proficient in the shared language. If one party fails to meet such a demand, the chance of establishing friendship is further reduced. Moreover, this frequently reported experience of loneliness and homesickness may lead to mental health problems and negatively affect international students' academic performance (Mori, 2000; Rajapaksa & Dundes, 2002), which are difficult to resolve especially when effective communication could not be achieved due to the lack of language skills. In other words, these aforementioned social adjustment issues might be, a large extent, related to international students' low level proficiency and their confidence to communicate using a foreign language. From a long-term perspective, a "vicious" cycle could be formulated: the lack of proficiency leads to embarrassment during communication breakdowns, negatively affecting international students' willingness to communicate using the foreign language, thus avoiding the opportunity to communicate and henceforth unable to further improve their language skills. They subsequently "retreat socially to the community in which they are most comfortable" (Huntley, 1993, p. 10), often among those who share the same first language.

For international Chinese students in particular, literature on how lack of English proficiency could profoundly affect their academic and social adjustment is not abundant. Nonetheless, Feng's qualitative study (1991) lent support to the cycle presented above as interviewed participant (N=52) commented they felt embarrassed when others asked "Pardon me?" "Could you say it again?". Further, it was commented that after experiencing such embarrassments on several occasions, international Chinese students would rather avoid communication at all. Similar viewpoint was put forward in a later qualitative study by Zou (2000), who reported that because of the low English proficiency and the lack of confidence in their language skills, in some extreme cases, international Chinese students became "afraid to meet people" (pp. 191-192). Similar behaviours were observed in academic settings as Sun and Cheng (1999) pointed out that many Chinese students never achieved full participation in American classrooms as they lacked language skills, listening and speaking in particular, to accurately comprehend the proposed questions or to actively voice their opinions through discussion.

2.2.2 Differences in learning environment and learning style

When international students travel to a different country to study, chances that they experience academic culture shock, "a case of incongruent schemata about higher education in the students' home country and in the host country" (Gilbert, 2000, p. 14) are very high, particularly upon their arrival. This academic cultural shock is closely associated with the learning environment of an academic institution, including the education system, lecture style, assessment, relationship between students and lecturers, and could cause subsequent academic adjustment issues that influence adversely on international students' academic performance. Rienties et al (2012) reported academic success of international students of non-western backgrounds in a Dutch university was primarily determined by academic integration, in particular by the degree of academic adjustment, which could be related to the language proficiency of international students.

As noted earlier, international Chinese students exhibit inactive classroom behaviour as they make less effort in participating in group discussions or debating in class and do not like to raise or answer questions (Li, Duanmu & Chen, 2009; Parker, 1999). Although this lack of active engagement could be attributed to the beliefs and preference international students have established through their previous education (Ma, 2014), i.e. academic cultural shock, it could also be relevant to their lack of confidence in their spoken language skills as international Chinese students in American universities pointed out their concerns over their pronunciation and their ability to speak fluently as well as express their ideas clearly in front of their classmates (Ma, 2014; Robertson, Line, Jones & Thomas, 2000). The preferred western teaching styles, which include lectures with individual student participation or lectures with group discussion (Beishline & Holmes, 1997), often assert high demand on student's listening and speaking skills that many international Chinese students may not have. It has been frequently reported that listening to lectures was difficult for international students due to the accent and speed of lecturers' spoken English, the choice of vocabulary and example, idiomatic styles, humour and choice of examples, vocabulary and the speech speed (Holmes, 2004; Ramsay, Barker & Jones, 1999; Robertson et al., 2000). Given these reported difficulties, one may say that language proficiency has a role to play in international students' reported lack of active classroom engagement. Although it is possible that the deeply rooted teacher-as-the-authority notion has somewhat prevented international Chinese students from challenging what the teacher has said, and encouraged note-taking and content-memorising (Gu, 2009), it is also probable that passive following, note-taking and memorisation is linguistically less demanding than voicing one's opinion on the spot during a heated discussion while comprehending what others have said. In a similar fashion, instead of arguing studying in western HEIs is challenging for international Chinese students because the academic environment is more characterised by independent learning and less instructor supervision and guidance than the traditional Chinese learning approach (Smith & Smith, 1999), the present study takes a different stand and proposes that these challenges may have be elevated by insufficient language proficiency.

In addition to differences in teaching and learning styles, classroom participation, teacher and learner interaction, international Chinese students may also find the assessment process of English speaking HEIs drastically different from their previous education. As frequently documented, the main, if not the only assessment in Chinese education is through exams (e.g. Yu & Suen, 2005; Kirkpatrick & Zang, 2011); however, in most western HEIs, a combination of assessment methods is often used. In addition to exams, students are often assessed through written tasks (e.g. essay, dissertation) or group projects, which not only assesses their knowledge of learning content, but also their writing skills.

Among the four language skills, English writing seems to be the most demanding aspects of academic language for international students (Angelova & Riatzantseva, 1999; Robertson et. al. 2000). To write in an unfamiliar academic discourse requires not only the conceptual understanding of the way of writing and of the meaning of using literature to develop written argumentations, but more importantly, a good command of linguistic knowledge of the target language, e.g. vocabulary and grammar (Gu & Brooks, 2008). International Chinese students are accustomed to indirect writing styles and unfamiliar with analysing the strengths and weaknesses of an argument, both of which could be considered features of academic writing in English speaking HEIs. This difference in writing styles and the demand for critical thinking further elevate the demand for excellent writing skills, because it serves as the premises of good academic writing, without which ideas could not be conveyed. Meanwhile, in more recent studies, the problem of plagiarism and Chinese students has attracted the attention of a growing number of researchers (e.g. Pennycock, 1996; Shei, 2005). Although most studies have explored this phenomenon from culture related perspectives, others also pointed out that plagiarism could also be a result of Chinese students' lack of adequate language skills. Edwards & Ran (2006, p.10.), for example, asserted that Chinese students "simply do not have sufficient command of English to explain what an author says in their own words. They are limited by their vocabulary and probably by their grammar as well".

2.2.3 Language as the main obstacle

The above review of the literature indicates that language proficiency is, in many cases, at the heart of the struggles and challenges faced by many international students. Many of the aforementioned issues and challenges international students experience could be, to various extents, magnified by their lack of sufficient language proficiency. For students whose first language is not English enrolled in English-speaking HEIs, the language barrier is regarded as “the most significant, prevalent problem” (Mori, 2000, p. 137), even for those who have gained admission to elite universities. For example, in a leading Australian university, approximately one quarter of the international undergraduate students (N=910, out of which 324 were international Chinese students) enrolled in the faculty of Accounting were regarded as not having acquired satisfactory academic English skills, based on the university’s Measurement of Academic Skills of University that required students to write a short essay or other genre (e.g. report) based on disciplinary content. For postgraduate students (N=278), it was worse; 88% of international postgraduate students (97% of whom were Chinese) in the faculty were considered to have unsatisfactory academic English skills (Paton, 2007).

Linking this lack of language skills back to international students’ social and academic adjustment issues, intuitively, one can assume that only when a certain level of proficiency has been achieved can international students combat these problems successfully. In line with this assumption, Senyshyn, Warford and Zhan’s (2000) research showed that international students who had higher proficiency upon admission, indexed by high TOEFL (Test of English as a Foreign Language) scores, experienced fewer adjustment difficulties, had more positive experiences and felt more satisfied with their academic progress than those with lower TOEFL scores. This provides further support for the notion that language proficiency could be the key to mitigating the negative impact of adjustment issues.

With regards to international Chinese students in particular, researchers have repeatedly asserted that the lack of proficiency is particularly concerning. This concern was also shared by Chinese students themselves (Sun & Cheng, 1999; Wan, 2001, Yuan, 2011). Even for those who obtained way above the required scores in standardized English tests e.g. Test of English as a Foreign Language (TOEFL) and International English Language Testing System (IELTS), when they arrive in the U.S., they were quickly faced with language-related problems. For example, international Chinese students enrolled in a U.S university stated that oftentimes, they were not able to gain a full understanding of the learning content (Edwards & Ran, 2006; Ma, 2014). Moreover, Du-Babcock's (2002) Hong Kong study (where English is stipulated as the language of instruction at HE level) found that Chinese students were not equipped with adequate listening skills to "follow the flow and missed nuances of normally paced lectures" (p.81), even when they had high enough scores to gain admission. This discrepancy between high enough scores on the required test and having sufficient language skills to communicate in or outside the academic setting leads to the reasonable speculation that maybe alternative means have been used to score high in the test, such as attending dedicated teste preparation courses or repeated test taking.

This speculation was confirmed by international Chinese students interviewed in Ma's study (2014). They commented that they attended dedicated test preparation courses with the sole purpose of getting ready for the test and these preparation courses were not perceived helpful in terms of enhancing their communication skills or their general proficiency. As more detailed discussion regarding test preparation and its effects in Chinese EFL contexts are provided in later sections of this chapter (section 2.6 and 2.7), here attention should be paid to significance of researching test preparation enabling Chinese students to achieve the required test score but not improving their overall communication skills, because it indicates the possibility that dedicated test preparation could be one of the key determinants associated with the low academic achievement of international Chinese students in English-speaking HEIs. However, so far, no study has been done to empirically examine this possibility.

Turning back to the discussion on international Chinese students' lack of adequate language skills, more research has found that reading in the context of western university education creates fundamental problems for Chinese students, both in terms of their reliance on the teacher for guidance as to what they should be reading and in terms of the sheer volume of reading recommended by lecturers (Edwards & Ran, 2006). International Chinese students not only read at a much slower pace than home students, they also need to read the same text multiple times in order to achieve a good understanding (Holmes, 2004; Ma, 2014; Trenkic & Warmington, 2017). Technical words and concepts, and terminology, are found particularly problematic (Gu & Maley, 2008; Lebcir, Wells & Bond, 2008). In addition, problems were present not only in the reading of academic literature and the contents of slides used during lectures, but also in comprehending examination tasks and assignment titles, both associated closely with their academic outcomes.

As stated earlier, international Chinese students are aware that the level of their language proficiency may be one of the most serious problems they face in cross-cultural learning (Mori, 2000). This issue becomes particularly clear when international Chinese students attend classes with native English students (Raymond & Parks, 2004), forcing them to reassess their English skills and later come to the realisation that even although they have managed to achieve the 'on-paper' language proficiency (i.e. they achieved the required language test scores), their communicative competence was still lacking. They were also aware of their comparatively less active participation in classroom discussion and their less efficient communication with their academic supervisors as well as their peers, although their English test scores seem to indicate that they should be capable and linguistically ready (Mori, 2000). Such discrepancy between test scores and communicative competence is one of the key areas the present study aimed to explore.

In a nutshell, there is no denying that language comprehension and competence are at the heart of difficulties for international students (Robertson et. al. 2000) and are fundamental for obtaining academic success, especially at higher education level (Paton, 2007). Given the detrimental effects insufficient language proficiency could pose on international students' social and academic integration and success, as denoted in the afore-discussed literature, and to prevent international students from failing their academic study due to the lack of sufficient language proficiency, it is common for the HEIs to set a language requirement as part of the admission criterion to screen potential applicants. To ensure the fairness of the screening process, most HEIs resort to standardised proficiency tests for the sake of reliability and validity. The two most widely recognised tests developed for such purpose are IELTS (the International English Language Testing System) for HEIs in UK, Australian, New Zealand, and TOEFL (Test of English as a Foreign Language) for HEIs in Canada and the US.

Before this chapter proceeds to elaborate on the development and the current use of IELTS and TOEFL, the following sections briefly present theoretical debates regarding the understanding of proficiency, as the conceptualisation of proficiency underpins both the consequent development of these two tests and the methods used in this present study.

2.3 Understanding L2 proficiency

2.3.1 Earlier frameworks(1960s-1980s): approaching proficiency as a general construct

Defining proficiency as a construct is not a trivial matter as the way how proficiency is understood relates closely to the way how proficiency is assessed. Since the 1960s, many L2 proficiency models have been put forward and accordingly, many measures and tests have been proposed. Among them, earlier models from Lado (1961) and Carroll (1961) consisted mainly of two components, one being linguistic knowledge (e.g. lexical, morphological, syntax and phonological knowledge) and the other being the four language skills (e.g. listening, reading, writing and speaking). What was missing from these earlier proficiency models was the realization that people listen, read, write and speak for the purpose of communication (Hulstijn, 2015). This is to say that earlier models did not take into account the communicative situations where such linguistic knowledge and skills were needed, i.e. learners' communicative competence was largely neglected. In accordance with this, proficiency tests developed at that time focused primarily at assessing learners' linguistic knowledge; for example, the original Test of English as a Foreign Language (TOEFL). When TOEFL made its first debut, it consisted only of multiple-choice items assessing vocabulary, reading comprehension, listening comprehension, knowledge of correct English structure and grammar. This test format corresponded with the mainstream linguistic theories at the time that considered language proficiency as a notion compromised of various components such as grammar, vocabulary and comprehension (e.g. Carroll, 1961; Lado, 1961).

To compensate the lack of attention to learners' ability to use linguistic knowledge for real life communicative purposes, Hymes (1972) proposed the notion of communicative competence to encompass knowledge of both the linguistic knowledge of a language itself and the appropriate use of such linguistic knowledge. From then on, the achieving of communicative competence has become the key goal of many second/foreign language learners. Later, communicative competence framework became the guiding principle upon which IELTS was built.

When communicative competence was first put forward, the concept remained rather vague; hence research was needed to probe deeper into its constitution. Building upon Hymes' theory, Canale and Swain (1981) developed their own communicative competence framework, specifying that to achieve communicative competence, one needs to acquire the grammatical, sociolinguistic and strategic competence of a language. In Canale and Swain's model (1981), grammatical competence includes lexical knowledge, morphology, syntax, sentence-grammar semantics and phonology; sociolinguistic competence encompasses both sociocultural and discourse rules; and strategic competence contained grammatical as well as sociolinguistic strategies. With the development in theoretical conceptualisation of communicative competence and proficiency comes the quest of how to measure L2 proficiency components, the focus of many language and education researchers (e.g. Bachman & Palmer, 1982)

With the notion of communicative competence broadening the understanding of proficiency, Bachman and Palmer (1982) set out to examine the three components that they assumed to constitute communicative competence: "grammatical (morphology and syntax), pragmatic (vocabulary, cohesion and coherence), and sociolinguistic competence (distinguishing registers, nativeness and control of non-literal, figurative language and relevant cultural allusions)" (p. 450) using a multi-components multi-methods research design among 116 L2 speakers of English with various background and diverse English learning history. To measure the competence in each of the three hypothesized components, four methods were used: interview, writing sample, a multiple-choice test and a self-rating. Results from confirmatory factor analyses rejected Bachman and Palmer's proposed structure of communicative competence, highlighting the need for more research effort to form a better understanding of the concept and its measurement.

2.3.2 Recent studies (1990s-2010s): approaching proficiency from four language skills

Fast-forwarding proficiency research to the late twentieth century, a series of language research projects were commissioned in Amsterdam to further explore the constitution of proficiency (e.g. Andringa, Olsthoorn & van Beuningen, 2012; Schoonen, Hulstijn & Bosser, 1998; Schoonen, van Gelderen, de Glopper, Hulstijn, Simis, Snellings & Stevenson, 2003). Different from Bachman and Palmer who approached proficiency as a whole, these Amsterdam projects focused on examining componential structure of each language skill (i.e. listening, reading, writing and speaking) separately. Findings of these studies are relevant to the way how language proficiency is understood and assessed nowadays, and are closely related to the methods and measures adopted in the present study.

For foreign language listening, two processes are evoked: decoding, i.e. the process of matching, assembling and identifying received acoustic input using learners' existing lexical knowledge, and meaning building, i.e. the processing of drawing inferences or schemata knowledge of the world and the topic of conversation (Field, 2008). The former often takes place at a local level while the later occurs a global level. For a learner to be proficient in listening, the need for accurate and automatic decoding is highlighted (Vandergrift, 2004) and the importance of linguistic knowledge, processing speed of linguistic information, and general cognitive ability has been explored (Andringa et al, 2012). Using data collected from 113 L2 Dutch learners who completed a number of tests to measure their L2 skills (e.g. discourse comprehension, vocabulary, grammatical processing, word monitoring, self-paced reading), working memory, and verbal reasoning ability; researchers found that success in L2 comprehension was correlated with both Knowledge (vocabulary, grammatical accuracy, and segmentation accuracy) and Processing Speed (semantic processing speed, grammatical processing speed, segmentation speed, word monitoring, and self-paced listening), and IQ. Subsequent regression analysis showed that Processing Speed did not add onto the model predicting L2 listening comprehension after what was already accounted for by Knowledge and IQ. This is to say that conjointly, vocabulary, grammatical accuracy, and segmentation accuracy, acted as the best predictor for listening comprehension success. Similar results were found by Mecartty (2000) who probed the contribution of lexical and grammatical knowledge to the listening proficiency. Among 77 L2 learners of Spanish, Mecartty reported, whilst both types of knowledge significantly correlated with listening, only lexical knowledge explained unique variance in listening comprehension, further highlighting the fundamental role of linguistic knowledge in listening (e.g. Mecartty, 2000; Vandergrift, 2007). These research evidence on L2 listening correspond to the listening problems international students experience as discussed in section 2.2.1-2.2.3. In addition, these evidence provide good grounds for one to hypothesize that if learners' linguistic knowledge, in particular lexical and grammatical knowledge, becomes improved, their listening proficiency can be elevated as well, and vice versa.

Similar findings have been reported from L2 reading research. Reading, as another receptive skill, also encompasses lower level processes and higher level processes (Grabe, 1991). In line with findings from L2 listening research, research indicates that success in L2 reading comprehension is, to a large extent, associated with learners' lexical knowledge. Analysing data collected from 416 Dutch EFL learners of three age groups: 1) grade 6 (first year of EFL education) 2) grade 8 (third year of EFL education) and 3) grade 10 (fifth year of EFL education), Schoonen et al (1998) concluded that for those who had relatively longer EFL learning experience, English vocabulary was the best predictor of English reading comprehension, accounting for 76% and 60% of the total variance in reading comprehension respectively. This is to say that learners with larger lexical knowledge are more likely to be skilled readers compared to those with smaller lexical reservoir, and the improvement reading comprehension should correspond to the increase in learners' lexis. This is important as it ties with the reading difficulties raised by international students in section 2.2.1-2.2.3 and relates to the methods used in this study and the later interpretation of research findings.

Research has also been conducted to explore what constructs constitute L2 writing ability. Schoonen et al (2003), for example, compared the importance of linguistic knowledge, metacognitive knowledge and fluency or accessibility of such linguistic knowledge in L2 writing. 281 L2 English learners were involved and their writing proficiency, vocabulary knowledge, English and Dutch (L1) orthographic knowledge, grammatical knowledge, metacognitive knowledge, speed of lexical retrieval and of sentence building were measured and structural equation modelling was used to analyse the data. Correlation analyses showed that all measured L2 components correlated significantly with writing proficiency, with medium to even large effect size and together, they explained a substantial proportion of the total variance in L2 writing. Further analyses showed that knowledge measures had a stronger correlation with writing proficiency than speed measures, which is reasonable because unlike listening which often requires on-spot, timely processing of information input, the time allowance for writing is often more lenient. Nevertheless, findings from Schoonen et al's (2003) inferred that learners with higher L2 writing proficiency were likely to have better knowledge of vocabulary, spelling and grammar, as well as a higher speed of processing. Linking this back to the writing difficulties reported in literature from 2.2.2, one can argue that the international students who found writing in English speaking HEI difficult may have insufficient lexical, orthographic, grammatical knowledge or efficient speed of processing.

Lastly, for L2 speaking, Hulstijn, Schoonen, De Jong, Steinel & Florijn's research (2012) on the relationship between individual differences in subskills (i.e., in skills hypothesized to be components of speaking proficiency) and individual differences in successfully conveying information through speaking (i.e. functional adequacy) offers valuable insights into international students' inactive classroom participation and their lack of verbal engagement as discussed earlier (section 2.2.1, 2.2.3). Using data collected from 181 L2 learners of Dutch, structural equation modelling revealed that the final model consisting of learners' lexical and grammatical knowledge, speed of lexical retrieval, articulation response, sentence building, and pronunciation fit the data well. Moreover, apart from response latency and response duration, all abovementioned measured components correlated substantially with the function adequacy and vocabulary knowledge and intonation were found to be significant predicators (standardized regression coefficients of .305 and .341, respectively). Given the established importance of linguistic knowledge (vocabulary and grammar), speed-of-processing skills (lexical retrieval and sentence building), and pronunciation skills (speech sounds, word stress, and intonation) for L2 speaking proficiency, there is good reason to believe that the afore-discussed lack of verbal participation and engagement observed among international students are intrinsically tied to the insufficient linguistic knowledge, the lack of speed-of-processing as well as pronunciation skills. Pronunciation skills are found particularly challenging for international Chinese students as they reported feeling embarrassed by their non-native pronunciation (Feng, 1991).

2.3.3 Acknowledged indicators of L2 proficiency

Regardless of how researchers have approached the notion of language proficiency, either as a single general construct or through analysing the four key language skills, the following consensuses have been reached from the afore-presented research. Firstly, vocabulary knowledge has been repeatedly put forward as a good indicator, if not the best predictor, of proficiency in L2 reading (Schoonen et al., 1998), L2 listening (Andringa et al., 2012), L2 writing (Schoonen et al., 2003) as well as L2 speaking (Hulstijn et al., 2012). Indeed, as "the building block of language" (Schmitt, Schmitt, & Clapham, 2001, p. 53), many linguists and educators consider vocabulary to be the single most important aspect of foreign language learning (Knight, 1994). In line with Schoonen et al's finding (2003), Laufer (1998), who looked at the relationship between L2 learners' vocabulary size, lexical text coverage that their vocabulary provides and their reading comprehension, also asserted that vocabulary correlated with holistic assessments of writing and general proficiency, and was the best single predictor of reading comprehension. More research has showed vocabulary size is one of best predicators of one's reading and comprehending ability (August, Carlo, Dresler & Snow, 2005; Qian, 2002; Read, 1988), subsequent vocabulary acquisition (Pulido, 2003; Verspoor & Lowie, 2003) and general proficiency (Grabe, 1991; Hermann, 2003; Zareva, Schwanenflugel & Nikolova 2005).

In addition, research has also shown that the speed of processing is closely associate with L2 listening (Andringa et al., 1998) and in L2 speaking proficiency (Hulstijn et al., 2012). Although it did not make unique contribution to the prediction of L2 listening proficiency after taking the knowledge factor into account in Andringa et al's research (2012), others have highlighted the need of efficient lower order processing, such as word identification and syntactic parsing for the development of L2 reading comprehension (Favreau & Segalowitz, 1983; Koda, 1996; Segalowitz, Poulsen, & Komoda, 1991). The rationale behind this need is that readers have limited working memory capacity, resulting in a competition between lower order decoding and higher order comprehension processing (Just & Carpenter, 1992; Inhoff, Pollatsek, Posner & Rayner, 1989). When lower order processing is slow and attention-demanding, the higher order processing needed for text comprehension suffers.

With research furthering the understanding of L2 proficiency, coupled with the fact that the impact of globalisation and economic development has made English the lingua franca of the world and a language of opportunity (British Council, 2013), the need to learn English and to become proficient users of English for communicative purposes have become the goal for many. Along with this global trend of English learning, the need to be certified by well-acknowledged assessment organisations as proof to showcase English language skills emerges. Meanwhile, as mentioned at the beginning of this chapter (2.1), rapid globalization and economic development have significantly boosted the number of Chinese students travelling to English-speaking countries, such as UK, seeking further education. As previous sections have demonstrated, international students' social adaptation and academic attainment at an English speaking HEIs are, to a large extent, associated with their English language proficiency, or rather, the lack of which. Given these, a reliable and valid measure of English proficiency is called for to demonstrate that they are linguistically equipped to study in an institution where English is used as the medium of instruction, and for the accepting institutions, such measure of English proficiency is demanded both for bar setting and for application screening.

2.4 Standardized English proficiency tests

2.4.1 ELTS and IELTS

Many tests have also been developed to serve the purpose of measuring the proficiency for the large population of L2 English learners in a standardized and reliable manner. The primary focus of this present study was on the International English Language Testing System (IELTS), developed and managed collaboratively by the British Council, the University of Cambridge Local Examinations Syndicate (UCLES) and International Development Program Education Australia (IDPEA). It is an examination “designed to assess the English language ability of people whose first language is not English and who need to study, work or live where English is used as the language of communication” (IELTS homepage, 2018).

The English Language Testing Service (ELTS), the predecessor of IELTS, was first created and introduced by the British Council during 1980s (Celestine & Ming, 1999). Between the late 1980s and early 1990s, ELTS underwent a series of changes; to start with, the word International was added in 1989 to acknowledge the involvement of the International Development Program Education Australia (IDPEA) on managing the test. Secondly, on the basis of research on the effectiveness of one-module general proficiency assessment approach conducted by International Editing Committee, the previous three subject-specific modules were replaced by one academic reading module and one academic writing module. This change in assessment approach indicated the following assumptions: 1) there existed a general language proficiency that could be tested on non-subject specific grounds, and 2) candidates who performed well on a general proficiency test should be able to cope with subject specific texts successfully during their future academic studies (Celestine & Ming, 1999). Moreover, the thematic link between reading and writing in ELTS was removed due to concerns over construct validity (detailed discussion on test validity is presented in 2.5.3). Researchers have argued that in the ELTS test where candidates were required to produce a piece of writing using the background schemata provided through the reading passages could confuse the assessment of writing skills with the assessment of reading skills. In addition, the band range of ELTS general training reading and writing was increased from six to nine to match the band range of academic reading and writing (Charge & Taylor, 1997).

Today, IELTS candidates, regardless of their disciplines and their level of study (undergraduate, postgraduate, or others) sit the same listening and speaking tests, while the reading and writing tests differ depending on whether the candidate chooses the academic or the general version. The academic reading and writing assess whether candidates are ready to study or train in the medium of English at an undergraduate or postgraduate level. The emphasis of the general training reading and writing is on communication skills in a broad social and educational context, suitable for those who are going to English-speaking countries to complete their secondary education or to undertake work experience or training programmes at pre-degree level (Charge & Taylor, 1997). For the present study, which is situated in an academic context, from hereafter, unless specified, IELTS refers to IELTS Academic, not IELTS General.

2.4.2 TOEFL

On the other side of the Atlantic, a battery of English proficiency tests has also been developed, among which is the long-standing Test of English as a Foreign Language (TOEFL). Under the guidance of the National Council[4], "TOEFL was first developed in the early 1960's to assess the English proficiency of nonnative speakers of the language who intend to study in institutions where English is the language of instruction" (ETS, 2018 p.2). Since1965, the responsibility of managing TOEFL has been shared by the College Board and Educational Testing Service (ETS). Like IELTS, TOEFL also underwent a series of changes since its conception in 1960s not only in terms of test formats (from a paper-based test i.e. TOEFL pBT, to a computer-based test i.e. TOEFL cBT, and in 2005, to an internet based test, i.e. TOEFL iBT), but also in the underpinning theoretical frameworks.

[4] The Council was formed through the cooperative effort of more than 30 public and private institutions concerned with the English proficiency of non-native speakers, especially those applying to English-medium academic institutions.

As stated earlier in section 2.3.1, the original TOEFL only assessed candidates' knowledge of vocabulary, reading comprehension, listening comprehension, knowledge of correct English structure and grammar. It was not until the 1970s that speaking and writing skills became formally assessed through TOEFL tests, which had expanded to include the Test of Spoken English (TSE) and the Test of Written English (TWE) in addition to the initial multiple-choice items. Such change in TOEFL was also a reflection of the change that took place in understanding proficiency as concept that involves not only the linguistic knowledge such as grammar and vocabulary, but also the appropriate production using such linguistic knowledge. In other words, the inclusion of TSE and TWE as part of the TOEFL test suite echoed the shift from linguistic knowledge being the centre of proficiency (Lado, 1961; Carroll, 1961) to communicative competence being equally as important as advocated by Bachman and Palmer (1996), Canale & Swain (1980), and Hymes (1972).

Under the guidance of communicative competence theories, focusing specifically on academic contexts, along with the latest development in applied linguistics, research methods and psychometrics as well as information technology, TOEFL today has become an internet based test consisting of academic tasks that require the integration of receptive and productive skills such as listening, reading and writing or speaking, as well multiple-choice items for listening and reading. Such test design and content are based on evidence drawn from various research studies conducted by ETS (e.g. Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt, & Schedl, 2000; Cumming, Kantor, Powers, Santos & Taylor, 2000; Bejar, Douglas, Jamieson, Nissan & Turner, 2000; Butler, Eignor, Dan, Stan, McNamara & Suomi, 2000; Rosenfeld, Leung & Oltman, 2001).

Comparing IELTS Academic with TOEFL, similarities can be found both in terms of current test formats, underpinning theoretical frameworks and their current status. At present, both tests act as a measure of general English proficiency with a focus on communicative competence and language features common in academic contexts. Although both are being taken world wide for various purposes, the majority of IELTS and TOEFL tests are taken by non-native speakers of English so as to obtain the proof of their English proficiency to study at an English speaking HEI, often at the request of their accepting institutions. IELTS statistics showed that more than three million tests were taken in 2016, among which 80.7% were IELTS Academic. IELTS Academic is now accepted by all universities in Australia and the UK and many of the leading institutions in the USA (IELTS homepage, 2018), signalling the importance of IELTS as a test and the stake it holds for non-native speakers of English with a study-abroad agenda. Likewise, ETS states "The *TOEFL* test is the most widely respected English-language test in the world, recognized by more than 10,000 colleges, universities and agencies in more than 130 countries, including Australia, Canada, the U.K. and the United States. Wherever you want to study, the TOEFL test can help you get there" (TOEFL homepage, 2018).

Given the pivotal role English proficiency plays in international students' social, psychological and academic well-being as discussed earlier in section 2.2.1-2.2.3, the request of standardized proficiency test results as part of the admission criteria for international students is justified. In addition to this gatekeeping function, what could also be implied from the setting of language requirement is that the receiving universities are using these tests not merely as a yardstick of measuring proficiency, but also a threshold beyond which international students' chances at achieving academic success should no longer be constrained by their language proficiency. In other words, the inclusion of language requirement through setting a proficiency test result threshold (e.g. minimum IELTS 5.5 or minimum TOEFL iBT 95) shows that accepting HEIs assume that international students who are able to meet such requirements would have the needed language proficiency or are be able to develop the needed language proficiency to fulfil their academic potentials, compared to those who fail to meet such requirement.

This worldwide acceptance of IELTS and TOEFL scores has made these two tests the centre of many academic research and social debates, especially given the stakes both tests hold among EFL countries, EFL leaners, and in particular, international students. One of the most frequently posed question for these two tests is related to their reliability and their validity as a measure of English language proficiency and as a "gate-keeper" for English speaking HEIs.

2.5 Test reliability and validity

Regardless of the stakes a test may hold, a good test, as a measure, needs to have both reliability and validity. Validity is often seen as the "hallmark of quality" (Newton & Shaw, 2014, p.1) and the "single most important criterion' as far as testing and assessment are concerned (Koretz, 2008, p.215). Over the decades, validity as an overall concept has been debated from a variety of perspectives; given the context and the goal of the present study, here, the discussion of validity is narrowed to focus primarily in the realm of educational testing. Before moving onto discuss whether the use of IELTS and TOEFL as a measure of proficiency to screen international students is valid, a general introduction of key test reliability and test validity theories is presented first.

2.5.1 Reliability

Test reliability refers to whether a test is able to produce consistent outcomes throughout different administrations (Newton & Shaw, 2014), a concept closely associated with test validity. In the context of proficiency testing, for a test to be considered reliable, the scores obtained by candidate A at a particular time at a certain test centre should not be significantly different from the scores she or he would achieve if she or he decides to repeat the same test again within a relatively short time interval (e.g. within 3 months) at another test centre. Reliability is a prerequisite for a test to have validity for simple reasons; if a test can not produce consistent outcomes when measuring a specific construct/attribute, there is no validity to be investigated.

For standardized proficiency tests used globally for multiple purposes, such as IELTS and TOEFL, it is essential for them to be reliable so as to be considered valid. According to IELTS statistics (IELTS Test performance, 2017), in 2017, the Cronbach's Alpha, a reliability estimate which measures the internal consistency of the test items in IELTS Academic Listening and Academic Reading modules, was reported with an average .91 and .90 coefficient[5]. Similarly, ETS (2018) has also published the reliability estimates for TOEFL iBT, 0.94 for overall TOEFL, 0.85 for Listening and Reading scores respectively, 0.88 for Speaking and 0.74 for Writing. In addition to reliability estimates published from test developers (i.e. ETS), Zhang (2008) also empirically examined the reliability of TOEFL scores through comparing TOEFL scores of more than 12,000 examinees who sat TOEFL iBT tests within a period of one month. The correlations of their scores on the two test forms were 0.77 for the Listening and Writing, 0.78 for Reading, 0.84 for Speaking, and 0.91 for their overall TOEFL score. Given these reliability statistics reported either by test developers themselves, or by independent researchers, there is good reason to conclude that on the reliability front,

[5] for more detailed breakdown regarding the calculation of this coefficient, see https://www.ielts.org/teaching-and-research/test-performance

both IELTS and TOEFL are reliable measures, capable of producing consistent test outcomes.

For high stake tests such as IELTS and TOEFL, being reliable is far from enough. As mentioned in the previous section, English-speaking HEIs globally have been using IELTS and TOEFL as part of their admission requirement for international students from non-English speaking backgrounds. Considering the current status of IELTS and TOEFL, attention has been paid to the discussion of the validity and validation of these two tests.

2.5.2 Validity

Validity is an overall evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment (Messick, 1989). Validity is not a property of the test or assessment per se, but rather of the meaning of the test scores. Hence, what is to be validated is not the test or observation device per se, but rather the inferences derived from test scores or other indicators (Cronbach, 1971) - inferences about score meaning or interpretation and about the implications for action that the interpretation entails. Among the various types of validity put forward since 1940s, the following sections focus on the discussion of *content*, *construct* and *criterion-related* (concurrent and predictive) validities; meanwhile, issues relevant to *washback,* a component of *consequential validity* are also included in later sections.

2.5.3 Content validity

TOEFL Content validity

For a test to have content validity, it needs to consist of “a representative sample of language structures and skills with which it is meant to be concerned” (Hughes, 1989, p. 22). Specifically, content validity encompasses content relevance and content coverage (Bachman, 1990). In the realm of language proficiency test, on the one hand, content relevance applies to both the language ability the test aims to assess and the test method itself; on the other, content coverage entails whether the test tasks mirror the tasks candidates would encounter in the target language context (i.e. authenticity).

Many research efforts have been made by ETS to ensure the content of TOEFL relates closely to the language construct under assessment and covers a variety of authentic tasks during the test's design and development stage. For example, Biber & Gray (2004) collected a corpus of 1.67 million words of spoken language at four universities to establish the representativeness and authenticity of the lectures and conversations that are used to in TOEFL Listening to assess candidates' listening comprehension. In addition, Cumming, Grant, Mulcahy-Ernt, & Powers' study (2005) also provided empirical evidence on the "relevance, authenticity, and educational appropriateness of integrated test tasks", i.e. tasks that require the integrated application of two or more language skills (ETS, 2008, p.5). 7 experienced ESL teachers were asked to rate whether their students' performance on the sample TOEFL speaking and writing tasks (a) represented the domain of academic English required for studies at English-medium universities or colleges in North America, (b) elicited performance from their adult ESL students that corresponded to their usual performance in ESL classes and course assignments. Results indicated that these tasks were realistic and appropriate simulations of tasks students would encounter in their academic studies and that students' performance on these sample TOEFL tasks were in line with their usual performance in classes, if not better. Regarding task variety and authenticity, in current TOEFL, a wide range of test tasks and formats has been incorporated (e.g. listen to lectures, classroom discussions and conversations, then answer questions; read academic passages and answer questions; discuss a familiar topic; speak based on reading and listening tasks) so as to mimic the tasks candidates may encounter in their subsequent academic studies in English-speaking HEIs (ETS homepage: test content, 2018).

IELTS content validity

Although detailed studies concerning how materials were collected and how test items could not be found, formats and contents were developed for IELTS, information provided on the IELTS homepage indicates that IELTS has content relevance as all four language skills are included in the test. Evidence on content coverage can also be found as various real-life task types are involved in the test content, e.g. listen to a conversation set in an everyday social context, a monologue set in an everyday social context, a conversation between up to four people set in an educational or training context; read academic passages for the gist, reading for main ideas, reading for detail, skimming, understanding logical argument and recognising writers' opinions, attitudes and purpose; write on topics are of general interest to, and suitable for entering undergraduate and postgraduate studies; speak on general questions such as home, family, work, studies and interests and discuss a particular topic of the examiner's choice (IELTS homepage: test format, 2018).

2.5.4 Construct validity and relevant threats

Construct validity, in simple terms, examines whether a test is indeed testing the *construct* or *attribute* that it claims to test and it is often regarded the most important aspect in test validation, i.e. "the process of making a case for the proposed interpretation and uses of test scores" (Bachman & Palmer 1982; Bachman 1990; Brown, 2000; Cronbach, 1984; ETS, 2008, p.3). The word *construct* here "refers to any underlying ability which is hypothesised in a theory of language ability" (Hughes, 1989, p.26). Broadly speaking, construct validity involves construct relevance and construct representativeness and the main threats to these two are referred to as construct irrelevance and construct under-representation. A typical example of construct under-representation could be the previous version of TOEFL, which only included solely multiple-choice items assessing vocabulary, reading comprehension, listening comprehension, knowledge of correct English structure and grammar, but excluded productive language skills such as speaking and writing. For the old TOEFL to be regarded as a measure of overall proficiency and communicative competence would be seen as construct-underrepresented.

Messick (1989) regarded construct-irrelevant variance as a contaminant to score interpretation if a test contains "excess reliable variance that is irrelevant to the interpreted construct" (p. 34). This type of invalidity, according to Messick (1989), takes two forms: construct-irrelevant easiness and construct-irrelevant difficulty, that is, contaminating influences such as an individual's background knowledge, personality, characteristics, test-taking strategies, and general intellectual or cognitive ability that tend to systematically increase or decrease test scores for an individual candidate or a group of candidates; hence effort needs to be made to keep influences such as these to a minimum (Jin & Yan, 2017; Zhen & De Jong, 2011). For example, a validated proficiency test can be subverted by test preparation practices or coaching emphasizing testwiseness strategies that might increase test scores without correspondingly improving the skills measured by the test. Although this would not compromise the validity of the uncoached test in general, the validity of the interpretation and use of the coached scores would be jeopardized.

By contrast, test preparation practices emphasizing test familiarization and anxiety reduction may actually improve validity: scores that formerly were invalidly low because of anxiety might now become validly higher (Messick, 1982).

For high-stake tests such as TOEFL and IELTS, it is reasonable to expect the tests to be optimally authentic with minimum construct under-representation and construct-irrelevance. Construct under-representation and construct-irrelevance may lead to the test being too narrow and excluding "important dimensions or facets of focal constructs" or the test being too broad "containing excess reliable variance that is irrelevant to the interpreted construct" (Messick, 1996, p.3). A series of construct validation studies have been commissioned by ETS (e.g. Biber & Gray, 2013; Brown, Iwashita & McNamara, 2005; Cohen & Upton, 2006; Swain, Huang, Barkaoui, Brooks & Lapkin, 2009) and IELTS (e.g. Moore, Morton & Price, 2012; Weir, Hawkey, Green, Unaldi & Devi, 2005) respectively to showcase that both tests, or rather, scores of both tests, are capable of provide inference as well as discriminate candidates of difference English proficiency level.

2.5.5 Criterion-related validity: concurrent validity and predictive validity

Under the criterion-related validity lies the concurrent validity, which "is established when the test and the criterion are administered at about the same time" (Hughes, 1989, p.23) and the predictive validity, which relates to whether the test could accurately predict the candidates' performance and behaviour in the future assessed through an external measure (Anderson, Clapham & Wall, 1995). To examine concurrent validity, correlation-coefficient is often calculated to see if two measures of different formats for the same or similar construct could yield similar results consistently. To examine predictive validity is much more complex because the target performance or behaviour we wish to predict in the future often rely on many other non-language related factors (Bachman, 1990; Sewell, 2009).

Sawaki and Nissan (2009) looked at how TOEFL Listening relates to listening to academic lectures at English-speaking HEIs (the criterion) and found substantial correlations between the criterion and the TOEFL Listening section score (Pearson correlation coefficients ranging from .56 to .74). Moreover, the validity of TOEFL Speaking as a measure of speaking ability in typical academic settings, such as speaking about academic course content, campus life, and familiar daily topics has been constantly examined and supported by Waters (1996), Douglas (1997), Rosenfeld, Leung, & Oltman (2001). Furthermore, Wang, Eignor and Enright (2008) investigated how TOEFL iBT relate to candidates' self-perceived language proficiency (the criterion) using a questionnaire consisting of "can do" statements. An average correlation between coefficient of .46 between the summative scores for each of the four self- assessment scales with test scores on the measures of four skills was observed along with a coefficient of .52 with the total test score, suggesting a moderate concurrent validity between TOEFL iBT and self-perceived language proficiency.

Apart from Zheng and De Jong's study (2011) which reported a .83 correlation coefficient between IELTS and Pearson Test of English Academic (another standardised, computer-based academic English proficiency test), evidence regarding IELTS's concurrent validity could only be inferred from Sewell's work (2008) which claimed in 2007, the inter-module correlation between IELTS reading and listening modules was 0.89. These figures, by themselves would support the test's claims of concurrent validity but, Bachman warns (1990), without evidence from an independent source supporting this interpretation of the criterion of the ability being tested, that there is no firm basis for interpreting this criterion as evidence of validity. The inter-module correlation of IELTS writing and speaking modules is unclear because they are not item-based, which also brings doubt to their concurrent validities.

It is clear that there are obvious differences in test formats and target candidature between IELTS and TOEFL; while IELTS consists of various task types and is targeted mainly at candidates who wish to study at UK, Australian and New Zealand HEIs, TOEFL consists mainly of multiple choice questions and is targeted at those who wish to study at North American HEIs. This, coupled with the high fees candidates need to pay for taking these tests ($200 for one TOEFL, £153 for one IELTS), as well as the large amount of time and preparatory efforts required for taking one test, indicates the probability that the population of candidates who took both tests is relatively small. Hence, few studies have examined the correlation between the two from concurrent validity perspective. Only one study could be found to link IELTS scores to TOEFL scores (ETS, 2010). Although this study reported a correlation between the two tests' overall scores (*r*=.73 between IELTS and TOEFL overall scores), the correlation coefficients between the four modules were comparatively weaker (*r*=.44 between

IELTS and TOEFL Writing, *r*=.57 for IELTS and TOEFL Speaking, *r*=.63 between IELTS and TOEFL Listening, and *r*=.68 between IELTS and TOEFL Reading).
So far, it is reasonable to conclude that the majority of the key literature on the topic of validity agrees that TOEFL and IELTS have demonstrated sufficient content, construct and concurrent validity to be widely used as a reliable measure of English language proficiency. What remains undiscussed is whether or not these two tests of proficiency, which are often used as a part of admission requirements for screening international students from non-English speaking backgrounds, are indeed capable of acting as a predictor for academic attainment. In the following sections, empirical evidence concerning the predictive validity of language proficiency as measured by TOEFL and IELTS are presented.

2.5.6 TOEFL as a predictor for academic attainment

Earlier studies (1980s-1990s)

Many researchers (e.g. Arcuino, 2013; Chen & Sun, 2006; Cho & Bridgeman, 2012; Person, 2002) have scrutinised the relationship between English language proficiency (indexed by TOEFL and GRE-Verbal) [6] and international students' academic attainment in various educational HEI contexts and mixed results have been yielded. One of the earliest predictive studies examining the relationship between language proficiency and international students' academic performance was carried out by Sharon in 1972. The aim of this study was to examine whether TOEFL added to the predictive validity of Graduate Record Examination (GRE)-Verbal test among 975 international students from 24 US universities. Unlike most later predictive studies that treat TOEFL as a primary linguistic predictor (e.g. Chen & Sun, 2006), in this study, Sharon hypothesized that TOEFL would act as a moderator for the relationship between GRE-V and university level academic performance in the sense that students scoring high on TOEFL would be more predictable by GRE-V than those scoring low. In other words, Sharon assumed that if a student did not have adequate English proficiency, a verbal aptitude test could not accurately predict his/her scholastic achievement. GRE-Q[7] scores

[6] GRE-Verbal Reasoning measures candidates' ability to analyse and evaluate written material and synthesize information obtained from it, analyse relationships among component parts of sentences and recognize relationships among words and concepts (ETS GRE homepage, 2018).

[7] GRE-Quantitative Reasoning measures candidate's ability to understand, interpret and analyse quantitative information, solve problems using mathematical models and apply basic skills and elementary concepts of arithmetic, algebra, geometry and data analysis (ETS GRE homepage, 2018)

were also included as an additional predictor and GPA (Grade Average Point) acted as an indicator of academic achievement. Initial correlation analysis found a moderate-strong correlation (.70) between GRE-V and TOEFL indicating that the two tests indicating that, to a large degree, measuring the similar ability and that there could be collinearity.

Regression analysis reported that GRE-Q was the single best predictor of GPA for all disciplines, with a coefficient of .32. Further, it was found the linear combinations of GRE-V or Q with TOEFL did not significantly improve the overall model fit, i.e. TOEFL did not add to the prediction of GPA after what was already accounted for by GRE-V or GRE-Q when all participants were analysed together.

Interestingly, when participants were separate into difference subgroups based on their disciplines (e.g. engineering, technology and mathematics, which could be seen as linguistically less demanding disciplines, and other[8]) and their TOEFL scores (low, middle, and high), TOEFL did act as a moderator and enhanced the predictive validity of GRE-V in some cases. For participants from linguistically less demanding disciplines, TOEFL moderated the relationship between GRE-V and GPA in the engineering, and the predictive validity of GRE-V increased from .22 to .35 in the low TOEFL group, and to .36 in the middle TOEFL group. Similarly, TOEFL also moderated the relationship between GRE-V and GPA among other students, increasing GRE-V's predictive validity from .35 to .44 in middle proficiency group.

This change in the moderator role of TOEFL when all participants were looked at together and when participants were separated into subgroups highlighted that when analysing the relationship between language proficiency and academic attainment, it is important to take into consideration the disciplinary differences, especially when the sample is of crossdisciplinary nature. From an analytical viewpoint, this is particularly important because it suggests that the "real" relationship between two variables, language proficiency and academic attainment in this context, could be, to some extent, masked by the noise from data created by uncontrolled variables such as field of disciplines. If this noise was not controlled for during analyses, reliability of findings

[8] The category "other' consists of all students not majoring in engineering, technology, mathematics, or natural sciences.

could become questionable. Given this importance, the following review of literature in this section and section 2.5.6 also looks at whether existing research has factored in the participants' field of disciplines.

Around the same time as Sharon (1972), with the aim to build a regression equation for prediction of success in a master's level program, Ayer and Peter (1977) made a similar attempt to address the relationship between academic success, TOEFL scores and GRE. 50 Asian masters students who were studying linguistically less demanding programmes such as engineering, chemistry or mathematics at the time at a USA university were involved. Their GPA were collected as an indicator of their academic performance, acting as the dependent variable; their TOEFL scores, GRE Verbal and GRE Quantitative scores were also collected as indicators of their language proficiency and mathematic skills, acting as the independent variables. Correlation between TOEFL scores with overall GPA reported a positive and significant relationship between these two variables at a 0.01 level (r=.40). Correlation between GRE-Q and GPA also yielded a positive and significant relationship but with moderate magnitude (r=.55, p<.05). The larger correlation coefficient between GRE-Q and GPA could be related to the bigger standard deviation of GRE-Q in comparison to that of TOEFL. Furthermore, comparing students who initially had scored one half standard deviation above the mean (N=14) with one half standard deviation below the mean (N=12) on the TOEFL, t-test indicated a significant difference a .05 level between the two groups. In other words, those who began their master's programme with higher English proficiency (i.e. higher TOEFL scores) were indeed more likely to perform academically better than those who entered their programme with lower proficiency (i.e. lower TOEFL scores). In addition, regression analysis found that a combination of scores from GRE-V and TOEFL predicted GPA reasonably well; additional variance explained by GRE-Q on the other hand was not significant.

Findings from both Ayer and Peter's and Sharon's studies, although 4 decades ago, indicated that even for international students studying linguistically less demanding disciplines such as engineering, chemistry or mathematics, their academic attainment was to various extent, affected by their language proficiency. However, as international students involved in both Ayer and Peter's and Sharon's studies were likely to be admitted on the basis that they had met the language requirement, as indexed by TOEFL scores, one could argue their subsequent academic attainment should no longer be affected by proficiency if the requirement was set at a correct level. This contradiction between achieving sufficient TOEFL scores and TOEFL still being a predictor for academic attainment leads to the question of whether the language requirement was set too low.

More research has been done on the relationship between language proficiency as indexed by standardised tests and international students' academic attainment. With an attempt to examine whether TOEFL scores or First Certificate of English (FCE) scores could predict their university academic attainment as measured by GPA, Al-Musawi and Al-Ansari (1999) recruited 86 first and second-year English Language and Literature (i.e. linguistically more demanding programmes) students at a Bahrain University where English is used widely as a foreign language. Participants' language proficiency was assessed using TOEFL and FCE; their academic performance in non-English courses was indicated through composite GPA while their English course performance was indicated through ENGPA. Given that participants were EFL learners studying English Language and Literature, it is reasonable to assume that both their GPA and their ENGPA were related to their English language proficiency. Initial correlation analyses found a weaker and of lower magnitude relationship between TOEFL and GPA (r=.50, p<.05), compared to that between FCE and GPA (r=.69, p<.01). Similar difference was observed in the correlation between TOEFL and ENGPA (r=.70, p<.05) and between FCE and ENGPA (r=.84, p<. 01). Furthermore, stepwise regression predicting student's GPA, and ENGPA, respectively, using the componential scores of the TOEFL and the FCE as independent variables revealed that TOEFL componential scores did not contribute to the overall prediction of GPA while the combination of three componential scores of FCE (namely, multiple-choice, cloze, and sentence transformation) predicted GPA reasonably well, explaining a total of 52% of the variances. Likewise results were found using ENGPA as the outcome variable, but in this regression, TOEFL section 2 contributed slightly to the overall model. TOEFL section 2, combined with the same abovementioned three componential scores of FCE, explained a majority of the total variance in ENGPA (R^2=.79). Based on these results, Al-Musawi and Al-Ansari concluded that for these participants, it seems that their academic attainment was "determined by their performance on the FCE exam rather than by their performance on the TOEFL test" (p.397).

However, upon closer examination, this conclusion seemed somewhat selfcontradictory. The contrast between TOEFL being an insignificant predictor and FCE being a significant predictor brought the validity of TOEFL into questions, rendering TOEFL somewhat less capable a predictor than FCE. However, if FCE appears to measure the same construct as that being measured by the TOEFL (Bachman et al., 1990), why would there be such a difference? To account for this difference, Al-Musawi and Al-Ansari further argued that the reason why TOEFL failed to contribute to the overall prediction could be that students of English majors had received preparation for the TOEFL. In other words, it was suggested that the relationship between TOEFL and academic attainment could, to some extent, be manipulated by test preparation activities. The nature and the effects of dedicated test preparation are discussed later sections of this chapter.

Albeit the potential interference from TOEFL related test preparation and its effect, it could also be argued that the difference in TOEFL's and FCE's predictive power was related to how data analyses were conducted. The seemingly plausible regression analyses could be problematic because, theoretically TOEFL and FCE are measures of very similar, if not entirely the same construct, i.e. English proficiency. Given this similarity, it is reasonable to assume these two predictor variables were collinear; under these circumstances, the use of stepwise regression could result in bias in parameter estimation, inconsistencies among model selection algorithms, an inherent (but often overlooked) problem of multiple hypothesis testing, and an inappropriate focus or reliance on a single best model (Whittingham, Stephens, Bradbury & Freckleton, 2006). Therefore, the more appropriate way to examine which measure of English proficiency, TOEFL or FCE, was a better predictor of academic attainment would be to compare the model fit between the one using TOEFL as the primary predictor and the one where FCE acted as the main predictor. In this sense, it is reasonable to pose questions as to the reliability of Al-Musawi and Al-Ansari's findings.

Although Al-Musawi and Al-Ansari's study adopted a very similar research design as Ayer and Peter's and Sharon's, the difference in research contexts meant findings from these three studies should be interpreted more carefully. To start with, participants involved in Al-Musawi and Al-Ansari's study were EFL learners enrolled in a linguistically more demanding programme in an EFL university and neither TOEFL nor FCE were set as an admission requirement. Because of the nature of the discipline, English language and literature, it is reasonable that students who were more proficient in English, indexed by TOEFL or FCE, were more likely to achieve better academic grades in English language and literature. By contrast, participants in Ayer and Peter's and partly in Sharon's research were international students studying linguistically less demanding programmes at an American university, accepted on the basis that they had met the TOEFL requirement, which, as stated earlier, should mean that they had acquired the demanded language proficiency to complete their academic study, and TOEFL should no longer be a predictor for academic attainment. Therefore, due to the difference in research context and sample makeup, although it is reasonable that TOEFL was found a good predictor in Al-Musawi and Al-Ansari's study, TOEFL being a good predictor in Ayer and Peter's and partly in Sharon's research should be interpreted with caveat.

It is important to point out that, in addition to the afore-stated problems, both Ayer and Peter's (1977) and Al-Musawi and Al-Ansari's (1999) findings were based on relatively small sample size, therefore limiting the generalizability of their results. At the same time, these three studies were relatively out-of-date. Given the changes TOEFL had undergone, more recent research was in need.

Recent studies (2000s-2010s)

In the light of that, Chen and Sun (2006) reexamined the relationship between language proficiency and academic attainment among 712 international undergraduate students from fall 1997 to fall 2003 in an American university. Similar to Al-Musawi and Al-Ansari's study, Chen and Sun also adopted a comparative approach, investigating whether or not international students were better prepared to study in a postsecondary program in an American university if they pass the TOEFL requirement or if they completed an English as a Second Language (ESL) programme. Participants' TOEFL scores were collected along with their first-year GPA. T-test were used to examine whether there were differences between students who entered their programme with TOEFL (TOEFL takers, N=651), students who entered after completing ESL programme (ESL completers, N=36) and students who could not complete the required level of the ESL program but passed TOEFL and eventually were admitted (ESL incompleters, N=55). Results showed that there was no significant difference in GPA between TOEFL takers and ESL completers but there was a significant difference in GPA between ESL completers and ESL incompleters. In other words, students who *initially* either met TOEFL requirement or completed ESL programme had statistically similar GPA; however, if students failed to complete an ESL programme, their first year GPA were significantly lower than those who completed an ESL program even although they could pass the TOEFL *eventually*. This finding is particularly interesting because it indicated that, to some extent, there existed discrepancy in participants' language proficiency between passing TOEFL requirement and completing ESL programme, as participants who failed ESL but *eventually* met the TOEFL requirement, probably through resitting the test, achieved less academic GPA then participants than participants who *initially* met the TOEFL requirement or completed the ESL programme. From this, it could be inferred that through repeating TOEFL, candidates may be enabled to *eventually* achieve the required TOEFL scores, but this does not necessarily indicate that their "actual" general proficiency has reached the required level, and thus there might be a discrepancy between what the scores indicate and what the candidates are actually capable of, linguistically. More discussion on this repeated TOEFL taking and its effect is presented in section 2.7.

Chen and Sun's study (2006), to a large extent, echoed with Person's smaller scale study (2002) that also looked at the relationship between academic performance and language proficiency among postgraduate international students (N=126) enrolled in a U.S. HEI, admitted either on the basis of their TOEFL scores or completing an ESL programme. Using a correlational research design, Person reported that a weak yet significant relationship between TOEFL and participants' first year GPA (r=.266, p<.05) and an insignificant correlation between ESL course performance (measured by Michigan English Language Assessment Battery, MELAB[9]) and first year GPA was insignificant (r=.127, p>.05). Because the difference in correlation coefficient was not tested for statistically significance, one could interpret this difference in different ways. For example, one could argue that the insignificant correlation between ESL course performance and first year GPA suggested that international students who completed the ESL programme had gained sufficient language proficiency and thus their subsequent academic attainment was not hindered; by contrast, the significant correlation between TOEFL and GPA revealed that students admitted through TOEFL might not have obtained sufficient language proficiency and thus their subsequent academic attainment was, to various extent, affected by the lack of such proficiency. However, because Person did not test for the statistical significance of the difference in correlation coefficient, it is also probable that this was caused by the limited variance in using ESL course completion as a predictor or measurement error, thus rendering the

[9] For more details on MELAB, see Person (2002) p.24.

findings inconclusive.

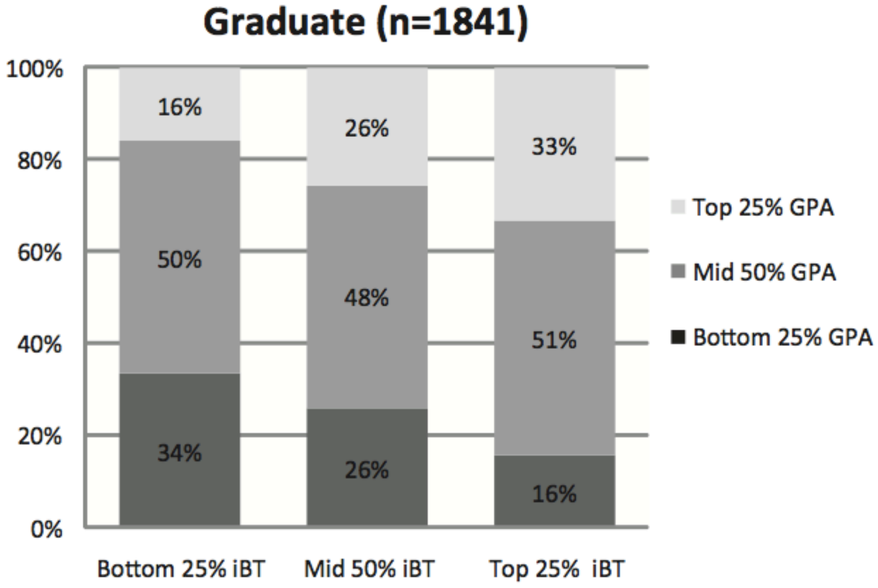
Furthermore, Chen and Sun (2006) found neither the relationship between TOEFL and last year GPA nor that between ESL and last year GPA was reported significant (TOEFL *r*=.009, ESL *r*=.029, *p*>.05). This change in correlation analysis results could mean that, over time, international students admitted through both TOEFL and ESL progammes developed sufficient language proficiency and their academic attainment was no longer impeded by proficiency. Meanwhile, it also highlighted that the predictive power of TOEFL on academic attainment was prone to change over time. In this sense, it is reasonable to speculate that, if analysis only looked at the relationship between TOEFL and academic attainment at the end of their academic journey, results might be incomprehensive. In the light of this, longitudinal research approach is called for and the probability of change over time should not be ignored.

It should be highlighted that findings from Chen and Sun's correlational study should be interpreted with caveat given the following rationales. To begin with, the variances in participants' disciplinary linguistic demand were not accounted for in this study and the considerably imbalanced group size might have contaminated the analysis outcome. Moreover, because of the specificity of ESL programmes, i.e. ESL programmes offered at one institution often differ from ESL programmes offered at another location, the comparison results between TOEFL and ESL programme in Chen and Sun's study might not be applicable to other programmes offered at other institutions. In other words, to what extent can Chen and Sun's study be generalized remained unclear. In addition, although t-test did provide insights into the relationship between language proficiency (indexed by TOEFL or ESL programme completion) by highlighting the differences in GPA among groups of students admitted with different language proficiency criteria, the proportion of variance in GPA explained by proficiency in each group was not clear, calling for more research efforts. Furthermore, the differences in coefficient could be attributed to measurement errors, limited variances in the predictor (ESL programme completion/non-completion) and/or the outcome measure (i.e. GPA).

Situated in similar contexts as Chen and Sun, Cho and Bridgeman (2012) carried out a large-scale cross-disciplinary study looking at the relationship between TOEFL scores and future academic performance as defined by GPA among a total number of 2594 students (1850 postgraduates and 744 undergraduates). Scores on admissions-related tests including TOEFL, GRE/GMAT, SAT/ACT were collected as independent variables while GPA (broken down by year) were collected as dependent variables. For postgraduates, it was found that TOEFL scores correlated significantly with GPA, although the relationship was relatively weak and differed from one discipline to another (*r*=.26 for business students, *r*=.24 for humanities and arts students, *r*=.17 for science and engineering students and *r*=.25 for social sciences students). Regression analysis revealed that TOEFL accounted for 6-7% of the total variance in GPA. Similar patterns were observed for undergraduates; overall, TOEFL explained about 3% of the variance in GPA with weighted average correlations between TOEFL and discipline-specific GPA ranging between .13 and .25.

Cho and Bridgeman further explored their data using expectancy chart, a method less common among predictive studies. As shown in Figure below, 34% of the postgraduate students in the low TOEFL iBT score group received a GPA in the bottom 25% while only 16% earned a GPA in the top 25%. Contrastingly, 16% of the graduate students in the high TOEFL group received a GPA in the bottom 25% range, and 33% received a GPA in the top 25%. The expectancy charts show that there was a much greater chance for students in the high TOEFL group to earn a top 25% GPA, and also that the chance of earning a bottom 25% GPA decreases substantially for the high TOEFL group. Moreover, this pattern was also found in all discipline subgroups of postgraduates and most undergraduate subgroups, except business students. Although business students showed slightly different patterns, expectancy charts still suggested that the chance of receiving a bottom 25% GPA in business courses is much smaller for students with high TOEFL scores than those with low TOEFL scores, which lends support to the overall conclusion that those with higher TOEFL scores were more likely to achieve better academic performance than those with lower TOEFL scores. In other words, academic performance was tied to international students’ language proficiency, regardless of their level of study or their field of discipline.

Figure 2.1 Percentage of the graduate students earning top 25%, middle 50% and bottom 25% GPA by TOEFL iBT score group by Cho and Bridgeman (2012)



Meanwhile, consistent with aforementioned research (e.g. Ayer & Peter, 1977), the correlation was the weakest among science and engineering students, but it is interesting to see that the coefficient for science and engineering students reported in this study (*r*=.17) was much smaller compared with Ayer and Peter's study which also involved engineering, chemistry or mathematics students (*r*=.40). What could have contributed to this difference in coefficient?

It could be related to difference in TOEFL requirement set for university admission. For example, the TOEFL requirement was comparatively lower in Ayer and Peter's study than that in Cho and Bridgeman's, which meant participants in Ayer and Peter's study have lower proficiency than participants in Cho and Bridgeman's; thus academic attainment was more affected by such lack of proficiency in Ayer and Peter's study than that in Cho and Bridgeman's. Alternatively, if TOEFL requirements in both studies were set at a similar level, this discrepancy in coefficient could be explained by different means used to achieve the TOEFL scores by participants both studies. For example, participants in Ayer and Peter's might have been engaged in test preparation activities to help achieve the scores and thus their TOEFL scores might not be a correct representation of their "true" proficiency, while participants in Cho and Bridgeman's study were not engaged in such activities and thus their scores were more likely to be their proficiency. In other words, participants in Ayer and Peter's may appeared to have gained the scores, but their proficiency could still be lacking, affecting their academic attainment to a larger degree, in comparison to participants in in Cho and Bridgeman's. This hypothesis of test preparation affecting the relationship between language proficiency indexed by standardised proficiency tests and international students' academic attainment is at the heart of the present study; relevant literature on test preparation is provided at section 2.6.

The use of expectancy charts in Cho and Bridgeman's study, a less-frequently used analytical method, seems to have provided a clearer picture regarding the relationship between proficiency as measured by TOEFL and academic attainment as measured by GPA, in comparison to the inconsistent findings drawn from conventional correlation or regression analyses. This inconsistency could be related to the noise brought by the heterogeneity of data if the research involved a cross-disciplinary sample. For example, in Cho and Bridgeman's, the sample consisted of both undergraduate and postgraduate studies of different disciplines as well as of different universities. It is very likely that assessment standards to evaluate participants' academic attainment varied from one discipline to another, one institution to another, thus introducing noise to the data that could affect the correlation or regression analysis outcomes. This assumption has also been brought up in Ho & Spkins's study (1985) where the researchers argued the use of composite criterion measures (e.g., GPAs), in which heterogeneous elements are included, has been the rule rather than the exception in many predictive studies. The problem of criterion heterogeneity is likely to be especially serious at the university level, "where various academic subjects demand divergent competencies or dispositions" (p.258). Serval means could be taken to counter this problem. Firstly, these noises could be effectively controlled through factoring in the institutional or disciplinary differences in the correlation or regression analyses, so as to improve the reliability of research findings. At the same time, visualisation of data through charts can also be of value when looking at the relationship between these two variables. Both means are applied in the present study and more details are provided in Chapter 5.

Many other researchers have also reported inconsistent findings regarding the relationship between language proficiency as measured by TOEFL and academic attainment; to present a holistic overall of relevant students without unneeded repetition, the following table is provided as a further support regarding TOEFL's inconsistent predictive validity.

Table 2.1 Summary of predictive study on the relationship between language proficiency and academic attainment using TOEFL as a measure

Researcher	Year	Participants	Measure of language proficiency	Measure of academic attainment	Key findings
Hwang & Dizney	1970	63 Chinese graduate students at an American University	TOEFL	GPA	Correlation between total TOEFL scores and first term GPAs were insignificant.
Ho & Spinks	1985	230 first-year Arts faculty students at the Hong Kong University where English is the language of instruction	Independent measures of English skills	GPA	English skill variate accounted for 10.6% of the variance of the examination GPAs. The correlations between English reading and GPA, speaking and GPA were significant, *r*=.17, .18 respectively, *p*<.05.
Light, Xu & Mossop	1987	376 international postgraduate students enrolled in an American university	TOEFL	GPA, credit hours	TOEFL correlated significantly with GPA (*r*=.14, *p*<.05) but the coefficient was too weak to have practical meaning in predicting GPA. There is a stronger relationship between GPA and TOEFL for Humanities/arts/social science students than for science/math business students. TOEFL also correlated significantly with the number of credit hours students earned during their first semester of study, *r*=.19, *p*<.01.
Johnson	1988	196 international students enrolled as undergraduates at an American university	TOEFL	GPA, credit hours	The overall mean TOEFL score correlated significantly with the mean GPA (*r*= .36, *p*<.01). Students (n = 68) with TOEFL scores below 500 earned significantly lower grades (*z*=-3.77, *p*< .01) than students (N=128) with TOEFL scores of 500 and above. Similar to Light, Xu & Mossop, there was a significant correlation(*r*=.80, *p*< .01) between credit hours earned and TOEFL.
Krausz, Schiff, Schiff & Hise	2005	54 international students enrolled in the initial graduate financial accounting course in an American university	TOEFL	GPA	TOEFL scores were not significantly correlated with the grade in the course, *r*=-.06, *p*>.05.

As shown in the summary table, the predictive validity of TOEFL varied from one study to another, both in terms of coefficient strength and magnitude. This inconsistency in research findings regarding the relationship between language proficiency and academic attainment not only exists when TOEFL is used as a measure; similar disparity has been observed among IELTS predictive research.

2.5.7 IELTS and academic attainment

Earlier studies (late 1990s until early 2000s)

Cotton and Conrow (1998) were among the first group of researchers to examine how language proficiency as measured by IELTS related to academic attainment among 33 international students studying at an Australian university. In addition to GPA, Cotton and Conrow also asked staff teaching these students and the students themselves to provide ratings regarding their academic performance. Correlation analysis between IELTS overall and GPA was reported insignificant with a coefficient of -.24 (N=26); although there was a positive moderate correlation between IELTS reading and GPA, the relationship was not significant either. In addition, correlation between IELTS and staff rating of students' academic performance was not found significant (N=30), neither was there significant correlation between IELTS and students' self-rating of their academic performance. This overall insignificant correlation suggested that academic attainment of participants involved in this particular study was affected by factors other than language proficiency; or in other words, they had gained sufficient proficiency and thus their subsequent academic performance was no longer affected by the lack of such.

Meanwhile, the negative correlation between IELTS and GPA reported by Cotton and Conrow (1998) may seem very odd, but if look closer at the sample makeup, one could find that 3 participants (12%) with high IELTS scores (7+) performed poorly during their exams while 2 participants (7%) with low IELTS scores (5.5) obtained good GPA. Given the small sample size (N=26), this could have resulted in the negative correlation; the small sample size also limits its findings from wider implication.

The correlation between IELTS and staff rating (overall r=.15, reading r=.36, writing r=.34, listening r=.07, speaking r=-.33, p>.05) should also be interpreted with caveat because these staff ratings were based on different criteria. Some were based on students' written work, which were assumed to be more reliable, but other ratings based on students' performance during tutorials could be regarded less reliable. Moreover, correlation between IELTS and students' own rating of academic performance presented an even more somewhat odd picture: the first student self-ratings (end of semester 1) were in general negatively associated with IELTS (N=32, overall *r*=-.28, reading *r*=-.25, writing *r*=.28, listening *r*=-.31, speaking *r*=-.16, *p*>.05) but the second self-ratings (end of semester 2) were positively associated with IELTS (N=22, overall *r*=.12, reading *r*=.46, writing *r*=.39, listening *r*=.16, speaking *r*=-.16, *p*>.05). This difference, according to the researchers, can be attributed to the change in perceptions about oneself; when students first arrived at an alien environment, they might have low academic expectation of themselves. Over time, they may become more confident and hence in the second term, they became capable of evaluating themselves more positively and more in line with their language proficiency. Cotton and Conrow further cautioned that the interpretation of students' self-ratings should also consider the effect of factors such as culture and expectations. In short, although efforts have been made to explore the relationship between proficiency (measured by IELTS) and academic attainment, Cotton and Conrow's study offered limited insights due to sample size, research design and other uncontrolled socio-cultural, psychological and personal variables.

Using very similar research design as Cotton and Conrow, Kerstjens and Nery (2000) also examined the relationship between IELTS and international students' academic attainment through correlation analysis, interview and questionnaire. First semester GPA and IELTS scores were collected from 113 first year international students who were studying in the Faculty of Business at an Australian university; very similar to what Cotton and Conrow found, the overall correlation between IELTS and GPA was not reported significant in Kerstjens and Nery's study either, *r*=.028, *p*>.05. However, the reading and writing module of IELTS, were nevertheless significantly associated with GPA, although the effect size was small, *r*=.286, 250 respectively. Follow-up regression analysis showed that the overall model entering four IELTS modules as predictors was significant, accounting for 8.4% of the total variance in GPA but only IELTS reading was a found significant predictor GPA (*Beta*=.263). Furthermore, after reading has been accounted for, the significant correlation observed between writing module and GPA was not found in the regression. This indicates that the significant correlation between writing module and GPA could be regarded as common variance between IELTS reading and writing, rather than unique variance in writing itself.

In addition, questionnaires were used to elicit participants' perceptions concerning the adequacy of their IELTS scores and their language proficiency for their academic performance while interviews were conducted with faculty staff to seek their opinions on students' IELTS scores and general proficiency. Interestingly, students and faculty staff involved in Kerstjens and Nery's study agreed that the overall students' proficiency was adequate for their academic studies, which formed stark contrast with the literature reviewed in 2.2 where international students were found to have inadequate language skills (e.g. Ma, 2014; Robertson et al, 2000). This differences in findings might be attributed to the Kerstjens and Nery's research sample being relatively more proficient than the average of international students, but this assumption could not be validated as Kerstjens and Nery did not reveal the mean IELTS scores maintained by their research sample.

Staff interviews also revealed that the most important language skills needed for achieving academic success were listening to lectures and interpretative reading, which was consistent with the finding that IELTS reading was a significant predictor for GPA; those who were better at reading academic passages were more likely to achieve higher academic attainment.

The main limitation with Kerstjens and Nery's study was the study's longevity, i.e. only first semester GPA were collected as a measure of academic attainment. As discussed earlier in Chen and Sun's research (2006), the relationship between proficiency and academic attainment is prone to change over time. In that study, it was found that the correlation between TOEFL and first-year GPA was significant but later the correlation between TOEFL and last-year GPA became insignificant, which was partially in contrast with Kerstjens and Nery's finding, i.e. IELTS and first-semester GPA was insignificant. Regardless of this incongruity, as Kerstjens and Nery only looked at one GPA, could it be possible that IELTS turned out to be significantly correlated with students' academic attainment at a later stage of their academic journey, when, presumably, the demand to read literature and to write essay become more challenging than at the beginning? The answer remains to be found.

In short, the sole use of first semester GPA might not be able to provide an accurate indication of international students' academic attainment or an accurate depiction of the relationship between proficiency and attainment. More longitudinal studies are needed so as to provide more insight regarding international students' academic attainment and how it is related to their language proficiency as measured by IELTS. Moreover, Kerstjens and Nery's findings might have been further accepted by the heterogeneity of the research sample as participants were of different nationality (hence different L1), different English learning history and more importantly, different previous academic aptitude. All these variables were not controlled for in analytical process, which could have interfered with the outcome of Kerstjens and Nery's research and its generalisability.

In a very similar vein, Feast (2002) also collected GPA from five semesters among 101 international students who were studying in five disciplines with most students (47%) enrolled in a business faculty at an Australian university. Participants' overall IELTS scores ranged from 4.5 to 8.5, with the majority sitting between 6.0 and 7.0. To account for the large variation in the research sample in terms of discipline area, home country and level of study, which could have introduced noise affecting the analyses, multilevel regression was used to estimate the impact of English language proficiency, as measured by IELTS test scores, over time as students progressed through their studies. Multilevel analysis permitted a more appropriate and detailed intra and inter-student analysis of the relationship between IELTS and GPA than is possible with simple regression analysis using mean GPA scores. Results showed that at between-students level, IELTS had a significant relationship with GPA (t=2.92, p<.01), with a weak-moderate coefficient of .39, indicating that higher GPA was associated with higher IELTS, which was in contrast with findings from Cotton and Conrow (1998), Kerstjens and Nery (2000). This significant relationship, although weak, brings up the question that if students were admitted on the basis that they had provided sufficient evidence supporting their language skills, i.e. language proficiency should no longer be a barrier hindering their academic attainment, why was IELTS still significantly related to GPA? Could it be possible that IELTS requirement were set too low? Or could it be possible that other factors such as test preparation have interfered with the reliability and validity of IELTS scores by boosting scores without improving students' actual proficiency? Were this assumption to be true, the scores submitted for admission might not be an accurate indicator of proficiency, but rather students' testwiseness (further discussion related to testwiseness is provided in later sections of this chapter); thus, students may *appear to* have acquired the required proficiency as indicated by their IELTS scores, while in practice, their proficiency is still lacking. Moreover, if the sample of existing predictive study involved participants who attended test preparation and those who did not, this could complicate IELTS's predictive validity and the interpretation research findings. So far, to the best of my knowledge, no study has empirically examined the latter possibility.

Feast's study also reported that the variable China (i.e. participants with Chinese nationality) had a weak regression coefficient of .99 (*t*=2.41, *p*<.05) with GPA, suggesting that all other variables being equal, Chinese international students were more likely to achieve higher GPA compared to their international peers, which is in contrast with Morrion et al (2005), Iannelli & Huang (2013) who reported Chinese international students were less likely to achieve as good academic attainment both compared to home students and other international students (see 2.2). This differential finding could be attribute to the small sample of Chinese international students in Feast's study (N=7), limiting the generalizability of Feast's findings. Nonetheless, Feast's study provides very valuable insights on the relationship between language proficiency and academic attainment; more importantly, the significant correlation between IELTS and GPA led to the question of IELTS validity and the effect of test preparation which had not yet been thoroughly explored.

Feast's finding on IELTS predictive validity was largely consistent with Yen & Kuzma's (2009) IELTS predictive study. As one of the relatively few predictive studies in the UK, Yen and Kuzma examined the relationship between IELTS and academic performance of a group of Chinese students who studied business and management course at a UK HEI in the academic years of 07/08 and 08/09 (N=61). Correlation analysis revealed significant correlation (*r*= .46, *p*< .01) between the overall IELTS score and students' first semester GPA, as well as a significant yet weaker relationship (*r*=.26, *p*<.05) between the overall IELTS score and second semester's GPA, similar to what Feast and Chen and Sun (2006) reported above. This positive relationships between Chinese students' overall IELTS scores and their academic performance indicated the probability that Chinese students were "constrained" from achieving more academic success because of their low language proficiency. Moreover, significant correlations were also found between first semester's GPA and individual IELTS modules: Listening (*r*= .45, *p*< .01), Writing (*r*= .41, *p*<.01) and Reading (*r*=.27, *p*< .05). However, the correlation between the second semester's GPA and students

IELTS tests was less permanent – significant correlation was only found against Listening (*r*=.26, *p*< .05). The interpretation of these correlation findings need to, first, take into account range of the data. It is possible that the limited variance of the two variables examined in Yen & Kuzma's study affected the analytical outcomes. In addition, the homogeneity of the research sample should also be noted; as discussed earlier, with the increase in sample homogeneity, variance decreases, which means that the correlation coefficient observed in Yen & Kuzma's research was more likely to be the "true" indication of the relationship between language proficiency (as measured by IELTS) and academic attainment (as measured by GPA), as the noise from the data (e.g. nationality, disciplinary differences) were controlled.

These correlation results from Yen & Kuzma's work, to a large extent, resonated with that from Chen and Sun's study, in the sense that both studies found that language proficiency measured by standardised tests (i.e. IELTS or TOEFL) were more predictive of international students' academic attainment at the beginning of their academic journey, but as time proceeded, this predictive power seemed to "wear off". This "wearing off" of predictive power of proficiency could be interpreted as that students initially who lacked the proficiency to sustain their academic study eventually caught up, and at later stages, their proficiency (or rather, the lack of which) was no longer hindering their academic attainment. Given this, one may wonder, why would international students, admitted on the basis that they had acquired the needed proficiency (i.e. the needed scores on IELTS, TOEFL or other language tests), still constrained by their proficiency level? Was the IELTS requirement set too low? Regarding this, this study hypothesized that dedicated test preparation and repeated test taking together might have contributed to international students achieving the required scores on a proficiency test without correspondingly improving their proficiency level to match what their scores entail. More detailed discussion on test preparation and repeated test taking is provided later in section 2.6 and 2.7.

Around the same time as Feast's work, more contradictory findings have been reported, further complicating the relationship between proficiency and academic attainment. For example, Dooey & Oliver (2002) reported an overall insignificant correlation between IELTS and international students' semester weighted averages (SWA) (N=49); however, IELTS reading was found to be significantly associated with SWA, *r*=.273, *p*<.05 for semester 1 and r=.340, *p*<.01 for semester 2, consistent with Kerstjens and

Nery (2000) but differed from Feast (2002). The increase in correlation magnitude between IELTS reading and SWA overtime also echoed with Kerstjens and Nery's interview findings that as one progressed further into his/her academic programme, the ability to read became critical; those who could read more effectively stood a better chance of achieving better academic attainment. Although Dooey & Oliver set out to compare IELTS's predictive validity between linguistically more demanding discipline (in this case, business) and linguistically less demanding discipline (in this case, science and engineering), as correlation analyses yielded insignificant results between IELTS in all three disciplines, the between-discipline predictive validity were not probed deeper. However, as discussed before, when the research sample is small and of cross-disciplinary nature, the noise from the data could "mask" the "real" relationship between proficiency and academic attainment. Thus, if Dooey and Oliver had factored in this "noise" and conducted follow up correlation analyses between IELTS and SWAs separating linguistically more demanding participants from linguistically less demanding participants, the outcomes might be different.

It should also be noted that in Dooey and Oliver's study, in addition to examining the relationship between language proficiency and academic attainment among international students, native speakers of English were also involved in this study (N=23), which seemed inappropriate for the examination of IELTS's predictive validity, as IELTS was designed a measure for non-native English speakers. As expected, native speaker participants scored very high on IELTS but their academic attainment was not as good; 15 failed to achieve the minimum pass mark in both semesters, and four of these had no recorded grades for semester 2, seemingly having terminated their courses. Based on these findings, the researchers went on to claim that "while a language test such as IELTS can indicate a minimum proficiency level which students should attain in order to ensure a reasonable chance of success, it must be acknowledged that students are influenced by a range of other factors, many of which are outside the control of the receiving institutions" (p.50). However, Dooey and Oliver's conclusion is deemed problematic in the following aspects.

Firstly, Dooey and Oliver's study was built on the assumption that good language proficiency (indexed through IELTS) was a guarantee for academic success, for both native speakers of English and international students. This assumption was, to a large extent, invalid because language proficiency is not the premises for obtaining academic success; instead, the lack of proficiency is likely to hinder one's ability at achieving such success. Similar argument had been put forward by Vinke & Jochems (1993) who looked at the academic performance of international students enrolled in a Netherland university where English is the language of instruction. They stated that if students had a thorough command of English, proficiency would hardly be an impediment to academic performance. In this case the relationship between attainment and proficiency was expected to be weak and further improvement of proficiency would hardly affect academic attainment. On the contrary, the lower the level of proficiency, the more it would stand in the way of academic performance. Thus, the analyses on native speakers' IELTS and academic attainment in Dooey and Oliver's research should be interpreted with great caution. Meanwhile, the research sample involved in Dooey and Oliver's study, especially the native speaker participants, achieved very poor previous academic records, which could have subverted the research findings considerably. Furthermore, the small research sample also limited the findings from a wider application. Taken together, it is reasonable to conclude that Dooey and Oliver's findings are questionable in terms of reliability and generalisability, thus more research is called for to further scrutinize the relationship between proficiency (indexed by IELTS) and academic attainment.

Situated in similar context as Feast (2002), Dooey and Oliver (2002), Woodrow (2006) also explored the relationship between proficiency as measured by IELTS (N=62) and TOEFL (N=10) and international students' academic attainment in an elite Australian university. Also using a correlational design, Woodrow reported that IELTS was significantly correlated to semester 1 GPA, r=.40, p<.01, but the correlation between TOEFL and semester 1 GPA was not significant. Further exploring the data, Woodrow found that for participants who had just met the entry IELTS requirements (overall 6.5 and below, N=26), their semester 1 GPA was moderately significantly correlated with their IELTS scores, r=.52, p<.01. By contrast, for participants who exceeded the entry IELTS requirement, scoring IELTS overall 7 and above, their semester 1 GPA was not significantly related to their IELTS scores. This differences in correlational outcomes lent further support to the afore-stated argument, which is also the core of the present study, that language proficiency alone does not ensure academic attainment; rather, the lack of sufficient proficiency is a hinderance.

To sum up the literature on the predictive validity of IELTS on international students' academic attainment, one easy conclusion could the drawn: IELTS is not a consistent predictor. The predictive validity of IELTS could be affected by a number of factors, such as disciplinary differences in linguistic demand, and hence actions should be taken to account for these influences. Meanwhile, time might also moderate the predictive validity of IELTS in the sense that IELTS might be predictive at the beginning of students' academic journey, but this power could wear off throughout time. Further, this inconsistency of IELTS as a predictor could be the complication of the ever-growing IELTS dedicated test preparation industry and repeated test taking if the sample involved in previous predictive studies included participants who had attended test preparation and those who had not, which are explored in details in section 2.5 and 2.6.

2.5.7 Alternative measures of proficiency and international students' academic attainment

Albeit the inconsistency in findings on the relationship between language proficiency and international students' academic attainment when TOEFL or IELTS was used as a linguistic measure, research probing the same relationship using alternative measures seemed to form a much more unified picture: language proficiency is one of the key determinants, if not the most important.

For example, Trenkic and Warmington (2018) compared the language and literacy skills of British-home undergraduate students and international Chinese postgraduate students at a UK university and explored how such differences related to these two groups' subsequent academic attainment. Participants' general intelligence, vocabulary measures (size, and expressive vocabulary), word-processing accuracy, reading comprehension, spelling, written summarisation skills, and phoneme awareness were measured using a battery of tests at the beginning of their academic year (T1) and eight months into their programme (T2). To start with, the researchers found that there was highly significant and large difference in vocabulary size between Chinese international students (N=63) and British home students (N=64) along with significant between-group differences in reading comprehension, and written summarisation, two key indicators of higher literacy skills central for academic work at university level.

In terms of processing speed, it was found that Chinese students took significantly longer time than British students to read and verify the truthfulness of simple sentences and their accuracy at such verification was significantly lower than their British counterparts. These gaps observed at the beginning of academic year were not narrowed over time as ANOVA found insignificant between-group interaction. In other words, these gaps in language and literacy skills proceed to exist even eight months later.

These clear differences, particularly the ones at the beginning of the academic year, suggested that international Chinese students, although accepted on the pre-condition that they had obtained the required IELTS scores, still lacked the needed proficiency to handle the linguistic demand of studying at an English speaking university, if the language proficiency of involved British home students could be seen as the norm. Moreover, it should not be forgotten that the comparison drawn in Trenkic and Warmington's study was between British home undergraduate students and international Chinese postgraduate students; thus one can assume that the gap in proficiency between British home postgraduate students and international Chinese postgraduate students were even wider than what had been reported. Taking this a step further, it is reasonable to hypothesise that this gap in proficiency may put international Chinese students at an disadvantage in their academic study.

Correlation analysis showed that for Chinese students in Trenkic and Warmington's study, their T1 vocabulary size, and expressive vocabulary knowledge, word-processing accuracy, reading comprehension, spelling, written summarisation skills, and phoneme awareness were significantly associated with their academic grades while for British students, their academic outcomes were not significantly associated with most language and literacy measures (except spelling errors). This difference indicated that language proficiency alone does not "guarantee" academic success but the lack of it could act as an obstacle hindering one's ability to do so. Following this logic, it is reasonable to expect the IELTS would be predictive of international Chinese students' academic attainment.

Indeed, IELTS had a robust association with academic grades ($F(2,59)=6.80$, $p=.002$), with each drop of half a point in IELTS band score corresponding to a drop of about 4 points in grades: participants entering with IELTS 7.5 (N=12) achieved a weighted average of 65.58 (SD=8.69), those coming with IELTS 7.0 (N=29) averaged 61.70

(SD=5.29), and those with IELTS 6.5 (N=21) just 57.24 (SD = 6.44). Regression model using general intelligence, vocabulary (composite score from vocabulary size and expressive vocabulary), high literacy (composite score from reading comprehension and written summarisation results), speed of processing, spelling and phonological processing measures as predictors (entered in the presented order) accounted for over 50% percent of the variance in Chinese students' academic outcomes, *F*(6,51)=8.87, *p*=.000. Among the entered predictors, the unique contributions of vocabulary (16.81%), higher literacy skills (9.55%), speed of processing in English (6.30%), and spelling (4.16%) were statistically significant. Contrastingly, the same model only accounted for 10.7% of the variance for the British group, *F*(6, 48) = 0.96, *p*=.46 and none of the entered predictors contributed unique significant variance to the overall model, showcasing again, that language proficiency alone does not guarantee academic success but the lack of it could hinder one from doing so.

Trenkic and Warmington's research findings highlight that significant differences existed in language and literacy skills between international Chinese postgraduate students and British home undergraduate students; in comparison, Chinese postgraduate students not only read more slowly, but also understood less of what they read due to their limited vocabulary knowledge and speed of processing. The correlation between IELTS and academic grades, coupled with the large proportion of variance in international Chinese postgraduate students' academic outcome accounted for by the linguistic measures highlighted that students admitted on the precondition that they had met the IELTS requirement, did not seem to have acquire sufficient language knowledge or skills, thus affecting their academic attainment. Regarding this, one could argue that the IELTS admission requirement was set too low; in addition, one may also infer that means adopted by these international students to achieve the required IELTS scores, such as test preparation and repeated test taking, might have contributed to the wide gap in language and literacy skills between international Chinese postgraduate students and British home undergraduate students.

Findings from Trenkic and Warmington's research corroborated with a number of other studies that adopted alternative linguistic measures in addition to TOEFL and IELTS when examining how proficiency related to international students' academic attainment. As discussed in section 2.3.3, vocabulary has often been put forward as a good indicator of overall proficiency (e.g. Andringa et al., 2012; Schoonen et al., 1998; Schoonen et al., 2003; de Jong et al., 2012); it does not come as a surprise that a handful of research had examined how vocabulary knowledge relate to academic attainment.

Daller & Xue (2009), for example, used C-test (more detailed information of this test is presented in the following Methodology chapter) to measure the vocabulary knowledge of Chinese students attending a British university in addition to IELTS. The same C-test was administered 6 months prior to the beginning of their postgraduate programme (N=21, February, 2004) and again in September when these Chinese students arrived in the UK (N=20). Participants were also asked to write an essay and their produced work were analysed for lexical diversity and lexical sophistication. Academic success was operationalised in two ways: 1) whether a student failed at least one module in the first year or passed all modules at the first attempt and 2) the number of failed modules in the first year. Initial analyses found a moderate correlation between IELTS scores and the number of failed modules (*r*=.382, *p*<.05, N=23) and an even stronger negative correlation was found between the C-test scores from February 2004 and the number of failed modules (*r*=.565, *p*<.01, N=21). Subsequent regression analyses revealed almost 40 per cent of the variance in the number of failed modules during the first-year taught master's programmes can be predicted from Chinese students' C-test scores. Furthermore, logistic regression using IELTS as the sole predictor found that failure of at least one module could be predicted by IELTS scores. The lower the IELTS scores, the higher the risk of failing at least one module. IELTS scores also explained about 11 per cent of the variance of the number of failed modules. These results were to a large extent consistent with what was found by Trenkic and Warmington and illustrated again, that international Chinese students accepted on the premises of their IELTS, seemed to lack the proficiency needed to achieve academic success. Nevertheless, it is important to note that Daller and Xue's research sample was very small, N=23, which means that findings could be limited in terms of generalisability.

In the light of the small research sample, Daller & Phelan's (2013) later re-examined the relationship between proficiency and academic attainment among 69 European countries, 4 from China and one from an Arabic country. C-test was used again in this study as a measure of general language proficiency (administered twice, both at the beginning and the end of the academic year); the subtest V5 of Sigma testing system[10] was used as a measure of verbal intelligence; IELTS listening practice test acted as a measure of listening skills; and lastly, an online writing task on environmental issues

[10] For more information on V5 of Sigma testing system, see http://www.sigmaassessmentsystems.com/assessments/multidimensional-aptitude-battery-ii/

was incorporated to assess participants' holistic writing ability as well as their vocabulary size. Apart from C-test, all measures were administered once at the end of the academic year, except IELTS listening, which was at the beginning. Through initial correlational analyses, no significant correlation was found between these linguistic measures and GPA, which seemed to suggest that this group of participants had indeed acquired sufficient proficiency and thus their academic attainment was no longer affected by the lack of which. However, once the students with GPA lower than 40 were excluded from the dataset (labelled GPA adjusted) because they were regarded as not making serious attempts at their examinations, there were significant correlations between GPA adjusted and some of the predictors such as C-test 1 (beginning of the year), r=.432, p<.01, C-test 2 (end of the year), r=.616, p<.05, listening score, r=.803, p<.01, and writing score, r=.353, p<.05. Regression analysis reported that the final model including vocabulary size (indexed by Guiraud Advanced), verbal intelligence (indexed by Sigma V5) and C-test accounted almost all of the variance in GPA, F=38.335, p=.000, R^2=.958.

This particularly large proportion of variances explained by vocabulary size, verbal intelligence and C-test is understandable as the majority of participants (N=61) from Daller and Phalen's study were engaged in linguistically more demanding disciplines, e.g. Humanities, Law and Business. It is reasonable to assume that studying in these disciplines involved a considerable amount of reading and it is likely that students' academic progress were assessed through written assignments. Given this, it is not surprising that participants who had higher proficiency achieved better GPAs. Having said so, alternatively, findings from Daller and Phalen's study also pointed out that language proficiency was indeed at the heart of international students' academic performance. Although they were accepted onto the programme on the prerequisite that they had achieved a certain level of proficiency, this level did not seem to suffice the linguistic need from their subsequent academic study, especially when students were engaged in linguistically more demanding disciplines.

Similar findings from predictive using alternative measures of proficiency could also be seen from Ho & Spinks (1985), Bayliss & Raymond (2004), which leads to the question, what could have contributed to this difference in predictive power between standardised tests and alternative measures of general proficiency?

To sum up the literature on the relationship between proficiency and international students' academic attainment, the following conclusions could be drawn. First and foremost, standardised tests such IELTS and TOEFL were not a consistent predictor for academic attainment even although language proficiency, clearly, plays a role in international students' academic attainment as many studies had found significant correlation between proficiency as measured through standardised tests such as IELTS (e.g. Feast, 2002; Trenkic & Warmington, 2018) or through independent linguistic measures, such as vocabulary (e.g. Daller & Xue, 2009). However, as participants involved in most of these review predictive studies, if not all, were international students enrolled in an English speaking HEI, it is reasonable to assume that the majority of them were admitted on the premises that they had acquired the needed proficiency level (usually indexed by scores on a standardised proficiency test) to handle the linguistic demand of their subsequent academic journey. This is to say that, for the admitted international students, if their test scores were "true" representation of what their proficiency level was, language should no longer affect their subsequent academic performance. Or simply, there should be no significant correlation between academic attainment and proficiency during their course of academic pursuit, which contradicted the findings of afore-reviewed empirical research. As stated earlier, this contradiction could be explained in two ways. On the one hand, it could be because the language requirement set for admitting international students was too low; on the other hand, it could also be because that participants had resorted to alternative means to somewhat achieve a score beyond their "true" proficiency. Were the latter to be the case, it offered a reasonable explanation as to 1) why students who achieved the needed scores on tests were still affected in their study by the lack of proficiency 2) why the predictive power of standardised proficiency tests such as IELTS and TOEFL were not consistent.

There are, of course, many means to achieve the desired scores on a test and here in this study, the focus is on the two most common ones: dedicated test preparation (as known as *coaching*) and repeated test taking. Relevant literature on these two practices are reviewed next in section 2.6 and 2.7.

2.6 Test preparation and relevant concerns

Test preparation, also known as *coaching*, involves a variety of activities ranging from short term cramming using sample test items on one extreme, to the "long-term instruction aimed at knowledge and skill development at the other" (Messick, 1981, p.3). Conventionally, coaching for scholastic tests tended to fall on the former side of the continuum, focusing principally on test familiarisation and test taking strategy instructions, while coaching for achievement tests was more geared towards the development of content knowledge of the subject. Messick (1981) published a meta review synthesising the findings from 39 key scholastic tests coaching research; although scholastic tests differ from language proficiency tests, the heart of the present study, given the lack of literature on proficiency test coaching and the important insights Messick's meta review provides, this section frames the discussion of proficiency test preparation following the Messick's guide (1981), beginning with the definition of what constitutes coaching/test preparation.

Quintessentially, coaching refers to "any intervention procedure specifically undertaken to improve test scores, whether by improving the skills measured by the test or by improving the skills for taking the test, or both" (Messick, 1981, p.13). From this, it is easy to conclude that the core of test preparation lies in the achievement of score gains.

The nature of test preparation practices falls onto a continuum of acceptability. According to Mehrens & Kaminski (1989), general instructions related to the test objectives and general testwiseness development are categorised as acceptable practices, while using published parallel tests for drilling practice is regarded unacceptable.

In a similar vein, Messick (1981) categorised test preparation instructions into three types. Type A preparation mainly focuses on "test taking sophistication or reduced test anxiety" (p. 49); for example, in a test preparation session, the teacher shows the candidates the test format and point out the time constraints set out for the overall test and each componential module so that the candidates know what to expect when they enter the test room. Linking this type of test preparation practices with validity discussed previously in 2.5, it is reasonable to infer that type A preparation does not pose serious threats to the validity of a test. In fact, it is very likely that type A preparation can enhance test validity because it cancels out the interference from test anxiety, thus making the test score a more accurate description of learners' ability. Type B preparation aims to help learner achieve genuine improvement in the construct that the test sets out to assess. For example, in a language proficiency test preparation session, the teacher teaches vocabulary that is not explicitly assessed in the proficiency test but is commonly found in the target context of the test. This type of preparation is also believed not to pose threats to test validity as score gains achieved through type B preparation would correspond to improvement of the ability. Type C preparation, which is deemed problematic, centres around "test-taking stratagems and answer-selection tricks, resulting in scores that are inaccurately high" indication of students' ability (p.49). For example, in a proficiency test preparation session, the teacher teaches candidates to recite chunks so that candidates can quickly produce a speech when taking the speaking module of the test or instructs candidates to analyse the grammatically relationship between the stems and options in a multiple choice question to determine the most likely answer. From this, it is clear that, contrasting to type A and B preparation, type C preparation poses threats to the validity of a test and complicates the inferences drawn from the test scores.

Messick (1981) further added that there are certain demographic patterns regarding which type of preparation is preferred than the others. Type A preparation is oftentimes used by students who are preparing a test on their own; type B is usually present among commercial coaching schools and type C prevails among professional test preparation organisations, where the present study was situated in.

Typically, in type C test preparation programmes, in addition to test familiarisation (e.g. familiarity with the test format and knowing the type of item that is being presented, time planning, physical and emotional preparedness), learners receive specific instructions on testwiseness strategies such as looking for clues, process of elimination direction following, knowing the type of item, grammatical relationship between the stem and the answer (Summers, 1983; Frankel, 1983; Phillips, 1983; Ligon, 1981; Diamond, 1972; Bangert-Drowns, 1983; Kilan, 1992; George, 1985; Mehrens & Kaminski, 1989). Additionally, instruction is often misdirected in high-stake test preparation classrooms, resulting in a of tasks that resemble tests and skills are often taught in isolation and mostly through drilling and practicing sample test items (Shepard, 1991). Furthermore, in type C test preparation, curriculum is narrowed to focus primarily on the contents that are likely to be assessed, leaving those unlikely-to-be-tested areas unaddressed (Tunks, 2001).

Heated controversies have emerged over the effect and impact of proficiency test preparation and what this could mean for proficiency test validity, one of the key discussions the present study attempts to engage with. Existing literature on the topic on proficiency test preparation follows two main strands, one being how tests affect teaching and learning (i.e. washback or backwash), the other being how test preparation affect test scores. The majority of test preparation is situated in the former strand, using mainly qualitative methods while the amount of literature on the latter is considerably more limited; together both strands provide valuable insights on the nature and impact of tests and test preparation. As the present study is more concerned with the quantitative outcome of test preparation in the realm of proficiency tests (i.e. the latter strand), only a brief discussion of washback is provided.

Before proceeding to the discussion of washback, test preparation and score gains, a summative description of a typical IELTS preparation course is presented first to depict the common preparatory practises. This description is necessary and important because it relates to the intervention investigated in the first quasi-experiment of the present study.

2.6.1 A typical IELTS preparation course

According to Mickan and Motteram's (2008) classroom observation research, for a typical IELTS preparation course, the pedagogy is teacher directed and test focused, dominated by test practice, skills-focused activities, and explanations of the format and content of the IELTS modules and of test-taking procedures. Students worked together on exercises, on analysis of sections of the test and on preparation of answers to test questions. Although collaborative learning is present, the focus of learning is still on learning of the test and gaining testwiseness.

Regarding each individual module, for speaking, there is substantial practising of English specifically related to the IELTS speaking tasks and giving advice on what to do in the test situation. The teacher, first, modelled talking on topics, then set discussion tasks on topics students might encounter in part 1 IELTS Speaking, followed by the learning of the structure of an argument genre, which is in line with part 2 & 3 IELTS Speaking. Moreover, explicit instructions were given concerning the expectations held towards students and specific techniques and strategies they might adopt in their responses such as the use of fillers to avoid communication breakdown or paraphrasing (Mickan & Motteram, 2008). On one hand, these practices reduce construct irrelevant variances such as the inferences from nervousness and hence enhances the validity of IELTS Speaking. On the other hand, narrowing the activities to focus on potential test tasks and questions could also be seen as learning of the test, which introduces construct irrelevant variances such as one's ability to memorise more potential test tasks, undermining the validity.

Similar pedagogy was observed in IELTS Writing lessons; in addition to test format familiarisation, sample essay was frequently used as a model for learners to imitate (Mickan & Motteram, 2008). This practice alone is controversial. On one hand, students, especially those with lower proficiency, could improve their overall writing proficiency because imitating and modelling can enhance lexical and grammatical knowledge. On the other hand, this could also be argued as learning of the test; for students with higher language proficiency in particular, exemplar essays may limit their creativity and constrain outside-boxing thinking, both of which are skills needed for achieving academic success at HEIs (Rickett & Rudd, 2005; Stupnisky, Renaud, Daniels, Haynes & Perry, 2008). By doing this, it brings in both construct irrelevant variances such as one's ability to imitate model essay into the IELTS Writing module and construct underrepresentation variances, i.e. students might perceive academic writing a simple task of model imitation, into the test, both posting risks towards the IELTS validity.

For IELTS reading, the emphasise of learning is on lexical knowledge (Mickan & Motteram, 2008), which is reasonable given the important role lexis plays discussed previously in 2.3.3. Local and global level reading strategies, such as reading to understand the gist, underlying key words to improve comprehension and using context to guess the meaning of unknown words, were also taught. What needs to be distinguished here is that almost all classroom activities were centred around sample tests from IELTS preparation materials. Moreover, learners took timed practice tests on a routine basis, both to familiarise themselves with the test format and time constrains as well as to master the newly taught test taking strategies.

Lastly, for IELTS listening preparation, substantial drill-and-practice was observed using preparatory IELTS listening material. In particular, learners were told through frequent and repeated practise, students would be able to be more confident in listening, "see the pattern in the exam" (Mickan & Motteram, 2008, p. 19). This suggests that the teaching is primarily focused on how to obtain the correct answer to the test item or in other words, learning of how to *beat* the test. Furthermore, "the teacher advised the students they would be able to find patterns for locating the information for their answers" (Mickan & Motteram, 2008, p. 19). In other words, learning and teaching in test preparation are product-oriented, i.e. the goal of listening was to obtain the correct answers; more importantly, students were encouraged to rely on the so-called patterns rather than improving their overall listening skills.

Based on Mickan and Motteram's research (2008), the following conclusions can be drawn: 1) there is no doubt that test preparation courses are distinctively different from "normal" English language courses; 2) practices and activities in test preparation are centred around test papers; 3) narrow learning of the construct is frequently observed and 4) improvement of proficiency is not a priority. All four conclusions are intrinsically related to test washback, or back wash, to be discussed in the following section.

2.6.2 Test washback and test preparation

The notion of washback, which relates closely with test impact (Baker, 1991), consequential validity (Messick, 1989) or systemic validity (Fredericksen & Collins, 1989), was put forward by Alderson and Wall (1993) through their washback hypothesis: "that tests influence teaching" (p. 120). Alderson and Wall (1993) believe that tests have the power of determining "what happens in the classrooms" (p. 41). Buck (1988) later referred to washback as "a natural tendency for both students and teachers to tailor the classroom activities to the demands of the test, especially when the test is particularly important for candidates" (p. 17). The washback effect of tests on teaching has attracted considerable attention over recent years, but the critical question of how this translates into washback on learning remains under-explored. Such influence is of relevance to the present study as it is closely associated with test validity; more specifically, if a test influences teaching, which consequently influences learning, then it is possible that test of learning becomes the learning of the test, which affects the construct under assessment. For example, for language proficiency tests, e.g. IELTS and TOEFL, if the hypothesis tests influences teaching stands, to ready the students for these tests, one would expect the content of teaching and learning may shift from focusing on the language skills, the intended constructs to be assessed, to the test-taking skills and testwiseness, the unintended construct irrelevance; this shift may consequently undermine the construct validity of IELTS and TOEFL.

In line with this hypothesis, classroom observation showed that there were substantial differences in the approach and organisation of test preparation and non-test preparation classes (Alderson & Hamp-Lyons, 1996, Erafni, 2012). Test preparation classes were characterised with more test taking practices, more teacher talking, less turn-taking and much less pair work time (Alderson & Hamp-Lyons, 1996). In terms of choice of language of instruction, it was found that test preparation courses prefer the use of learners' L1 while non-preparation courses made more use of the language being assessed (Erafni, 2012). Although the use of L1 can be more effective for testing test taking strategy, it limits the linguistic target language input necessary for learners' proficiency development. However, this is not to say that the observed test preparation classes were necessarily less preferred than non-preparation classes as the researchers did not make attempts to elicit perceptions from learners. Moreover, due to the observational nature of Alderson & Hamp-Lyons' work, it is difficult to determine the degree of influences test preparation exerts on validity.

In addition, reasons and expectations of learners enrolled in test preparation courses were found different from those enrolled in other types of English courses. Analysing data elicited from 108 mainland Chinese EFL learners preparing for university study in the UK (75 engaged in six non-IELTS courses and 33 studying on seven dedicated IELTS preparation courses), there were clear differences in learners' reasons for taking the programmes they were enrolled in. While 69% of non-IELTS students indicated that their primary aim was learning useful skills for university, the IELTS preparation students were evenly split between those studying for the purpose of obtaining good IELTS scores and for acquiring useful skills for university. Apparent gaps were also observed in expectations both groups held. Non-IELTS students were more motivated to learn general academic English skills, such as "*how to communicate ideas effectively in writing, how to find information from books to use in writing essays, quick and efficient ways of reading books in English and how to write in a formal, academic style"* (Green, 2006, p. 121). Although IELTS students also expected teachers to provide error corrections and learning of how to write successful test essays, they expected classroom activities to be in line with IELTS tasks, so as to achieve test familiarisation. Interestingly, IELTS students did not expect to be solely engaged in IELTS-like tasks and they stated test taking strategy learning was not their only concern. In fact, they prioritized the development of writing skills over test practice. From this, one could map the expectations IELTS preparation learners had to the type A and B preparation practices presented earlier by Messick (1981) and were these expectations met during the investigated preparation courses, validity would not be compromised.

However, these expectation were not found met. Questionnaire data collected at the end of both programme revealed substantial differences in terms of learning contents. For instance, while non-IELTS students learned skills of argument organisation, the use of references and the production of extended texts, IELTS learners learnt specific words and phrases for describing graphs and diagrams; both are closely related to IELTS Writing task 1, which requires candidates to describe iconic data in the form of a graph, a table or a diagram (Green, 2006). This test task oriented learning, also regarded as narrow learning or learning of the test, maps to Messick's (1981) type C test preparation. Therefore one may speculate the validity of IELTS writing was compromised because unintended construct-irrelevance variances such as students' ability to memorise test specific lexis had been introduced. Nonetheless, such speculation remains untested because Green's work did not examine how much improvement in learners' writing proficiency was indeed achieved through attending preparatory programmes.

Further on the note of test preparation and general proficiency, test preparation course teachers did not believe that what they taught actually helped learners improve their general English skills (Erafni, 2012). Teachers viewed the courses as highly test-oriented, emphasizing memorization rather than communication. Moreover, teachers commented that even learners themselves did not care about communication skills and they were simply after a passing score in a short period of time. This resonates with international Chinese students perception from Ma & Cheng's research (2015) on the value of TOEFL test preparation, of which detailed discussion is presented in 2.8.2. Together, perceptive evidence from both teacher and learners indicate that improvement of proficiency is not prioritised during test preparation, although whether test preparation may lead to incidental proficiency improvement remains unclear.

To sum up, given the descriptive nature of many washback studies, it remains unclear whether such learning-of-the-test can indeed boost students' performance on IELTS with or without the corresponding improvement of their general proficiency. To explore the effects of test preparation on test validity, studies that quantitatively measure the change in test performance are needed to establish whether pedagogical changes made due to test washback pertain to candidates' test outcomes.

2.6.3 Preparation courses for proficiency tests and effects on test scores

When candidates attend dedicated courses to ready themselves for the test, it is only natural for them to expect improvement on their test performance through such study. Regarding this expectation, IELTS recommended that "individuals can take up to 200 hours to improve by one IELTS band" (Green, 2005, p. 22); this recommendation was accompanied by caveats; score gains were said to be affected by learner characteristics such as age, motivation, first language and educational background and gains were said to vary with level of proficiency (Green, 2005). Similar recommendations are made in *The BALEAP Guidelines on English Language Proficiency Levels for International Applicants to UK Universities* (Bool, Dunmore, Tonkyn, Schmitt, & Ward Goodbody, 2003, p. 5) which indicated that three months of intensive English study is said to be sufficient to prepare a student presenting an IELTS score of 5.5 for entry to a "linguistically demanding" course with an entry requirement of IELTS band 7. Nonetheless, it should be highlighted that on what grounds the 200-hour learning or the three-month training suggestion was based remained unclear.

The amount of literature examining the coachability of proficiency tests such as IELTS and TOEFL is very limited, and even less is focused on the effect of score improvement. Among this limited pool of research, a considerable proportion is situated in English-speaking countries and only focused on one component of the test, instead of covering all componential modules (e.g. Archibald, 2001; Brown, 1998; Green, 2005, 2006, 2007; Hayes & Read, 2003). However, given that most test preparation practices take place in ESL/EFL countries prior to candidates' gaining entrance into an English-speaking country as IELTS and TOEFL scores are often set as part of their academic visa application requirement, this lack of ESL/EFL proficiency test preparation research indicates that the effects of test preparation on proficiency test performance is not well-understood.

Test preparation research in English speaking countries

Using data collected from a sample of international students (N=476) preparing for academic study at fifteen UK institutions, Green (2007) examined whether dedicated IELTS preparation course (N=85) was more effective in boosting candidates' scores than a university non-IELTS EAP course (N=336) and a combination course covering both IELTS writing and EAP (N=60).

Results revealed that participants attending all three courses made significant progresses in their IELTS Writing scores, achieving an average of .207 band increase. Participants from combination course who scored the lowest at the pre-course IELTS Writing, made the most gain in their scores, which indicated that the ability to achieve score gains can be related to candidates' pre-course proficiency level. In other words, those who had lower proficiency when they enrolled in the preparation courses, are hypothesized to make bigger score gains than those who had higher pre-course proficiency. More importantly, no significant difference in writing score gains achieved between course types was found. Further examining the predictive contribution of participant-related variables revealed that those "who had low initial writing and grammar scores, studied on longer courses, were educated beyond secondary level and believed that they were good at learning to write in English, would achieve the highest writing score gains" (p. 91), which is in line with the hypothesis. Pre-course Writing scores, grammar and vocabulary scores alone accounted for 38% of the variance in scores gain achieved, which supported Elder & O'Loughlin's claim that the language proficiency one had prior to the beginning of the course is the most constant indicator of how far one is likely to "travel" (2003, p. 226). By contrast, adding course type did not contribute to any improvement in the prediction of score gains, which is to say that participants attending IELTS course or combination course did not obtain any measurable advantage over those attending EAP courses in terms of test score gains.

In addition, degree of regression to the mean was observed: the higher the pre-course writing score, the lower the gain, which was in line with the aforementioned hypothesis. Regression to the mean is a statistical phenomenon that refers to the extreme cases at the tails of a normal distribution when measured for the first time will move closer towards the mean at further measurements (Barnett, Van Der Pols & Dobson, 2004). For example, in Green's study, participants that scored much lower than the mean score of the group prior to attending these courses were likely to score closer to the mean score at the post course measurement. In other words, the bigger the difference between the individual pre-course score and the group pre-course score, the more likely the participants were to achieve greater score gains so as to be closer to the mean.

This regression to the mean seems consistent with the frequently observed plateau effect in language learning (Brown, 1998): at lower levels of ability, relatively short periods of instruction can result in measurable improvements in proficiency, but at higher levels considerably longer periods are usually required; however, can score gains be regarded as improvement in proficiency? Is it probable that score gains achieved through test preparation are not reflection of improved proficiency, but rather testwiseness (e.g. enhanced test familiarity, better test taking strategies, as discussed in section 2.6.2)?

Answers to this question can inferred from the significant correlation reported between test-related practices ("The activities we did in class were similar to the ones on the IELTS test", Green, 2007, p. 90) and achieved score gains. As the only aspect of courses that had a significantly positive relationship with score gain, it shows that higher score gains achieved by participants were associated with more test related drill and practice. The nature of such practice is often testwiseness-oriented (i.e. type C preparation, Messick, 1981), and is believed to post threats to the construct validity of the test.

Interpretation of Green's findings should take into account that his study was situated in an English-speaking context, which means that the observed score improvement might be relevant to factors other than attending courses, such as sharing a house with English-speaking roommates, frequency of using library, number of close English-speaking friends. If the study was situated in a non-English speaking country, where the majority of IELTS preparation population is, it may be also to provide a much clearer and more accurate picture regarding the effects of IELTS preparation.

Similar findings were reported by Brown (1998) who also compared the effectiveness in inducing IELTS writing score gains through IELTS dedicated preparation and general EAP course provided by an Australian university. Through the ten-week intensive IELTS preparation, the average score increased achieved by the IELTS group (N=14) was almost a whole band, (M=.94, SD=.83, Range=0-2.70). EAP group (N=9), however, did not make significant progresses. Regarding this between-course differences in terms of score gains, Brown argued that it could be attributed to the disparity in participants' pre-course proficiency, i.e. preparation participants had lower proficiency and hence were more susceptible to test preparation than EAP participants who already had a high proficiency. Though this resonates with the afore-discussed plateau effect, closer examination of pre-course proficiency of EAP participants revealed that Brown's argument may not be true, as the majority of their IELTS writing sat between 5 and 5.5 (M=5.32). This average pre-course score can hardly be regarded as indication of high proficiency, thus casting doubt on Brown's explanation. Drawing reference from literature discussed in previous sections (e.g. Green, 2007), it can be argued that the differences in score gains were attributed to the different practices used in these two course. For example, it is very possible that IELTS preparation courses in Brown's study involved more test-oriented practices whereas the EAP course did not, resulting in bigger score gains.

Apart from writing, other modules of IELTS have also caught researchers' attention. For example, Issitt (2008) also examined how subjective IELTS speaking was to specific IELTS speaking teaching among 35 international students enrolled in a UK university. Similar to the typical IELTS preparation course presented in 2.6.1, strategies relevant to improving confidence and familiarising with marking criteria were taught explicitly during the course in Issitt's study. Although 35 students attended the course, only 13 actually sat the IELTS test and only 8 of whom had a pre-course IELTS score.

Results showed that of the eight students who had a pre-course speaking score, five increased their grades by 1 band; two did not change their scores and one student's score decreased by 1 band. These findings were largely limited in implication and generalisability due to its small research sample; nonetheless, clear improvements in speaking scores were observed as a consequence of IELTS learning. More importantly, although Issitt argued that explicit learning of marking criteria used in IELTS speaking could be regarded as "a demystification process", "one way of inducting students into the UK academic culture and at the same time addressing their immediate needs, that of scoring highly in the IELTS exam" (p.136), one could also argue for the possibility that learning of marking criteria could help students to game the test (i.e. gaining testwiseness), allowing them to know what to say and how to say it. Regarding this, evidence from Issitt's was insufficient to offer a clear explanation as to why explicit IELTS learning could results in score gains; was it due to the increase of testwiseness, the familiarisation of future academic contexts and demand or simply, the improvement of general language ability? Answers to these questions have not been provided by previous researchers.

Unlike Green (2007), Brown (1998) and Issitt (2008) who only looked at the effects of test preparation on one particular IELTS module, Elder and O'Loughlin (2003) made an attempt to examine how susceptible IELTS is to intensive EAP courses. Although the goal of EAP courses differ from that of dedicated IELTS preparation, with the former focusing broadly at improving general academic language skills while the latter targeted specifically at boosting candidates' test scores, findings from Elder and O'Loughlin's work shed valuable lights on the coachability of IELTS and thus is deemed worthy of inclusion in this section.

Elder and O'Loughlin found that on average, an .598 (SD=.545) score increase in IELTS overall was observed, followed by a .781 (SD=.972) mean score increase in listening, a .545 (SD=.948) in writing, a .402 (SD=.729) average increase in reading, and .500 (SD=.930) in speaking; all these increases were reported statistically significant at .000 level. When participants' pre-course IELTS band was taken into consideration, it was found that the mean score gains achieved at post-course decreased as the pre-course score band increased, lending support to the aforementioned plateau effect.

The most predominant score improvement observed in Elder and O'Loughlin's study was in the only centre that included IELTS preparation component. Thus there is good reason to hypothesise that explicit IELTS preparation can lead to more score gains.

Research reviewed in this section shared the same problem: they were all situated in an English-speaking country, which meant that observed the score increases might be attributed to other non-course related factors, such as whether the students were living at an English-speaking homestay, or the students made many English-speaking friends, hence more exposure to the target language and bigger score gains. Given that most test preparation takes place in a EFL context prior to taking the test, as students often need results from IELTS and TOEFL for visa application so as to entre an English-speaking HEI, it would be more apt to investigate the effects of test preparation in such context, controlling for the intervening affect from factors such as exposure.

Test preparation research in EFL countries

Even less literature can be found on the effect of test preparation for proficiency tests in EFL contexts, particularly for IELTS. Studies that are there, are often limited either by the incomprehensive research design (e.g. lack of control group which means there was no base line against which the effect of intervention can be compared to) or simply, by the small research sample. Hence, findings from research reviewed in this section should be interpreted with these caveats in mind.

Situated in Iran, Bagheri and Karami (2014) looked at the effect of explicit teaching of listening strategies on Iran EFL learners' IELTS listening scores. 40 advanced EFL learners who attended a 3-month IELTS course took part in this study, of which 20 formed the control group and the rest formed the experiment group. A practice IELTS listening test paper was used to assess the pre-course listening proficiency of these learners so that to ensure the sample was homogenous; all leaners scored around Band 5.5 at pre-course, although specific data were not provided. Similar to the typical IELTS preparation course described in 2.6.1, learners were taught explicitly how to underline key words in given tests and then predict what they need (e.g. noun, adjective, number, special name) and guess the related information to answer the questions; they then were given opportunities to practise these strategies repeatedly throughout the course.

Results revealed that there was a significant difference in the number of correct answers obtained by both groups at a .05 level, *t*=-13.114; the preparation group answered 35.45 (SD=1.504) items correctly (out of a total of 40 question items), which according to IELTS listening scoring guide corresponds to Band 8, while the control only got 28.20 (SD=1.963) items right, Band 6.5 (IELTS scoring in detail, 2018). Although the this study seems to indicate that to a certain extent, explicit teaching of listening strategies could improve leaners' IELTS listening score and had positive effect on their final performance, due to the small sample size, the incomprehensive research design, and the limited information provided by the researcher, the reliability and validity of the findings can be questionable.

Also situated in test preparation programmes in EFL countries, Robb and Ercanbrack (1999) adopted a comparative approach and investigated whether targeted TOEIC (Test of English for International Communication) preparation course was more effective in inducing score gains than non-TOEIC preparation courses (a Business English course and an EAP course). Among a group of Japanese candidates (N=365), using a pre-test/intervention/posttest design, results showed that, in general, participants enrolled in all three types of courses achieved very similar score gains. Further analyses revealed that for non-English major candidates, presumably of low proficiency, TOEIC preparation did result in significantly more reading score gains, in comparison to the other two courses. By contrast, for English major participants, who were presumed to have high proficiency, score gains achieved through attending all three courses were similar, i.e. TOEIC preparation course was not more effective in boosting scores than Business English or General English courses. This discrepancy can be related to the previously discussed plateau effect (e.g. Elder and O'Loughlin, 2003; Green, 2007), suggesting that the effectiveness of test preparation in terms of score gains might rely on the pre-course proficiency candidates had.

So far, literature on proficiency test preparation suggests that proficiency test scores are indeed subjective to explicit test learning and it seems that substantial score gains are achievable within a relatively short period of time. What remains unclear is whether such gains are a result of improved proficiency and higher language skills or simply, increased testwiseness? Moreover, if score gains are found to be testwiseness-related, what implication can this have on IELTS test validity? Although no research to date has directly addressed these two questions, references can be drawn from Gan's study on IELTS preparation course and IELTS score change (2009).

Situated in Hong Kong, Gan used a questionnaire to elicit information about IELTS preparation experience and test performance among 146 undergraduate students who sat IELTS. 56 (38%) of the sample had attended IELTS preparation and results revealed those who attended IELTS preparation (N=56) had significantly lower pre-course proficiency (indexed by their A-level English grades) than those who did not attend preparation (N=84). At the end of their three-year university studies, participants of both groups took the IELTS test as an exit English language proficiency test and mean overall score of IELTS preparation participants was almost identical to the mean of non-preparation participants. Moreover, there were no significant differences in these IELTS component scores between groups; in fact, preparation participants scored higher (M=6.40, SD=.90) in speaking than non-preparation group (M=6.28, SD=.80), though the difference was not found statistically significant. These findings seem to indicate that there was significant narrowing of proficiency gaps (indexed by test scores) between groups over time but it remains unclear whether this gap-narrowing is a reflection of improved proficiency or enhanced testwiseness. Meanwhile, because the pre-course measure (A-level English) differed from post-course measure (IELTS), the interpretation of the effect of test preparation is further complicated.

Further analyses using questionnaire data revealed that the narrowing of proficiency gaps (indexed by test scores) can be related to specific IELTS related learning and practices as preparation participants used IELTS preparation software more often, practiced IELTS more frequently and accessed more IELTS on-line resources (Gan, 2009) than non-preparation participants. However, it is also probable preparation participants indeed improved their proficiency over the years and thus the pre-course differences ceased to exist. Nevertheless, Gan's research failed to provide an answer as to how score gains were achieved, meaning that the effects of test preparation are still unclear.

Also noteworthy is that Gan's study used self-reported data which could render the results less reliable. Although questionnaire was frequently used in social science studies as a valuable method, the accuracy and reliability of self-report might also limit the generalizability of research findings. Apart from this, the observatory nature of Gan's study meant that there were many uncontrollable variables that might have intervened the research outcomes; for example, as the IELTS preparation took up only 10-20 hours of 3-year full-time undergraduate education, to conclude that IELTS preparation was capable of producing positive effect on test performance might be too much of a stretch given how short the preparation period was. One could also argue that the narrowing of between-group differences could also be attributed to students' confidence in themselves and overall matureness in test taking, both of which may be effects of receiving higher education. As stated above, from a methodological perspective, the use of A-level as a pre-course measure and IELTS as a post-course measure could be problematic. Although both tests were widely acknowledged as standardised tests of proficiency, because of the difference in scoring system (i.e. IELTS having much less band scores/grades than A-level), it could be possible that the differences in A-level does not translate to significant differences in IELTS, which then cancels out the later observed gap-narrowing. It should also be highlighted that the context of Gan's study, Hong Kong, a previous colony of the UK, uses English and China both as its official languages. This context complicates the interpretation and limits the generalizability of Gan's findings as students in HK are likely to have more exposure to English compared to other students also from EFL contexts. Hence, it is reasonable to conclude that although Gan's provided valuable evidence as to the effect of IELTS preparation, more controlled, if possible, experimental studies in typical EFL contexts are still called for.

Gan's research findings posed the probability that test preparation might be able to improve scores as well as the ability the test assesses. If this probability is confirmed, then one can map this practice to type B preparation practice according to Messick's categorisation (1981), which does not pose detrimental threats to IELTS's validity. However, evidence from Xie's research (2013) on College English Test Band 4 (CET4)[11] test preparation showed that type C test preparation prevailed among professional test preparation organisations and had negative implication on test validity.

Using a pre-test/intervention/post-test design, participant's pre/post intervention CET level were assessed through two sets of past CET4 papers (N pre-test=847, N post-test=833); in addition, participants reported their use of strategies in preparation for the CET4 through a questionnaire (Xie, 2013). Six key subscales of strategies were looked at: test taking skills, test-oriented practices, repetitive drilling, narrow memorisation, social strategies to seek support from teachers and peers and finally, general learning strategies that enhance a broader range of language skills. These strategies had different implications on test validity. While general learning strategies fall under Messick's type B test preparation practices, posing no threats to test validity, test-oriented practices, repetitive drilling and narrow memorisation can be seen as typical samples of type C practices as these strategies may inflate scores without improving proficiency. Use of social strategies, however, can be seen as beneficial to test validity as it helps reduce test anxiety and boost test-taking motivation, thus reducing the impact of construct-irrelevant variances on test taking.

Analyses of strategies used in preparation revealed that test taking strategies were used with the highest frequency, followed by test preparation practices, drilling and memorisation. This distribution of strategy use, coupled with the fact that post-course CET scores were significantly higher than pre-course scores (overall and by skill), infers that the more test taking strategies candidates used, the more test practice, drill and memorisation were practiced, the more likely they were to achieve significant score gains.

[11] a standardized Academic English language proficiency test developed by the National College English Test Committee under Chinese Ministry of Education, administered twice a year to all non-English-major undergraduate students of all universities across Mainland China, each administration involving more than 2 million Chinese candidates (Zheng & Cheng, 2008)

To gain a better understanding as to which type of practices made unique contribution to score gains, structural equation modelling was used. The final model, which fitted the test perfectly, showed that test preparation practices, memorisation, drilling made significant contribution to post-course CET scores, after controlling for pre-course scores. In comparison, the effects of general language learning strategies and social strategies were not significant. To conclude, test preparation did affect test scores, and these effects primarily came from preparation practices via narrowing the curriculum, especially memorisation and drilling.

Based on these findings, there is good reason to conclude that the test preparation pattern observed in Xie's study (2013) to be negative and unintended because they are unlikely to help candidates develop their language abilities. The fact that test preparation practices, memorisation, drilling and cramming can substitute almost one third of the effects generated by candidates' prior English language ability (index by pre-course CET scores) raised concern about the extrapolation validity of CET4 scores, that is, the strength of the inference link from the tested to the untested behaviours (Kane, 1992). Given how effective test preparation practices, memorisation, drilling and cramming were shown to be, to what extent can scores gains on proficiency test be seen as indication of candidates' proficiency progress and to what extent can score gains be seen as indication of candidates' testwiseness are important questions waiting to be answered.

Among existing proficiency test preparation literature, Xie's study (2013) is the only one that explored effects of proficiency test preparation practices on test performance and test validity explicitly using a quantitative approach. Yu (2014) also made an attempt to address how Chinese EFL learners prepared tests using a qualitative approach by interviewing 15 Chinese students who had achieved a minimum of IELTS 7. Interview data revealed that all participants stated that they had a good understanding of the test itself and the task items. Also on the note of test preparation strategies, it was found that candidates frequently looked at posts written by previous IELTS takers to find tips, a means of familiarising themselves with the test. From this, one can infer that IELTS related knowledge acted as the key to success in achieving high IELTS scores. Although Yu's study offered valuable insights concerning the preparation practices Chinese EFL learners undertook in order to perform well on IELTS, the qualitative nature of this study, to certain extent, limited the study from presenting a clear picture as to how effective the aforementioned practices and strategies could be in increasing IELTS scores. In addition, whether findings based on interview data from 15 participants with high IELTS scores could be generalised to represent the wider Chinese IELTS taking population remains unclear, which calls for future study with bigger research sample.

To conclude, although research efforts have been made to probe the impact of high stake proficiency tests and the nature and the effects of dedicated test preparation, still little is known on this topic. Close examination on existing literature further reveals that most of the studies done in the field were situated in university contexts, such as in Gan's study (2009), the IELTS preparation programme was incorporated as part of students' university curriculum. By contract, test-preparation centres have not received the same amount of attention, despite the fact that they are arguably sites where the most egregious negative washback can potentially be observed. Tunks claimed that these centres tend to focus on raising test scores by developing "testwiseness' or becoming familiar with test directions, wording, format or item type, etc., and adopting strategic test-taking methods such as looking for clues, eliminating wrong choices, planning one's time, noting key words and grammatical relationships (2001). It is of importance to verify such claim using empirical evidence because it relates to the validity of the test as well as has implication for the users of the test.

To achieve more generalisability, I decided to situate such quest in the Chinese test preparation context on the following groups. Firstly, as presented in Introduction, Chinese students constitute the biggest proportion of international students enrolled in English-speaking HEIs as well as the biggest proportion of IELTS candidature globally (section 2.1.2, 2.8.1). Meanwhile, international Chinese students were reported to experience struggles with their English skills and they were constantly outperformed by other non-native English speaking international students (section 2.2.3). Furthermore, combining the high stakes IELTS holds among Chinese candidates, the gatekeeping role IELTS plays for English speaking HEI, and the popularity of test preparation in China, it seems logical to set out a research attempt to connect the dots between test preparation, score gain, language proficiency (or the lack of which), and Chinese students' academic attainment at English speaking HEIs. By doing so, this study also enriches the discussion of IELTS's construct and predictive validity and provides implication for IELTS users e.g. English speaking HEIs.

2.7 Repeated test taking and test performance

In addition to participating in dedicated test preparation, IELTS and TOEFL candidates noted that they often needed more than one attempt to achieve the required scores (e.g. Ma, 2014). According to Zhang (2008), approximately 250,000 candidates took TOEFL between January and August 2017, among which about 10% repeated TOEFL at least one more time. Some 12000 candidates resat TOEFL within 30 days of their first attempt and a small percentage resat TOEFL three times or more. This repetitive test taking is not a new phenomenon； Wilson reported some 28% of all first-time TOEFL candidates between July 1977 and June 1978 had sit TOEFL twice or more. Moreover, between 40 and 50% of candidates from South East Asia regions were TOEFL repeaters (e.g. Taiwan, Hong Kong, Korea, Thailand and Japan). Although statistics on IELTS repeaters are not available, a survey conducted by the test preparation giant in China, the New Oriental School, revealed that the majority of Chinese IELTS candidates needed around 3 attempts to achieve their desired scores (Li, 2013).

There are many reasons that could have contributed to this repeated test taking behaviour. Firstly, both IELTS and TOEFL scores are only valid for 2 years, meaning that candidates have no choice but to resit IELTS/TOEFL once the score expires (IELTS Homepage, 2018; ETS Homepage, 2018). Secondly, as discussed in previous chapter, neither IELTS nor TOEFL allow score merging. For example, candidate A achieved an IELTS overall 7 on his/her first attempt with 5.5 in Writing, an IELTS overall 6.5 on his/her second attempt with 6 in Writing, if the accepting HEI specifies that applicant should have an overall 7 with 6 and above on each componential test, this candidate has to resit IELTS again. More importantly, some candidates hold the belief that tests such as IELTS and TOEFL has a *primary task bank* with relatively few new tasks added on every year. Thus, they assume, through multiple attempts, they might encounter tasks that they have memorised the answers to and achieve a higher score. This is supported by the popular use of IELTS and TOEFL *Ji Jing* (which means test taking experience in Chinese), a widely used task bank built upon Chinese IELTS and TOEFL candidate's timely refection of the tasks they just encountered (Ma, 2014).

To the best of my knowledge, there is no dedicated study on repeated IELTS taking to date, a gap this present study aims to fulfil. In comparison, more efforts have been made on the TOEFL side, although research literature is by no means substantial. In 1987, Wilson analysed the pattern and score change among TOEFL repeaters and reported with more time and 'effort' (indexed by the number of attempts made, which according to Wilson (1987), "may be thought as reflecting" p.7), greater score gains could be achieved. Moreover, mean score gains were found to increase with each additional attempt; three time TOEFL repeaters were more likely to achieve bigger score gains than two time TOEFL repeaters and so on so forth. This, to some extent, lends support to the assumption of repeated test taking could lead to better scores commonly held among Chinese candidates; nonetheless it remains unclear whether the reported score gains were a reflection of improvement of general English proficiency based on the time and effort candidate put in as claimed by Wilson, or an outcome of accumulated testwiseness, which introduces construct irrelevant variance into the validity of the test.

It was also revealed that the tendency to repeat tests was not as linear as the hypothesised those who achieved lower first-attempt TOEFL were more likely to repeat. Rather, the tendency to repeat was more closely associated with the origins of the candidates; for example, the percentages of repeaters from the far-Eastern continents were almost six times greater than that of repeaters from Indian, African and major European contingents. This repeater tendency could be attributed to the culturally different role tests play in different educational context; in other words, candidates from countries that place more emphasis on tests are more likely to repeat test taking. On the basis of this assumption, it is reasonable to infer that for Chinese candidates, the majority of who receive most of their school level education in a highly test oriented fashion, are very likely to become test repeaters. Given the large amount of TOEFL and IELTS tests taken by Chinese candidates, an investigation of the effects of repeated test taking of this population are of significant importance as it not only bridge gap in existing literature on the topic but also provides valuable insights on the validity of the tests.

2.8 Context of this present study

2.8.1 High-stake English proficiency tests and Chinese candidates

Since its introduction into China in April 1990, IELTS has become popular throughout the country (China Daily, 2001). IELTS statistics show that between 2016 and 2017, 3 million IELTS were taken globally, revealing the growing popularity and the wide acceptance of IELTS as the world's leading test of English for international higher education and migration (IELTS homepage, 2018). In 2011, over 150 million candidates around the world have taken IELTS, among which there are over 300,000 Chinese candidates (in mainland China) who received this international test in 48 test centres. Mainland China takes 50% of IELTS candidates in Asian area and applicants even have to wait in lines for a long time to register. In 2014, the population of IELTS candidates quadrupled to 600,000, accounting for more than one fifth of all candidate population for the examination worldwide (South China Morning Post, 2015).

Based on the mean overall and individual band scores achieved by candidates of the year 2107 from different places of origin data (IELTS demographic data, 2018; IELTS test taker performance, 2016), Chinese candidates on average achieved an overall 5.73, which could be rounded up to the nearest band 6 *Competent user* category, i.e. s/he has an effective command of the language despite some inaccuracies, inappropriate usage and misunderstandings. S/he can use and understand fairly complex language, particularly in familiar situations. This average overall IELTS band ranks 202th among the 230 listed places of origin, placing Chinese IELTS candidates' overall IELTS performance at the bottom 15% of the whole IELTS candidature. Moreover, almost half of the Chinese IELTS 2016 candidate population achieved an overall lower than 5; interesting, the percentage of band 7 achievers (23.81%) was very similar to that of band 4.5 achievers (23.01%) and the percentage of band 7.5 achievers (15.24%) was almost the same as that of band 4 achievers (15.15%). Regarding individual module performance, Chinese candidates scored the highest on IELTS reading (M=6.08), followed by listening (M=5.84), speaking (M=5.39) and the lowest on writing (M=5.34). Comparison between Chinese candidates' IELTS performance and that of candidates from other EFL countries that shared similar cultural and educational characteristics e.g. Korea (mean IELTS overall=6.19), Mongolia (mean IELTS overall=6.04), Viet Nam (mean IELTS overall=5.95) and Japan (mean IELTS overall=5.81) further illustrated that Chinese candidates are particularly vulnerable in terms of their English proficiency. These reported statistics were mostly in line with the language barrier encountered by international Chinese students as stated in previous sections, lending support to the assumption that language is not their strongest suit. However, there also exists the possibility that Chinese candidates' lower IELTS performance could be related to their unskillfulness at taking the IELTS test, their unfamiliarity with the IELTS format or simply, nerve. In other words, their ‘actual’ proficiency may be slightly higher than what IELTS outcome may entail; given the high stakes IELTS holds among Chinese candidates, especially for those with an agenda to study at an English-speaking HEI, it is very common for potential IELTS candidates to attend dedicated test preparation courses to ready themselves for the test, as indicated in Ma’s study (2014).

2.8.2 Test preparation for English proficiency tests in China

Test preparation is by no means new in China. The testing and examination history in China can be traced back to the imperial period nearly two thousand years ago (Cheng, 2008) and Chinese education system attaches great importance to standardised tests as a tool to determine its students’ future. In comparison, the testing of foreign languages started much later. In recent decades, as the population of Chinese students seeking overseas education in English-speaking HEI has increased significantly, test preparation for high-stakes English language tests has become a prevalent social and educational phenomenon in China (Ma, 2013) and has become a fast-growing, profitable industry (Matoush & Fu, 2012; Wang, 2007; Xu, 2007). Consider, for example, New Oriental, a leading private educational institution in test preparation in China. It is reported that 70% of Chinese students in universities in the United States and Canada were trained at this institution prior to their overseas studies (Tang, 2010). New Oriental is only one of an enormous number of test preparation centres in China.

There has been reports highlighting the unethical practices of some Chinese test preparation centres. In 2015, a group of 357 students had their Upper Level Secondary School Admission Test (SSAT) examination scores cancelled as a punishment of memorising and cramming old test questions. Interestingly, this memorisation and cramming preparatory technique did not seem to cause much concern for many Chinese candidate and they do not consider it to be unethical or cheating; it was further revealed that many Chinese language centres and other cramming centres earn their living by compiling test questions and then signing up students, who memorise the answers rather than understand them (Yan, 2015). It is worth mentioning that this cancellation of SSAT was not a singular incident; in the same year, IELTS announced to permanently withhold the score reports from about 350 Chinese candidates because “their result is not a true reflection of their English language skills” (South China Morning Post, 2015). Moreover, in 2001, China's most prestigious overseas English exam training school, occupying over half of the market share on mainland China—the New Oriental Education Group—was sued to pay 10 million yuan in compensation for copyright and trademark infringement to the Educational Testing Service and the Graduate Management Admission Council owing to the repeated illegal publication of TOEFL, Graduate Record Examination (GRE) and Graduate Management Admission Test (GMAT) exam papers as preparatory materials without appropriate authorization (Zi, 2004).

In spite of these reported risky practices of some Chinese candidates and Chinese test preparation centres are engaged in, only a handful of research studies have been done to shed some light on the nature of test preparation industry and test preparation practices for high-stake English proficiency test in China; even less literature is available on IELTS and TOEFL test preparation in China in particular, regardless of the crucial gatekeeper role these two tests play. Ma & Cheng (2015) looked at the perceived value of TOEFL iBT preparation course through interviewing 12 Chinese students enrolled in a Canadian university on the premises of achieving the required TOEFL scores. All interviewed Chinese students acknowledged their main goal and motivation was to become familiar with the content of the test and how the test is administrated through these preparatory courses; they also considered such preparatory courses "the most time-efficient method to prepare for the TOEFL" and that they "would take test preparation courses again" if they were to sit the test again (p. 69). Ma & Cheng (2015) further reported that, for some students, their "only purpose for taking test preparation courses was to familiarize themselves with the test and that they had no expectation of improving their English skills" (p. 70), consistent with findings from Ma (2014). Given this instrumental motivation, it is not surprising that most, if not all of their attention was directed to test-related instruction, rather than the improvement of overall English proficiency. Further, students claimed that when personal constraints (e.g., a narrowed attitude toward learning) were confounded with contextual constraints (e.g., a limited timeline), it was unlikely that the students would shift their orientation to the need to develop English skills. One interviewee explicitly stated, "I have a very tight schedule [for preparing for the TOEFL iBT], and I do not want to improve my English proficiency through this course; I *just* need a score [emphasis in original tone], and this is most important to me" (p.70).

Ma and Cheng's study, although very limited in sample size, showed very clearly that Chinese candidates perceived test preparation course as a means to achieve the needed scores and were less concerned whether these courses could improve their general English proficiency, which could be problematic because this may result in scoring overrepresenting the proficiency of the candidates. However, due to the qualitative nature of Ma and Cheng's study, it remains unclear whether these perceptions from candidates did correspond with the outcomes of their learning. In other words, we now know that Chinese candidates *thought* test preparation could help them achieve their desired TOEFL scores but what we do not know is the extent of such effects; could test preparation boost candidates' scores significantly? More importantly, and perhaps more of relevance to stakeholders such as accepting universities, is the question whether this help from test preparation could endanger the validity of a standardised proficiency test? If the increase in test scores does not correspond to the improvement in candidates' general proficiency. These questions are of great research importance because their answers are closely associated with the aforementioned academic performance of international Chinese students enrolled in English-speaking universities, their social and psychological adjustment issues and the validity of these widely used high-stake proficiency tests.

2.8.3 Chinese candidates' repeated test taking behaviour and relevant concerns

When reviewing the literature on Chinese students' test preparatory practices, one particular test taking behaviour caught my attention. It has been documented in previous research that many students need to sit more than once in order to achieve the scores they desire (Ma, 2014; Ma & Cheng, 2015; Zhang, 2008). For example, in Ma & Cheng's study (2015), 75% of the participants (N=12) attempted TOEFL more than once before they stopped retrying, having either achieved the needed TOEFL scores or found alternative ways to bypass this TOEFL requirement, e.g. attending pre-session courses or pathway programmes.

Although up-to-date official statistics on the number of IELTS and TOEFL repeaters could not be found, as discussed earlier, the New Oriental Education Group conducted an independent survey across 19 Chinese cities and reported on average, Chinese IELTS candidates need around 3 attempts to achieve the scores they desire. This report also revealed that 62.7% of the surveyed Chinese IELTS candidate set an overall IELTS 7 as their target score (Li, 2013). The survey did not provide detailed analyses regarding reasons why Chinese IELTS candidates repeated IELTS taking behaviour, but based on the candidates' expected IELTS scores and Chinese candidates' average IELTS performance discussed previously (see section 2.8.1), it is reasonable to infer that such behaviour is an effort to increase their scores, which begs the question of can repeated test taking assist IELTS candidates in achieving higher scores? More importantly, can repeated test taking increase candidates' IELTS score to a level beyond their actual proficiency? If so, what implication could that have on the validity and reliability of IELTS? These questions are worth exploring because they are closely associated with the aforementioned international Chinese students' academic achievement in English speaking HEIs given the role IELTS plays in the admission of international students.

2.8.4 The present study

This study was designed to address the gap in existing literature and explore the effects of Chinese IELTS candidates' test preparation and repeated test taking on their IELTS performance and general proficiency; it also aimed to re-examine the predictive and construct validity of IELTS taking into accounting of such effects. It set out to answer the following five research questions:

- To what extent does IELTS preparation course improve candidates' IELTS scores (overall and by skill)?
- To what extent does IELTS preparation course improve candidates' general proficiency?
- To what extent does repeated IELTS taking affect Chinese candidates' IELTS scores and their general proficiency?
- To what extent does IELTS predict international Chinese students' academic attainment at a UK university?
- Have test preparation and repeated test taking eroded the validity of IELTS as a predictor for academic attainment?

Answers to these questions extend the literature on the effects of test preparation and repeated test taking in the context of IELTS and fill the gaps discussed in the reviewed literature. To answer these questions, two separate quasi-experiments were designed and conducted, methodology of which are explained in the following two chapters.

Chapter 3 Quasi-experiment 1: the effects of test preparation on Chinese candidates' IELTS test results and their general proficiency

This chapter discusses the methodology and methods relevant to the first quasi-experiment of the present study. Details regarding the research design, research instruments, testing procedure, and data analyses are provided, along with information on participants and ethical considerations of experimental research.

3.1 Methodology and methods

3.1.1 Research Design

To examine the effects of IELTS preparation on both students' IELTS performance and their general proficiency, a pretest-intervention-posttest design was adopted and two groups of participants were recruited, one acting as intervention and the other acting as the control. The intervention involved here was not in fact a manipulation; rather, data were collected through observing participants who signed up for test preparation courses out of their own interests. Because of this quasi-experiment's observational nature, one could argue that intervention participants were of lower general proficiency because they chose to attend dedicated IELTS preparation course (i.e. the intervention) and their motivation to achieve score gains was likely to be higher than the control.

Both groups were similar in age, education background, language proficiency level and length of English learning so as to control the interference from these variables on the effect of test preparation, achieving higher reliability of research findings. Pretest took place before the intervention participants began their preparation course and the posttest took place shortly after their had completed their course so as to ensure the timeliness of the collected data. Testing for the control group was arranged along the similar time frame.

While effects on IELTS test performance were measured through IELTS mock tests, participants' general proficiency was measured using a battery of non-IELTS related linguistic measure, including an online standardised proficiency test (Oxford Quick Placement Test, OOPT), a vocabulary test and an accuracy/speed of sentence processing test. The last two measures were included in this quasi-experiment because test preparation on one standardised proficiency might transfer to another (e.g. preparation on IELTS might also affect performance on OOPT), which complicates the interpretation of research findings. As lexical knowledge and speed of processing are both found reliable indicators of proficiency (section 2.2.3), the inclusion of these two measures was believed to enhance reliability of the outcomes.

3.1.2 Participants

Quasi-experiment 1 involved a total of 89 Chinese native speakers who learned English as a foreign language, recruited on a voluntary basis through advertising at the training centre where this quasi-experiment took place. Among these 89 participants, 45 of them were enrolled in an IELTS preparation course (i.e. the intervention of this quasi-experiment) at a typical test preparation and training centre in Shanghai, China, thus, forming the *Intervention Group* in this quasi-experiment. The sampling strategy used to recruit intervention participants was criteria-based sampling; specific criteria such as first time attending 1-month IELTS preparation course, no previous knowledge of IELTS, no previous IELTS tests taken were used to screen potential participants accessible at the training centre where data were collected.

The average age of intervention participants was 21.76-year-old with a SD of 3.75; while the oldest participant from this group was 34-year-old, the majority of this group was aged between 18 and 24. Given this age spectrum, it is not surprising that most of the intervention participants were full-time students; 38 (84%) were last year high school students and university undergraduates studying across a variety of disciplines; 6 were university graduates who had recently completed their undergraduate studies and were in full time preparation for their future overseas study at an English speaking university; only one participant was working full time in the field of marketing while attending the IELTS preparation course. It is worth pointing out that the majority of intervention participants were studying/had studied non-English related disciplines. This could be related to their motives for attending IELTS targeted preparatory course in the sense that they might not have learned sufficient English knowledge or adequate test taking skills through their current/previous undergraduate education, leading to the need for preparatory courses.

The remainder of the sample consisted of 44 Chinese native speakers who did not sign up for any kind of English related training at the time of data collection, thus forming the control group for this quasi-experiment. The inclusion of a control group set a baseline against which intervention participants' performance on tests could be compared. Moreover, given that this quasi-experiment adopted a repeated-measures pre-test post-test design, the control group also served to cancel out the effects of repeated testing and to further illustrate the effects of the intervention under investigation. The control participants were recruited through snow-ball sampling in the sense that they were brought in this quasi-experiment through intervention participants. In other words, intervention participants were asked to invite their coursemates or friends of similar age and education history to take part in this quasi-experiment and act as control. This helped to control the effects of variables such as age, previous education and English learning history.

Taking into consideration the chosen sampling strategy, as expected, the average age of control participants was very similar to that of intervention with a mean of 22.30 (SD=4.76). The difference between group was not significant, $t(87)=.595$, $p>.05$. 39 out of 44 control participants were full-time students either at the end of their high-school education or at various stages of their university education; the remaining 4 control participants were full-time employed in non-English dominant fields such as Medicine, Nursing, Journalism and Car Engineering. It is worth-noting that although 3 (0.7%) of the full-time undergraduate control participants were enrolled at English-related field of studies (e.g. English and English Education), the majority of control participants were studying/had studied non-English related disciplines, which resembles that of the intervention group. This between-group similarity in current/previous education could be seen as an indication that English education at university level received by most participants were similar; i.e. the variable *English learning history* was controlled in this quasi-experiment.

As mentioned earlier, all participants were from Mainland China, spoke Mandarin Chinese dominantly for communications, received their education in Mandarin Chinese and they all learned English as a foreign language. None reported having previous experience with IELTS tests or IELTS test preparation courses. In terms of their outside classroom English exposure, most of them listened to English music on a daily basis, watched English television or films on a weekly basis while several reported they read English books and journals frequently. They rarely used English social network and they did not communicate with English native speakers in their social life regularly. In terms of tourism/living experience in English speaking countries, 4 intervention participants and 2 control participants reported they had spent a relatively short period of time in the UK, USA, Canada and Australia for tourism. One intervention participant and 1 control participants spent 6 and 4 months living in an English-speaking country respectively as a part of their universities' exchange student programme. In the following data and results sections, analyses done with or without these two participants did not yield different results; thus, for the sake of data completeness, all participants were included in this thesis.

3.1.3 The Intervention

The intervention involved in quasi-experiment 1 was a common IELTS preparation course offered at a typical test preparation training centre in Shanghai, China. This particular course lasted for 4 weeks; students were required to attend four-hour focused teaching sessions Monday to Friday and there were voluntary self-study sessions in the afternoon in a designated classroom. The content of this course included IELTS specific teaching and learning on all four IELTS modules, i.e. listening, reading, writing and speaking and the practices in class were largely in line with the typical preparation course afore-presented in 2.6.1. As there is no literature documenting the specific pedagogy used for these preparation courses in the Chinese IELTS context, here, I drew reference from my previous IELTS teaching experience when I was employed as a full-time IELTS instructor at this particular training centre. Typically, for a four-week course, each module had approximately 18 hours of instructions and was taught by different IELTS instructors whose expertise and teaching experience varied, i.e. some teachers are specialised in IELTS listening, some in IELTS reading. In each module, the first 2 hours (session 1) were often spent on test familiarisation with the teacher demonstrating all potential task types that the candidates were likely to encounter in the test and explaining the details that the candidates need to pay attention to when answering those tasks. For example, in IELTS listening, the teacher would first showcase the various task types involved in the test (e.g. short answers, multiple choice, labelling a diagram/plan/map) and point out that in some of these tasks candidates were only allowed to answer the questions with ONE word only. Following test familiarisation, the remaining 16 hours were often spent on testwiseness development. Take the True/False/Not Given judgement task in IELTS reading for example; the teacher would instruct the candidates to first read through the statements and then judge statements that contain words such as "only" "all" as False because these statements are too “absolute” to be True. They may also teach the candidates to guess the answer by analysing the proportion of answers already produced; for instance, if there were three True/False/Not Given items, and if the candidates had already given True for item 1, and False for item 2, it was very likely that item 3 would be a Not Given. For IELTS speaking and writing, sessions following test familiarisation involve mostly of preparing past topics and memorising pre-written chunks. For example, the teacher would categorise past speaking topics into “person” “event” “object” and “experience” and ask candidates to prepare a 1-2 minute speech relevant each one of these topics. Then, the teacher would show the candidates how to adjust their prepared speeches on one topic to another. For instance, the candidate had prepared a 2 minute speech about his/her mother and the teacher would show the candidate that the speech on mother (person) could also be used for a memorable photograph (object) as the candidate could talk about a photograph of his/her mother. Similar pedagogies could be observed in the teaching of IELTS writing. Because of the large amount of time spent on test taking skills instructions and testwiseness development via narrow learning, primarily using past test papers, one could say that a substantial proportion of this pedagogy fits into Messick’s Type C test preparation previously discussed in section 2.6.

As data were collected from the intervention participants enrolled in the centre at different time points, the courses under investigation were taught by different instructors with different approaches and teaching styles. Nonetheless, the overall teaching and learning objectives, the course lengths as well as the main materials of these courses remained the consistent. Thus, it is reasonable to conclude that all interventions involved in this quasi-experiment were controlled and of similar nature.

3.1.4 Research Instruments

As the main goal of quasi-experiment 1 was to examine the extent to which dedicated IELTS preparation courses could improve Chinese candidates’ IELTS scores and their general proficiency, a battery of proficiency measures was used to gauge the pre-course and post-course IELTS levels and general proficiency levels of participants involved in the quasi-experiment. In addition, a questionnaire was used to collect participants’ demographic data.

Questionnaire

Questionnaire survey is amongst the most widely used data collection instruments in educational research. It is easy to use and generates a considerable amount of data within a short period of time (Dörnyei & Taguchi, 2009). This questionnaire (see Appendix 3) included questions concerning participants' education background, language background, English learning activities, familiarity with IELTS test and experience with test preparation. Participants were also asked to evaluate their own level of English proficiency using IELTS band system using a self-rating scale. This question facilitated better understanding of how participants viewed their own proficiency.

IELTS mock test[12]

To examine the extent to which IELTS preparation course could improve candidates' test performance, IELTS mock tests were conducted. The test paper used for mock tests were taken from Cambridge IELTS book series (book 1 and 2) (Cambridge ESOL, 2000; Jakeman & McDowell, 1996). Each book in this series consists of authentic past IELTS examination papers; thus, when used in a controlled environment, these test papers are believed to provide accurate indication of candidates' IELTS performance in real IELTS tests on the premises that they had not received preparatory training prior to taking the test. Book 1 and 2 were chose for the reason that the intervention course provided book 3-10 as the learning material; hence any test paper from book 3-10 would be familiar to the intervention participants, undermining the reliability of the mock test scores.

Same as a real IELTS test, the mock test in this quasi-experiment included a listening module, a reading module, a writing module and a speaking module, which in total lasted for approximately 4 hours. For listening, there were four sections, each with 10 questions. The topics of section 1 and 2 were situated in social contexts (e.g. opening an account at a local bank) while section 3 and 4 were situated in academic context (e.g. listening to an academic lecture). In terms of speech format, section 1 and 3 were dialogues between multiple speakers while section 2 and 4 were monologues. Candidates encountered a variety of task types in IELTS listening such as multiple

[12] Experiments involved in this thesis only looked at the IELTS Academic version; hence hereafter in this thesis, IELTS refers to IELTS Academic.

choice, short-answer questions, sentence/notes/form/table/summary/flow-chart completion, labelling a diagram/plan/map, classification and matching. In IELTS listening, recordings were only played once, as in the real IELTS test; thus, candidates needed to answer the questions while listening, which could have considerably increased the difficulty of the test. At the end of the listening module, candidates were allowed 10 minutes to write their answers on a given answer sheet. On average, listening module lasted for about 40 minutes in total (IELTS, 2018).

The listening module was followed by the one-hour reading module, which consisted of 3 passages and 40 questions. These passages were taken from magazines, journals, books and newspapers on topics of general interest and they contained logical argument. Similar to listening, task types candidates encountered in reading vary; on top of all above-presented task types, there were choosing suitable paragraph headings from a list, identification of writer's views-*yes, no, not given* or identification of information given in the passage-*true, false, not given* (IELTS, 2018).

The last module of IELTS written test was writing, which consisted of two tasks (20 minutes on task 1, 40 minutes on task 2). For writing task 1, candidates were required to produce at least 150 words using information extracted from a diagram or some representation of data (e.g. table, figure or chart), which resembled the *data* or *result* chapter in a dissertation to some extent . Their writing was assessed on their "ability to organise, present and possibly compare data, describe the stages of a process/an object or an event or explain how something works" (IELTS, 2018). In Task 2 candidates were asked to compare and contrast evidence to justify a given argument/statement, to discuss and present solutions to a particular problem, and to evaluate and challenge ideas, evidence and arguments (IELTS, 2018).

IELTS speaking takes approximately 10-15 minutes and is conducted on a 1-on-1 basis between the candidate and a well-trained IELTS examiner in real tests. In this mock test, the examiner role was played by 2 experienced IELTS speaking instructors; the reliability of their rating is discussed later in this chapter. Same to the real test, mock IELTS speaking consisted of three parts; part 1 was the warm up part where candidates introduced themselves and answered questions of general topics such as hometown, family/friends, study/work. Part 2 focused on one particular topic presented to candidates on a prompt card of the examiner's choice; candidates were given a minute for preparation and they are required to talk on this given topic on their own for 1-2 minutes then the examiner asks a series of follow-up questions to end this part. Part 3 was a 4-to-5-minute discussion where candidates and the examiner discussed abstract issues that were thematically related to the topic in part 2 (IELTS, 2018).

General proficiency

As this quasi-experiment intended to determine the extent to which IELTS preparation courses could improve candidates' general proficiency, a standardised measure of general proficiency is needed for such purpose. Here, the Oxford Online Placement Test (OOPT) was used.

Unlike IELTS which is a paper-and-pen test, OOPT is a computerised adaptive test. This means that it adapts to the ability level of each candidate and only presents the candidate with questions that are at his or her ability level. It does this by selecting each item for a candidate, based on how they answered the previous question. Getting a question correct means that candidate's next question will be a more difficult one. Getting a question wrong will result in the system selecting an easier question. In this way, the test homes in on each candidate's ability and level and can thus builds tests for each candidate to quickly identify their CEFR level (Purpura, 2009).

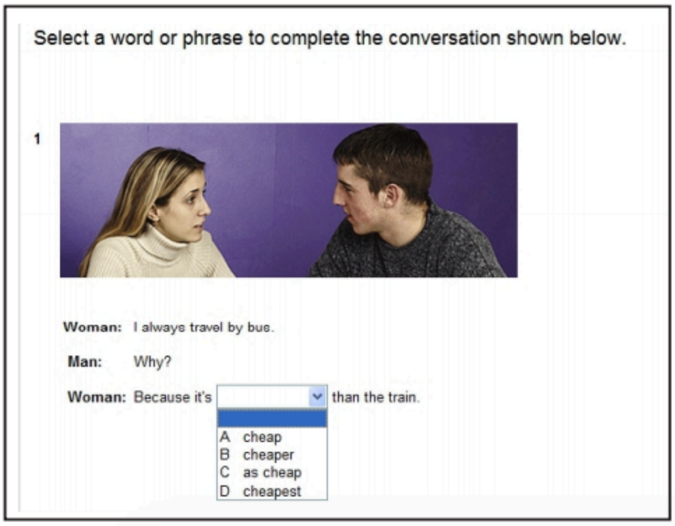
At the very beginning, OOPT asks candidates to choose their starting level from *lower level* for beginner and elementary learners, *mid-level* for intermediate learners, and *upper level* for upper-intermediate and advanced learners. The first test item is based on the chosen starting point and following test items are selected by the system depending on how successful the candidate is in terms of answering questions correctly. Because of its adaptive nature, no two candidates see the exact same series of test items during the test, which helps prevent cheating. The length of the test a candidate needs and the amount of questions one may encounter depends on how successful the candidate has been. In other words, if a candidate continues to answer questions correctly, it would take him/her much less time to finish the test than others who keep making mistakes.

OOPT is designed to measure candidates' grammatical and pragmatic knowledge of a second/foreign language as well as the ability to use such knowledge to communicate a range of meanings while listening and reading. It adopts Purpura's (2004) model of language knowledge which specifies two components of grammatical knowledge (i.e., grammatical form and semantic meaning) that jointly serve as the basis for conveying a range of implied meanings (e.g., contextual, sociocultural, etc.) in language use contexts as the test's theoretical framework. This model, underpinned by the communicative language competence theory, is believed to share some similarities with the model upon which IELTS is built, i.e. Hyme's communicative competence framework (1972) as discussed in the previous chapter (2.3-2.4). To recap, IELTS also, is "designed to assess the English language ability of people whose first language is not English and who need to study, work or live where English is used as the language of communication" (IELTS, 2018). This similarity in design and the commonality in underpinning theoretical frameworks indicate that IELTS and OOPT are likely to be assessing the same construct, i.e. candidates' general language proficiency; the key construct this quasi-experiment set out to assess.

Different from IELTS's four language modules, OOPT only consists of two sections: Use of English and Listening. The exclusion of productive skill modules (i.e. speaking and writing) in OOPT could influence the correlation between OOPT overall and IELTS overall scores reported in section 3.1.5.

OOPT Use of English, is mainly geared towards how much learners know about grammatical forms, and the meanings at word, phrasal or sentence level behind these forms (Purpura, 2009). Candidates saw four types of language knowledge tasks in this section. Type 1 measures the candidate' knowledge of grammar at their appropriate level of proficiency. As shown in Figure 3.1, in this task, candidates are asked to read a dialogue with a gap and then complete the dialogue by selecting one of four option choices. Type 2 (example shown in Figure 3.2) is designed to measure candidates' ability to "use grammatical forms in order to understand the meanings communicated by speakers in a short, minimumly-contextualized exchange" (Purpura, 2009, p. 17). The meanings vary on a cline from very explicit, where the meaning can be determined from the words in the sentence, to very implicit (also referred to as "implied" or "pragmatic"), where the meaning of the utterance can only be determined from the context. Being able to understand not only the words of an utterance, but also their collective meaning(s) as intended by the speaker in context, is a critical feature of being able to communicate in a language (Purpura, 2009) and an important component of proficiency under the guidance of communicative competence. Type 3 (example shown in Figure 3.3) assesses whether candidates can understand a long passage with gaps and whether they have sufficient knowledge of grammar and vocabulary to correctly complete these gaps. In other words, candidates' knowledge of both grammatical form and meaning is assessed. Task 4 looks at the students' knowledge of the pragmatic (i.e., implied) meanings encoded in situated interactions (Purpura, 2009).

Figure 3.1.1 Example for Task 1 in Use of English, OOPT: Testing knowledge of grammatical forms at the A2 level by Purpura (2009, p. 17)



Select a word or phrase to complete the conversation shown below.

1

Woman: I always travel by bus.

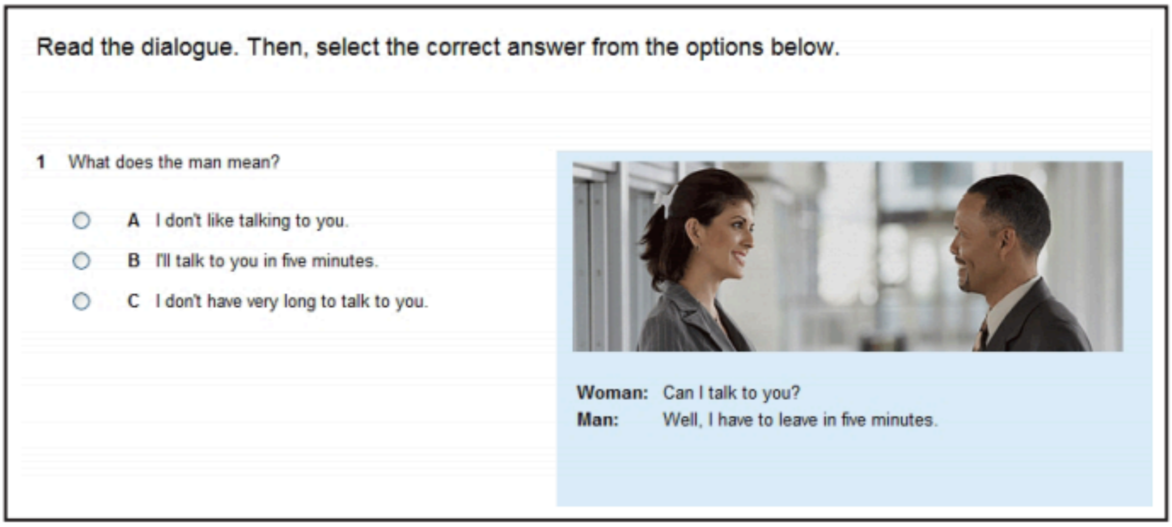
Man: Why?

Woman: Because it's ▾ than the train.

A cheap
B cheaper
C as cheap
D cheapest

This item tests the learners' knowledge of the comparative form.

Figure 3.1.2 Example for Task 2 in Use of English, OOPT: Testing knowledge of implied meaning at the A2 level by Purpura (2009, p. 18)



Read the dialogue. Then, select the correct answer from the options below.

1 What does the man mean?

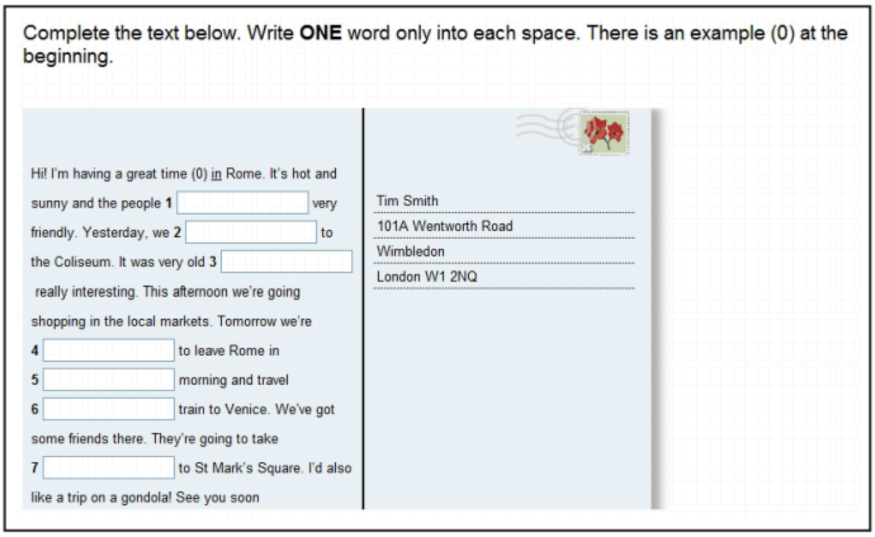
○ **A** I don't like talking to you.

○ **B** I'll talk to you in five minutes.

○ **C** I don't have very long to talk to you.

Woman: Can I talk to you?
Man: Well, I have to leave in five minutes.

Figure 3.1.3 Example for Task 3 in Use of English, OOPT: Testing knowledge of grammatical form and meaning across a passage at the A2 level by Purpura (2009, p. 19)



Complete the text below. Write **ONE** word only into each space. There is an example (0) at the beginning.

Hi! I'm having a great time (0) in Rome. It's hot and
sunny and the people **1** ☐ very
friendly. Yesterday, we **2** ☐ to
the Coliseum. It was very old **3** ☐
really interesting. This afternoon we're going
shopping in the local markets. Tomorrow we're
4 ☐ to leave Rome in
5 ☐ morning and travel
6 ☐ train to Venice. We've got
some friends there. They're going to take
7 ☐ to St Mark's Square. I'd also
like a trip on a gondola! See you soon

Tim Smith
101A Wentworth Road
Wimbledon
London W1 2NQ

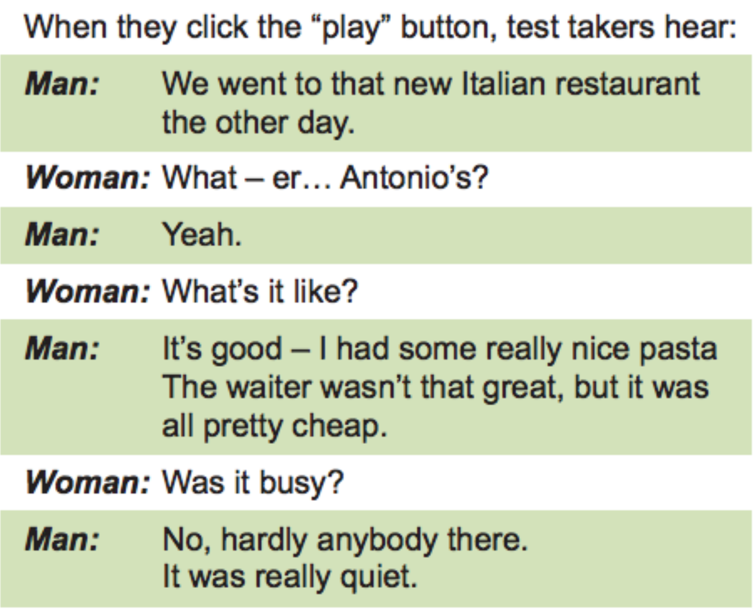
The correct answer in gap 1 is "are/seem/appear." In this case the learners need to determine what word would meaningfully complete the idea being expressed by the writer. The learners must also know that the verb should be in the present tense and given the subject "people", the verb would require a plural form. Given these two dimensions, we say this item measures both form and meaning.

From task descriptions and illustrations of OOPT Use of English, it is easy to note that they are very different from that in IELTS. IELTS does not have an explicit module designed to assess how competent candidates are at extracting or inferring meanings at a pragmatic level, although in IELTS Reading module, candidates may encounter a task where they would need to pick a title for a passage that they just read. However, the focus of this IELTS task is mostly on the literal meaning of the passage, not pragmatic. Given this differences in task types and task focus, it is reasonable to expect that the correlation between IELTS and OOPT may not be perfect, even although both tests are designed to measure the same construct and guided by the same theoretical framework, i.e. communicative competence.

The second section of the OOPT presents candidates with different types of listening passages from which they are expected to identify the literal, intended, and implied meanings being communicated. In other words, candidates need to understand what is said (literal meaning) in the passage, what is understood "between the lines" (intended meaning), and what is communicated "beyond the lines", drawing on the individual, social, cultural, affective, or attitudinal meanings of the situation (Purpura, 2009, p. 20).

OOPT Listening includes three tasks; the first presents candidates with a number of short dialogues, each followed by a single four-option multiple-choice question. The second task type presents candidates with a longer dialogue; the third with a monologue; an example is provided in Figure 3.1.4. After candidates listen, they are asked to answer one or two multiple-choice questions. Candidates are given approximately 15 listening questions, depending on their level and they are allowed listen to the recording twice.

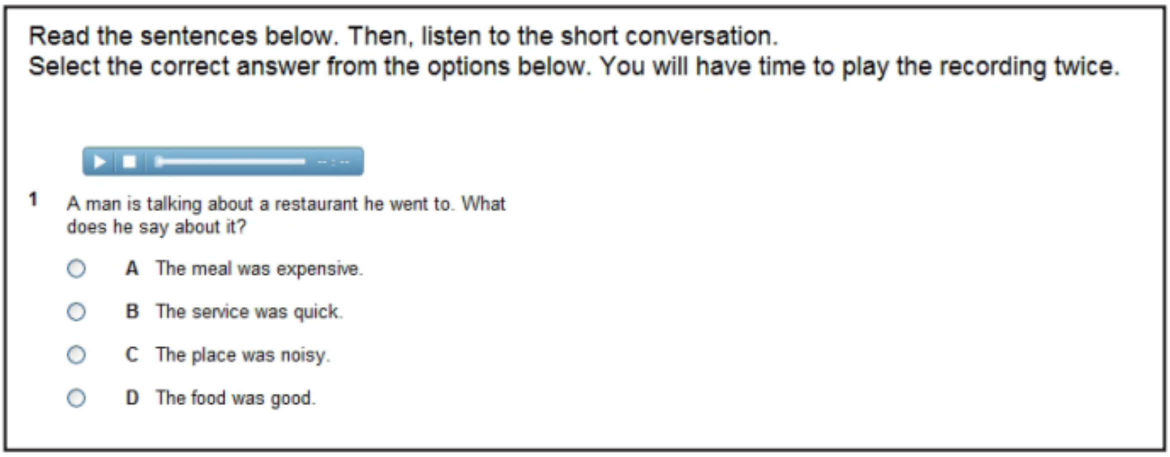
Figure 3.1.4 Example of OOPT listening tasks: Understanding literal meaning at B1 level by Purpura (2009, p. 20)



When they click the "play" button, test takers hear:

Man:	We went to that new Italian restaurant the other day.
Woman:	What – er… Antonio's?
Man:	Yeah.
Woman:	What's it like?
Man:	It's good – I had some really nice pasta The waiter wasn't that great, but it was all pretty cheap.
Woman:	Was it busy?
Man:	No, hardly anybody there. It was really quiet.

Candidates see the question below



Read the sentences below. Then, listen to the short conversation.
Select the correct answer from the options below. You will have time to play the recording twice.

1 A man is talking about a restaurant he went to. What does he say about it?

○ **A** The meal was expensive.

○ **B** The service was quick.

○ **C** The place was noisy.

○ **D** The food was good.

Although listening is assessed explicitly as an individual skill in both IELTS and OOPT, it is clear that while IELTS focuses mainly on the literal understanding of the recording and candidates' grammatical and lexical knowledge through tasks such as filling the blanks, answering short questions, OOPT integrates the assessment of both grammatical and pragmatic knowledge via tasks that require candidate to infer the implied meanings. This difference should be bared in mind when interpreting the correlation between IELTS listening scores and OOPT listening scores presented later in section 3.1.5.

Lexical knowledge

A measure of lexical knowledge was included in this quasi-experiment for the following two reasons: firstly, as reviewed in chapter 2 (section 2.3), lexical knowledge is found as a good indicator of one's general language proficiency, thus providing answers to the second research question (i.e. to what extent does test preparation improve candidates' general proficiency?). Secondly, given that both IELTS and OOPT are standardised proficiency tests aimed to assess very similar construct, it is reasonable to assume that testwiseness gained through the IELTS intervention may affect performance another standardised test, i.e OOPT. In this sense, only using IELTS and OOPT may not provide sufficient answers for the second research question; thus, two more measures of proficiency: lexical knowledge and sentence processing accuracy & speed were incorporated in this quasi-experiment.

Lexical knowledge was measured through a "lexical decision test" called Spot-the-word (Baddeley, Emslie & Nimmo-Smith, 1993, p. 58) and both versions of Spot-the-word (A and B) were included in this quasi-experiment. In each version, there are 60 pairs of words; each pair contains a genuine word with real meanings and one nonsense word that might look and sound like a real word but bears no real meanings. The non-words were invented so as to be approximately similar in length to the real words, and to follow English orthography so as to be readily pronounceable. For example, '*puma*' and '*laptess*'. These words are selected from a large pool of words with various familiarities ranging from every day common words (e.g. *kitchen, sofa*) through to less common ones (e.g. *levity, cuticle*) and very rare ones (e.g. *shako, xylophone*). Success of this test relies on a combination of different factors such as "lexical recognition, physical appearance, semantic meaning, and feeling of familiarity of a word", which provides a comprehensive evaluation of one's lexical ability (Crowell, Vanderploeg, Small, Graves, & Mortimer, 2002, p. 124).

Before the test began, candidates were given the following instructions:

> "Each of the pairs of items below contains one real word and one nonsense word, invented so as to look like a word but having no meaning. Please tick the

> item in each pair that you think is the real word. Some will be common words, most will be uncommon and some very rarely used. If you are unsure, guess, you will probably be right more often than you think. Before you begin the main test try the following six examples, then wait for the instruction to start." (Baddeley, Emslie & Nimmo-Smith, 1993, p. 57)

Although in the IELTS preparation courses under investigation there was no dedicated vocabulary session, it is reasonable to assume that vocabulary was embedded in preparatory teaching and learning contents given its acknowledged importance. Thus, it was hypothesized that the intervention participants' performance on Spot-the-word test after the IELTS preparation course could shed some light on whether test preparation could improve general proficiency indexed via lexical knowledge. Scoring of this test is provided in section 3.1.8.

Sentence processing: accuracy and speed

A comprehension accuracy and speed test was included in this quasi-experiment on similar grounds as Spot-the-word. As discussed earlier in Chapter 2, the ability to read and to comprehend accurately and efficiently has been found a good indicator of one's general proficiency; thus, by including accuracy of sentence processing and speed in this quasi-experiment, more insights could be gained as to whether IELTS preparation courses could improve candidates' general proficiency.

This quasi-experiment adopted the Speed-of-comprehension test, part of the Speed and Capacity of Language-Processing Test (Baddeley, Emslie & Nimmo Smith, 1992), to measure candidates' ability to read through English sentences quickly and accurately. This test consists of 100 sentences; half of these sentences are true and half of these sentences are false. True sentences are all obviously true, e.g. *'dogs have four legs'* or *'birds can fly'* and false sentences are obviously false e.g. *'dogs can fly'* or *'birds have four legs'*. False sentences are made by combining two true sentences. All sentences are made up using knowledge that is likely to be accessible to participants and there are no tricky questions (Baddeley, Emslie & Nimmo-Smith, 1992). Same as Spot-the-word test, before the test starts explicit instructions were given and 6 practise items were provided to make sure candidates understood the instructions clearly. They were told to verify these sentences as quickly as possible using common sense, putting a tick on sentences which they thought were true sentences and a cross on those they thought were false sentences. Scoring of this test is provided in section 3.1.8.

3.1.5 Reliability

Instrument reliability was evaluated through test-retest correlation analyses among all tests included in this quasi-experiment and through calculating Cronbach's Alpha for Spot-the-word A/B and Speed-of-comprehension A/B. Commonly considered to be a measure of item homogeneity, Cronbach's alpha ranges from 0 to 1.00 (1.00 indicating high consistency). Professionally developed high-stakes standardized tests should have internal consistency coefficients of at least .90 while lower-stakes standardized tests should have internal consistencies of at least .80 or .85 (Wells & Wollack, 2003). Cronbach's alpha provides a measure of the extent to which the items on the Spot-the-word-test provided consistent evaluation of participants' breath of vocabulary knowledge and their general English proficiency.

Correlation coefficients, summarised in Table 3.1.1, showed that at performance on IELTS correlated significantly with performance on the other three linguistic measures although the strength of correlation varied. As expected, performance on IELTS was more strongly correlated with that on OOPT given both tests were aimed at measuring similar constructs. The correlation between IELTS and Speed-of-comprehension was slightly weaker and that between IELTS and Spot-the-word was the weakest. This difference in correlation strength suggests that spot-the-word and speed-of-comprehension tests were indeed measuring more specific sub-constructs of foreign language proficiency, which were in line with the reasons why these two tests were involved in the first place.

Table 3.1.1 Correlation coefficients calculated using all participants' T1 performance on all linguistic measures used in quasi-experiment 1 (N=89)

	IELTS overall T1	OOPT overall T1	Spot-the-word T1	Speed-of-comprehension accuracy T1
OOPT overall T1	.467**	-		
Spot-the-word T1	.274**	.105	-	
Speed-of-comprehension accuracy T1	.374**	.246*	.109	-

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

Correlation analyses were also conducted for individual IELTS modules and OOPT sections at T1 and T2, using only the control participants' data, results of which were summarised in Table 3.1.2 and 3.1.3. As shown, IELTS and OOPT used in this quasi-experiment had acceptable test/retest reliability.

Table 3.1.2 Test-retest reliability calculated using participants' T1 and T2 performance on all IELTS modules (N=44)

	Listening T1	Reading T1	Writing T1	Speaking T1
Listening T2	.951***			
Reading T2		.902***		
Writing T2			.918***	
Speaking T2				.931***

****p*<.001

Table 3.1.3 Test-retest reliability calculated using participants' T1 and T2 performance on both OOPT sections (N=44)

	Use of English T1	Listening
Use of English T2	.504**	
Listening		.679**

***p*<.01

Table 3.1.4 Results from reliability analysis for both versions of Spot-the-word and Speed-of-comprehension at T1 and T2 (N=89)

Cronbach's α	Spot-the-word		Speed-of-comprehension	
	T1	T2	T1	T2
Version A	.715	.717	.830	.864
Version B	.735	.730	.783	.812

As two versions of Spot-the-word and Speed-of-comprehension were used at Time 1 (T1) and Time 2 (T2), separate internal consistency analyses were conducted and results summarised in Table 3.1.4. All participants' Spot-the-word-test and Speed-of-comprehension answers at both times were entered into SPSS item by item, using 1 indicating that their answer was correct, 0 incorrect, and 0.5 if the item was left unanswered (reasons for this scoring scheme are provided in 3.1.8). Reliability analyses showed a relatively high degree of test internal consistency for the both versions of Spot-the-word test and Speed-of-comprehension at both times, with a mean alpha of .724.

3.1.6 Ethics

As this quasi-experiment involves human participants, ethics approval was sought after. Prior to the commencement of data collection, this quasi-experiment gained ethic approval from Department of Education, University of York. Participants from both groups were given a letter of informed consent, detailing the expectation and requirements of participating in this quasi-experiment. Participants were also made aware of the rewards. For the intervention participants, they were entered into a prize draw, and the winner was entitled to a free IELTS test. For the control, every participant would receive a small payment upon the completion of second time testing. Following explanation of the content, participants were asked to sign the copy of consent which indicated their willingness to participate.

They were assured anonymity as their real names were replaced by a participant ID number and they had the right to withdraw at any point of the data collection. The intervention participants were also ensured that their participants in this quasi-experiment would not in any means affect their teaching and learning during the preparation course. The collected data would only be used for academic purposes and in future publication, only non-identifiable aggregated data would be presented.

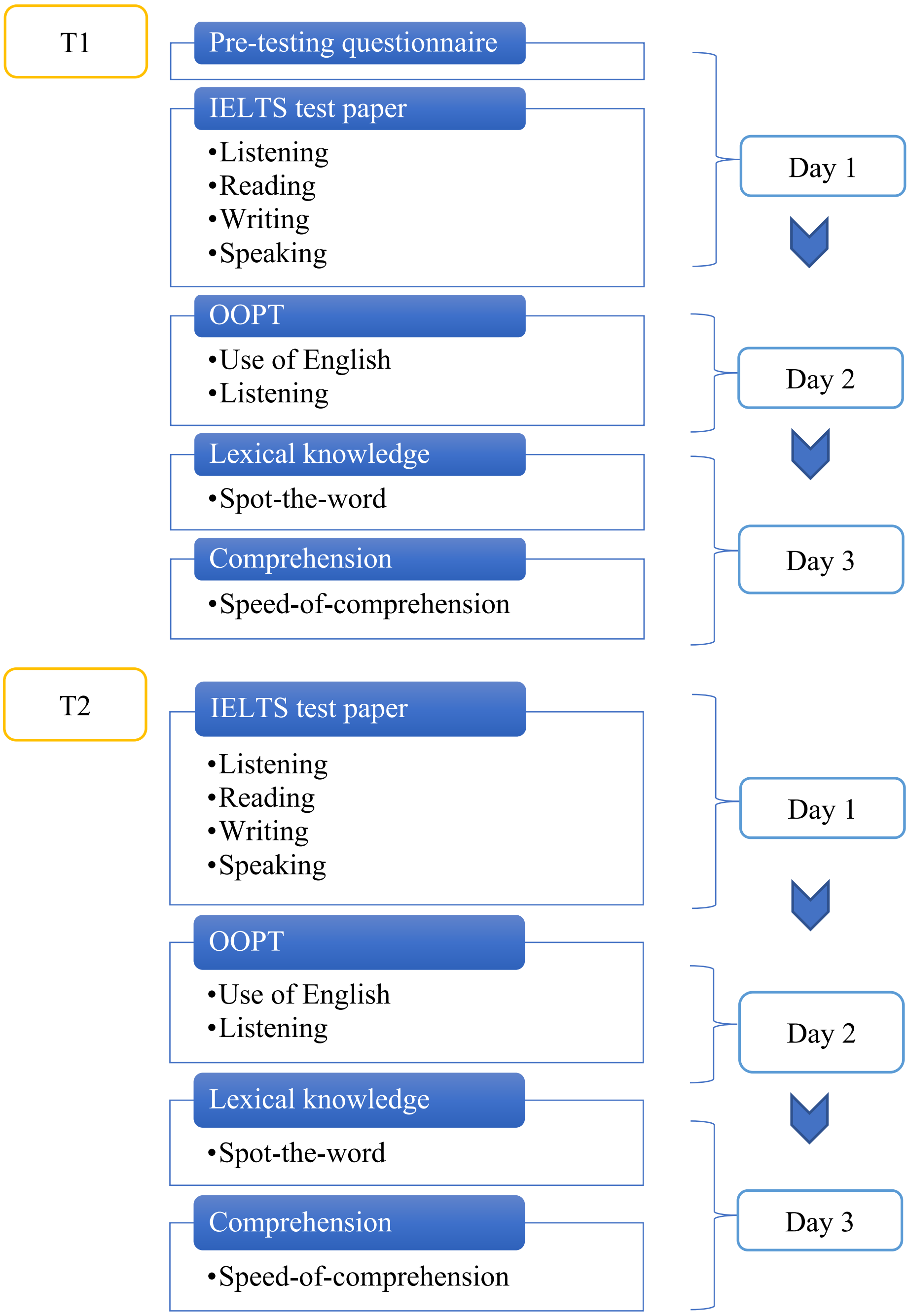
3.1.7 Procedure

The following steps were taken to recruit and test participants: potential intervention participants who met the criteria were contacted via email and invited to partake in this quasi-experiment. Control participants, on the other hand, were recruited through the network of the intervention participants, i.e. snowball sampling. Once willing participants who met the desired criteria showed their interests in participation, they were offered the chance to select the testing slots according to their availability; at this point, I briefly explained the general aim and the requirement of this quasi-experiment and answered relevant questions. Participants were presented with the written consent form and was informed that by signing this form, they were willing to take part in this quasi-experiment.

Testing at time1 was completed over several days; on day one, participants answered the pre-course questionnaire, and then sat IELTS listening, reading and writing modules in a controlled classroom setting as a group. In the following afternoon or the next morning, they sat the IELTS speaking module on a 1-on-1 face-to-face basis with an experienced IELTS speaking instructor; the speaking module of IELTS was recorded for later scoring, which is explained in the following section. The following day, OOPT was administered, also in a controlled classroom setting as a group test. On day three, participants took the Spot-the-word test and Speed-of-comprehension test on an individual basis. 4 different versions of IELTS test papers and both versions of Spot-the-word and Speed-of-comprehension were used; all test papers were arranged in a counterbalanced order so as to balance the difficulty and familiarity level of test contents. Testing was spread out over several days to avoid effects of over-testing and fatigue and to improve the reliability of data.

Testing at T2 followed the same procedure as testing T1, excluding the background questionnaire. To better illustrate the testing procedure at both T1 and T2, the following flow chart is provided:

Figure 3.1.5 Testing procedure for participant ID001 of quasi-experiment 1



3.1.8 Scoring

IELTS

There are two ways to calculate an overall IELTS score depending on whether one chooses to apply the *rounding convention*. Normally an overall IELTS band score is generated by equally averaging all four-module scores then reported to the nearest whole or half band. To avoid confusion, the following rounding convention applies: if the average across the four skills ends in .25, it is rounded up to the next half band, and if it ends in .75, it is rounded up to the next whole band. For example, a candidate achieving 6.5 for Listening, 6.5 for Reading, 5.0 for Writing and 7.0 for Speaking will be awarded an overall score of 6.5 (25 ÷ 4 = 6.25 = Band 6.5) while a candidate achieving 6.5 for Listening, 6.5 for Reading, 5.5 for Writing and 6.0 for Speaking will be awarded band 6 (24.5 ÷ 4 = 6.125 = Band 6). When a candidate receives his/her IELTS report, the score shown on that report is the one with rounding convention applied. However, since all subtests scores are also provided in the report, an actual overall score without rounding convention can also be calculated. In this quasi-experiment, as analyses using both the actual overall and the rounding conversion applied overall did not lead to difference in results, in the next section, I chose to report all IELTS overall scores with the rounding convention to resemble the reporting of real IELTS tests.

Participants' answers on IELTS listening and reading T1 and T2 were checked with the Answer Keys provided with the Cambridge IELTS book and the scores were generated based on how many items the candidate answered correctly using the guidance provided by IELTS assessment criteria (2018). Participants' performance on writing task 1 and 2 was rated by two experienced IELTS writing instructors using the marking procedure and criteria provided by IELTS[13]. According to IELTS assessment criteria (2018), examiners award a band score for each of the four criteria: Task Achievement (for Task 1), Task Response (for Task 2); Coherence and Cohesion; Lexical Resource; Grammatical Range and Accuracy; these criteria are weighted equally and the score on the task is the average. A final writing score in this quasi-experiment was generated by averaging the two markers' bands on both tasks and rounded up applying the aforementioned IELTS rule. As two examiners were involved in scoring the writing tasks at two times, inter-examiner reliability was calculated so as to examine whether their rating had been consistent and whether scores were reliable. A high degree of reliability was found between examiner 1 and examiner 2; the average measure ICC for T1 IELTS Writing Part 1 was .978 with a 95% CI from .966 to .985, for T1 IELTS Writing Part 2 was .972 with a 95% CI from .958 to .982, for T2 IELTS Writing Part 1

[13] For IELTS writing full assessment criteria and band descriptor, see https://www.ielts.org/-/media/pdfs/writing-band-descriptors-task-1.ashx?la=en and https://www.ielts.org/-/media/pdfs/writing-band-descriptors-task-2.ashx?la=en

was .986 with a 95% CI from .979 to .991, for T2 IELTS Writing Part 2 was .976 with a 95% CI from .964 to .984.

Performance on IELTS speaking was assessed by one experienced IELTS speaking instructor using procedure criteria provided IELTS (2018); in short, performance was assessed on the following four criteria: Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy, Pronunciation[14]. Same as the real IELTS test, participants were given the opportunity to appeal for a second marking for their speaking module if they felt the band they were given were not justified, yet so far, no participant had initiated an appeal.

Oxford Online Placement Test

As discussed earlier, OOPT is a computer-based adaptive test, scores of which are generated automatically on a scale of 120 at the end of the test. An in-depth explanation of the score calculation algorithm and how scores were linked to CEFR could be found in Pollitt's study (2010); thus, here in this section, a short summary of key facts related to score calculation is provided. According to Pollitt (2010), each time a question is answered correctly the system raises its estimate of the candidates' score, and presents the candidate with a more difficult item. When the candidate gets one question wrong, then the estimate and the selection of following items to present are both lowered. Through this OOPT quickly "brackets participants' performance level" (p. 5).

As OOPT consists of two parts, Use of English and Listening, candidates begin the test with Use of English and the system starts scoring them as they move forward with the test. When they have completed the Use of English test, the system uses their Use of English score to choose what Listening test items need to be presented next. When Listening test is completed, participants get their Listening scores. An overall OOPT score is generated by averaging both Use of English and Listening scores equally. Apart from receiving a numeric score at the end of the test, all scores were reported together with a corresponding CEFR level and descriptors[15]. Only the scores are used for later analyses.

[14] For IELTS speaking full assessment criteria and band descriptor, see https://takeielts.britishcouncil.org/sites/default/files/IELTS_Speaking_band_descriptors.pdf

[15] More for details on how OOPT scores map to CEFR levels, see Pollitt (2009).

Spot-the-word

Performance on Spot-the-word was scored by myself; 1 point was given when a participant ticked the right word; 0 was given when the answer was incorrect. Given that this task was included to assess whether a participant can discriminate a real word from a made-up word, if an item was left unanswered, 0.5 point was given because there was a fifty-fifty chance that participants could answer this item correctly. An overall score was generated by adding up all the points participants achieved out of 60 test items.

Speed-of-comprehension

Speed-of-comprehension performance was measured in two ways: accuracy and speed. Accuracy was scored by myself. 1 point was given when participants put a tick on a true sentence or crossed off a false sentence. If a participant put a question mark before a sentence, half a point was given because there was a chance that this question mark was due to the lack of vocabulary, i.e. he/she did not understand certain words in that sentence. If that was the case, there was a fifty fifty chance that he/she could answer correctly, hence the 0.5. An overall Accuracy score was generated by adding up all the points participants achieved out of 100 test items. Participants' processing speed was also measured during this test using a timer. Time needed by a participant to complete all the sentences was noted down and converted into seconds by the end of the test.

3.1.9 Hypotheses

RQ 1: To what extent does IELTS preparation course improve candidates' IELTS scores (overall and by skill)?

As noted in the previous literature review chapter, although the effects of IELTS preparation on IELTS performance varied from one study to another, overall, test preparation was found to have positive impact in terms of improving candidates' IELTS performance. On the basis of this, it was hypothesized that the 4-week intensive IELTS preparation under investigation could boost candidates' IELTS scores both overall and by skill. Moreover, this increase in IELTS performance was predicted to be statistically significant, which means, given the typicality of the observed intervention in Chinese test preparation context, it can be assumed that within a short period of time, by attending dedicated preparatory courses, Chinese IELTS candidates in general could achieve higher scores at the end of their courses, in comparison to their pre-preparation IELTS scores.

Given that this quasi-experiment adopted a pre-test intervention post-test design, it is important to put forward the effects of repeated measures (i.e. the practice effect), where performance is likely to increase at post-test simply because participants had gained experience in the pre-test. As discussed earlier in this chapter, to avoid noise from the practice effect, a control group is included in this quasi-experiment. If the practice effects were to take place, there would be increase in both groups' time 2 IELTS performance and these increases were to be of similar degree.

RQ2: To what extent does IELTS preparation course improve candidates' general English proficiency?

For this research question, it was predicted that IELTS preparation courses might not lead to significant improvement in candidates' general English proficiency on the basis of the following rationale. Firstly, the overall goal and the nature of the intervention under investigation, as previous presented, the goal of the 4-week intensive preparation courses was to ready candidates for the test, not to enhance their overall proficiency. This aim was also reflected in the design of course content in the sense that most teaching and learning activities were centred around test familiarization, using retired IELTS test papers for practices, similar to the typical preparation course content in 2.6.1. Although it could be argued that such activities could also provide value input to enhance general proficiency, these activities were more inclined to boost testwiseness, which was in line with findings from previous literature on test preparation (e.g. Tunks, 2001; Xie, 2013). Thus, it is hypothesized that while a significant increase is likely to be observed in the intervention participants' IELTS performance, there would not be significant changes in their general proficiency as indexed by OOPT. Furthermore, it was hypothesized that participants' lexical knowledge as indexed by Spot-the-word or their accuracy of sentence processing and speed as indexed by Speed-of-comprehension would not be significantly improved either.

Were these two predictions to be confirmed, there would also be implications on the reliability and the validity of IELTS, particularly for the construct and the predictive validity. Widely used as a measure of English language proficiency among ESL/EFL learners, IELTS is generally acknowledged as reliable. However, this reliability could be threatened by dedicated test preparation because if scores could be boosted within a short period of time, the reliability of IELTS could be weakened. Moreover, if this boost in scores is not the result of improved general proficiency but only testwiseness, the construct validity of IELTS could be undermined as construct irrelevant variances were introduced by test preparation practices. Further, once the construct validity had been eroded, the predictive validity of IELTS is likely to be affected as well, which could consequently lead to IELTS being an inconsistent predictor for international students' academic attainment as demonstrated in the previous predictive literature.

On a more explorative note, in this quasi-experiment, it is hypothesized that the effect of IELTS preparation would be different on those of comparatively higher proficiency and those of comparatively lower proficiency. This is based on the plateau effect denoted in a few test preparation studies (e.g. Brown, 1998; Green, 2007) where researchers have claimed, at lower levels, relatively shorter periods of instruction are needed in order to achieve measurable improvements in proficiency (indexed by score gains), but at higher levels, considerably longer periods are usually required. Thus, it is suspected that for participants who began the preparation course with a lower proficiency, they were more likely to achieve bigger score boost compared with those who began with a comparatively higher proficiency.

3.1.10 Analyses

Both SPSS 24 and R Version 3.5.0 were employed for statistical analysis and data visualisation. Data were checked for normal distribution using Shapiro Wilk test and results indicated T1 and T2 IELTS scores (overall and by skills), T1 and T2 reading processing speed, T1 OOPT Listening were not normally distributed for both groups, p<.05 in all cases. Also, control group's T1 OOPT use of English scores, T1 and T2 vocabulary scores, the intervention group's T2 OOPT overall scores, T2 sentence processing accuracy scores, were not normally distributed. In the light of data distribution, a combination of parametric and non-parametric tests were used in the subsequent data analyses.

Descriptive statistics (i.e. Mean, SD, Median and Range) were produced first to give a general understanding of the research sample and participants' performance on each measure at both time points. Mann-Whitney U test were used to examine the between-group differences in IELTS and other general proficiency measures at two time points, followed by the repeated measures ANOVA test to examine the effect of time, group and the interaction between time and group, as literature suggested that ANOVA is robust even for non-normally distributed data (e.g. Glass, Peckham & Sanders, 1972).To better visualise the data and the trends, statistics for central tendency (i.e. Mean, SD and SE) were used in producing the charts.

Descriptive statistics and results from the analyses were reported in the next section in relation to the proposed research questions. Confirmatory analyses were presented first to examine these afore-stated hypotheses, followed by exploratory analyses that revealed some unexpected findings.

3.2 Results and Analysis

3.2.1 To what extent does IELTS preparation course improve candidates' IELTS scores (overall and by skill)?

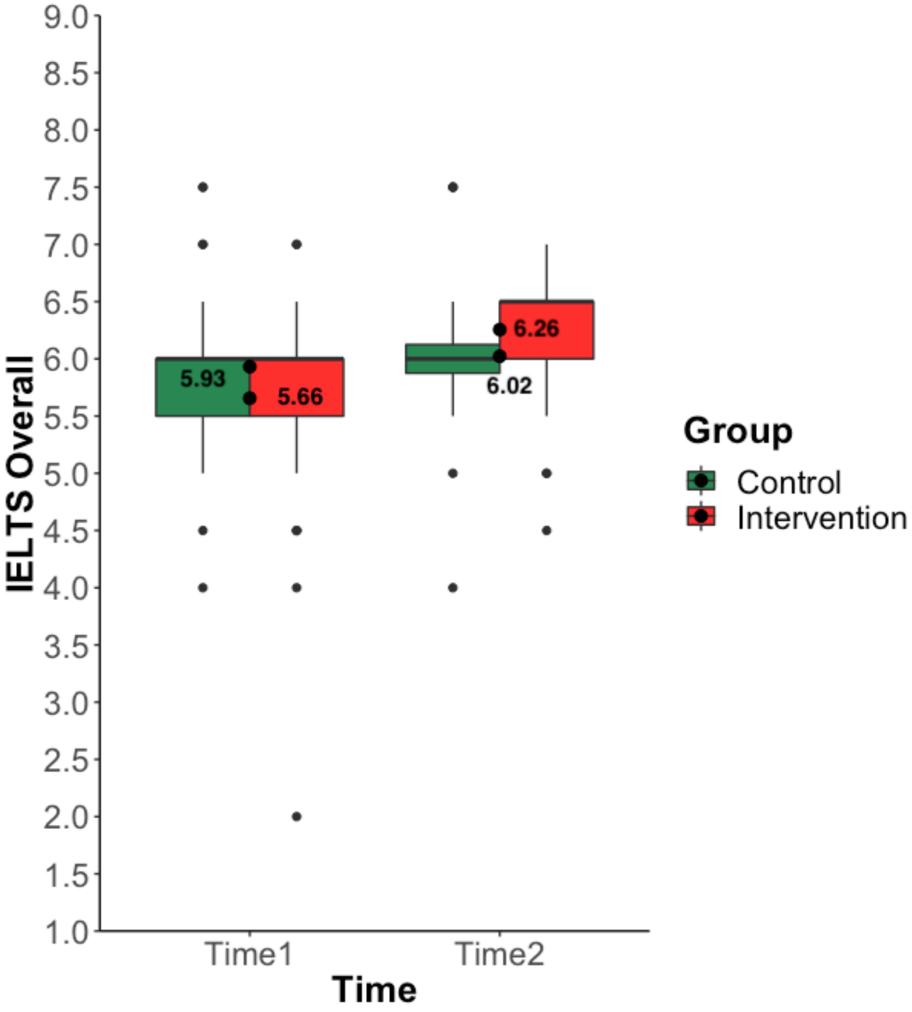
To answer this question, descriptive statistics were presented first to provide an overall picture of the participants' score change from T1 to T2, then Mann-Whitney U tests and ANOVA were performed to statistically examine the effect of test preparation.

The control and intervention participants' performance on IELTS overall at time 1 (T1) and time 2 (T2) were summarised in Table 3.2.1, along with a visual depiction of data density and between-group differences on IELTS overall at both timepoints (Figure 3.2.1).

Table 3.2.1 The intervention and control participants' T1 and T2 IELTS overall scores and between-group differences

Measures		Intervention (N=45)				Control (N=44)				*r*
		Mean	SD	Median	Range	Mean	SD	Median	Range	
IELTS Overall	T1	5.66	.86	6.0	2.0-7.0	5.93	.64	6.0	4.0-7.5	-.14
	T2	6.26	.60	6.5	4.5-7.0	6.02	.61	6.0	4.0-7.5	.25

Figure 3.2.1 The intervention and control participants' T1 and T2 IELTS overall score distribution and both groups' mean scores at two time points (N=89)



The subsequent boxplots follow the same format as this boxplot; thus to avoid repetition, detailed explanations are only provided once here. This boxplot displays how data from both groups' scores on IELTS overall were distributed at two time points. The bold horizontal lines are the median. The whiskers above the median represent the top 25% of the data while the whiskers below at the bottom of the box represent the bottom 25% of the data. The solid dots above/below the whiskers are outliers who achieved unusually high or low scores. The solid dots within the boxes along with the exact statistics are the mean scores for each group at both timepoints.

As shown, at T1, the intervention participants' IELTS overall score had a larger range (Min=2, Max=7) than that of control (Min=4, Max=7.5), suggesting that candidates with low IELTS scores were more likely to attend test preparation so as to ready themselves for the test than those with higher scores, which was in line with the overall goal of the preparation course. At T1, control group's mean was higher than that of the intervention and there were more participants with an overall 6, the median, (N=23, 52%) in the control than that in the intervention (N=18, 40%). As data were not normally distributed, Mann-Whitney U test was carried out to test the statistical significance of the observed between-group differences in T1 mean score. U-test found the difference to be insignificant, U=842.00, p>.05, effect size r=-.14, which suggests that both groups' overall IELTS performance was well-matched at T1, setting an even playground for the following examination regarding the effect of the intervention.

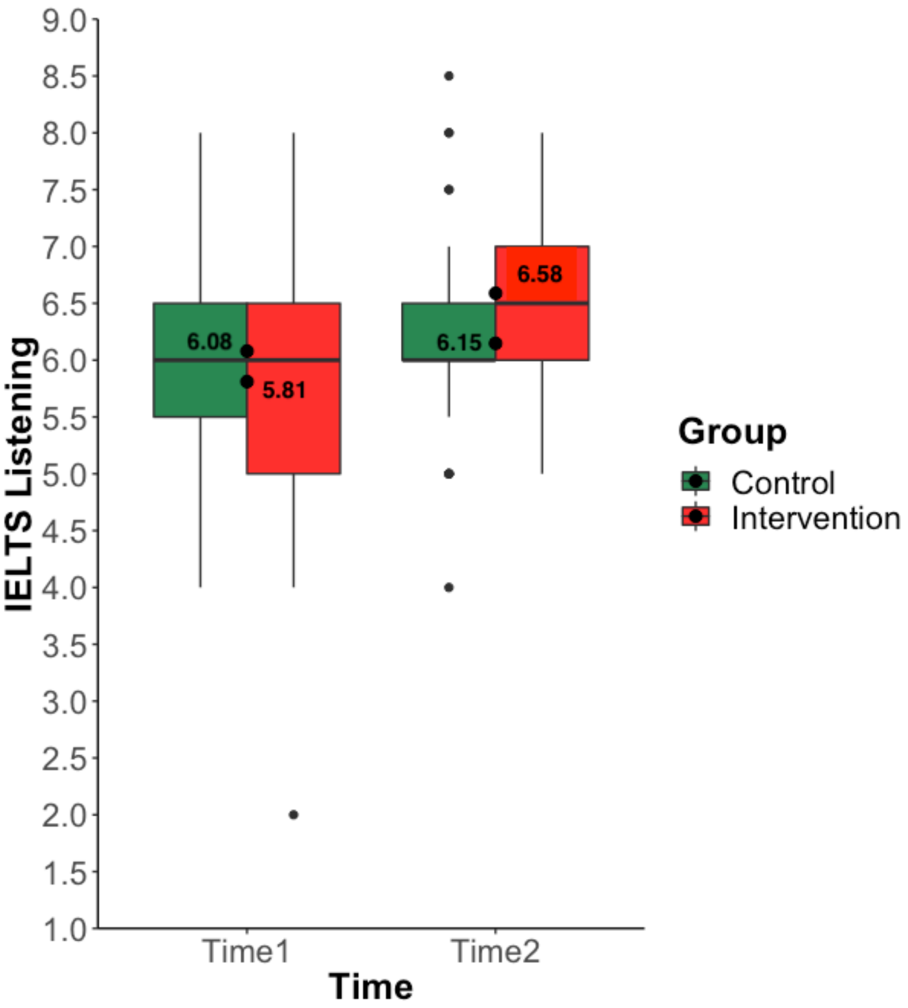
At T2, the intervention's score range became smaller (Min=4.5, Max=7) while control's range remained the same (Min=4, Max=7.5). This shortening of score range could be interpreted as an effect of the intervention, suggesting that by attending test preparation, candidates with very low IELTS scores had succeed in improving their scores. This time, the intervention participants' mean and median IELTS overall were both higher than that of control and this between-group difference was found significant, U=1226.00, p<.05. Although there were still more control participants achieving an overall 6 (the median) at T2 (N=22, 50%) in comparison with the intervention (N=14, 31%), the proportions of the intervention participants achieving 6.5 (N=14, 31%) and 7 (N=10, 22%) were much bigger than that of control (overall 6.5 N=8, 18%; overall 7 N=0).

Breaking down the overall IELTS scores into specific modules, Table 3.2.2 summarises all participants' performance on IELTS listening at T1 and T2 by group, the effect size of between-group differences at both timepoints, along with a visual display of data distribution at T1 and T2 (Figure 3.2.2). As shown, the trend observed in IELTS listening was very similar to that of IELTS overall as presented above.

Table 3.2.2 The intervention and control participants' T1 and T2 IELTS listening scores and between-group differences

Measures		Intervention (N=45)				Control (N=44)				*r*
		Mean	SD	Median	Range	Mean	SD	Median	Range	
IELTS listening	T1	5.81	1.14	6.0	2.0-8.0	6.08	.90	6.0	4.0-8.0	-.13
	T2	6.58	.73	6.5	5.0-8.0	6.15	.87	6.0	4.0-8.5	.31

Figure 3.2.2 The intervention and control participants' T1 and T2 IELTS overall score distribution, and both groups' mean scores at two time points (N=89)



At T1 IELTS listening, there were more intervention participants scoring lower than 6, the median (N=22, 49%) than that of control (N=13, 30%) and on average, there was a difference between group with the intervention participants lagging behind. U-test found the T1 between-group differences to be insignificant, *U*=843.50, *p*>.05, suggesting that both groups' listening performance was well matched at T1.

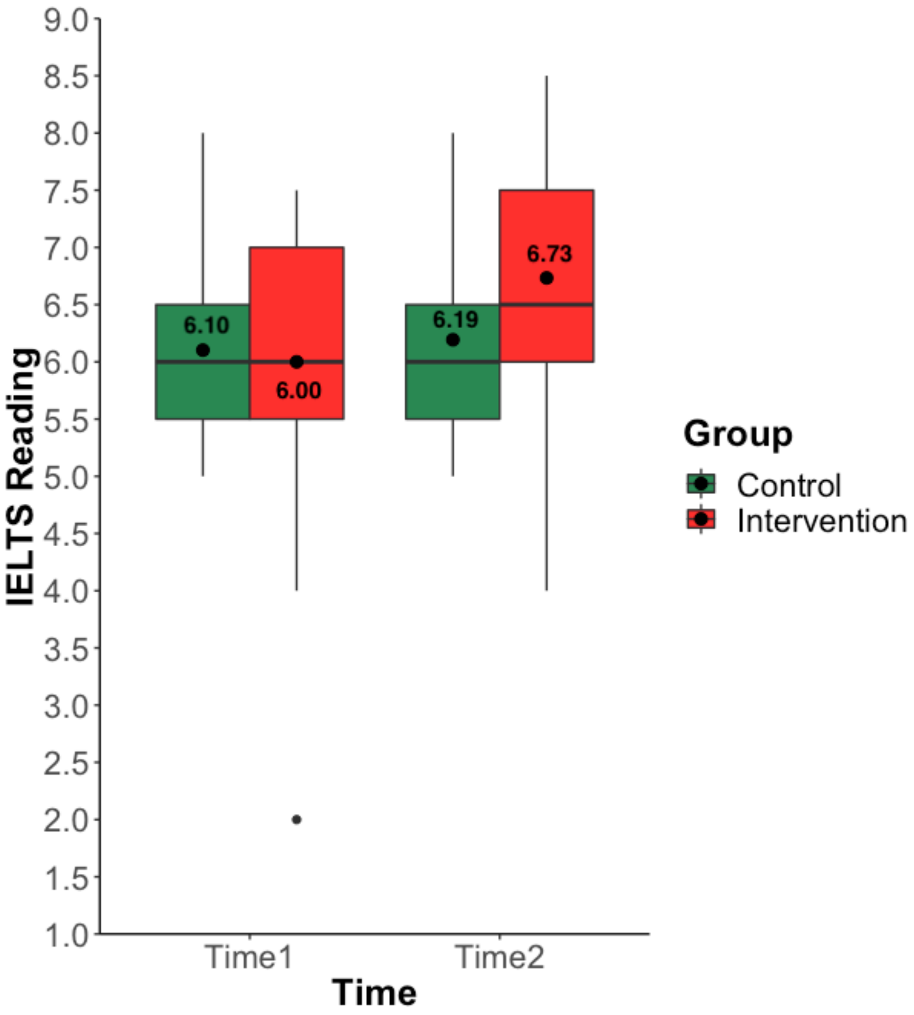
As shown in Figure 3.2.2., at T2 when the intervention was completed, there were noticeable improvement in the intervention group's listening scores (Mean and Median) while that of the control remained unchanged throughout time. Moreover, the proportion of participants scoring lower than 6 (the Median for both groups at T1) was much smaller in the intervention group (N=5, 11%) than that in the control (N=10, 23%). Further, this time, the intervention participants were no longer lagging behind; they outperformed control and this T2 between-group differences were found significant, *U*=-1341.00, *p*<.01, effect size *r*=.31.

With regard to IELTS reading and the effect of test preparation, performance on T1 and T2 IELTS reading was summarised in Table 3.2.3 and visualised in Figure 3.2.3.

Table 3.2.3 The intervention and control participants' T1 and T2 IELTS reading scores and between-group differences

Measures		Intervention (N=45)				Control (N=44)				*r*
		Mean	SD	Median	Range	Mean	SD	Median	Range	
IELTS reading	T1	6.00	1.11	6.0	2.0-7.5	6.10	.76	6.0	5.0-8.0	.02
	T2	6.73	.94	6.5	4.0-8.5	6.19	.71	6.0	5.0-8.0	.31

Figure 3.2.3 The intervention and control participants' T1 and T2 IELTS reading score distribution, and both groups' mean scores at two time points (N=89)



At T1, although participants' from both groups performed at a very similar level on average, *U*=1007.50, *p*>.05 and that the proportion of participants scored lower than the median, 6, was the same in control as that in the intervention (N=16, 36% for both groups), there was one intervention participant scoring as low as 2, thus introducing more disparity to the overall intervention group's reading performance at T1. In addition, there were more high-achieving control participants, with 2 scoring 8, while the highest score intervention participants achieved was only 7.5.

At T2, while the control's mean and score distribution remained almost the same as T1, obvious changes could be observed among the intervention participants. To begin with, they had significantly improved their average reading score by over half a band from 6.00 to 6.73, resulting in a significant between-group difference, *U*=1343.00, *p*<.01.

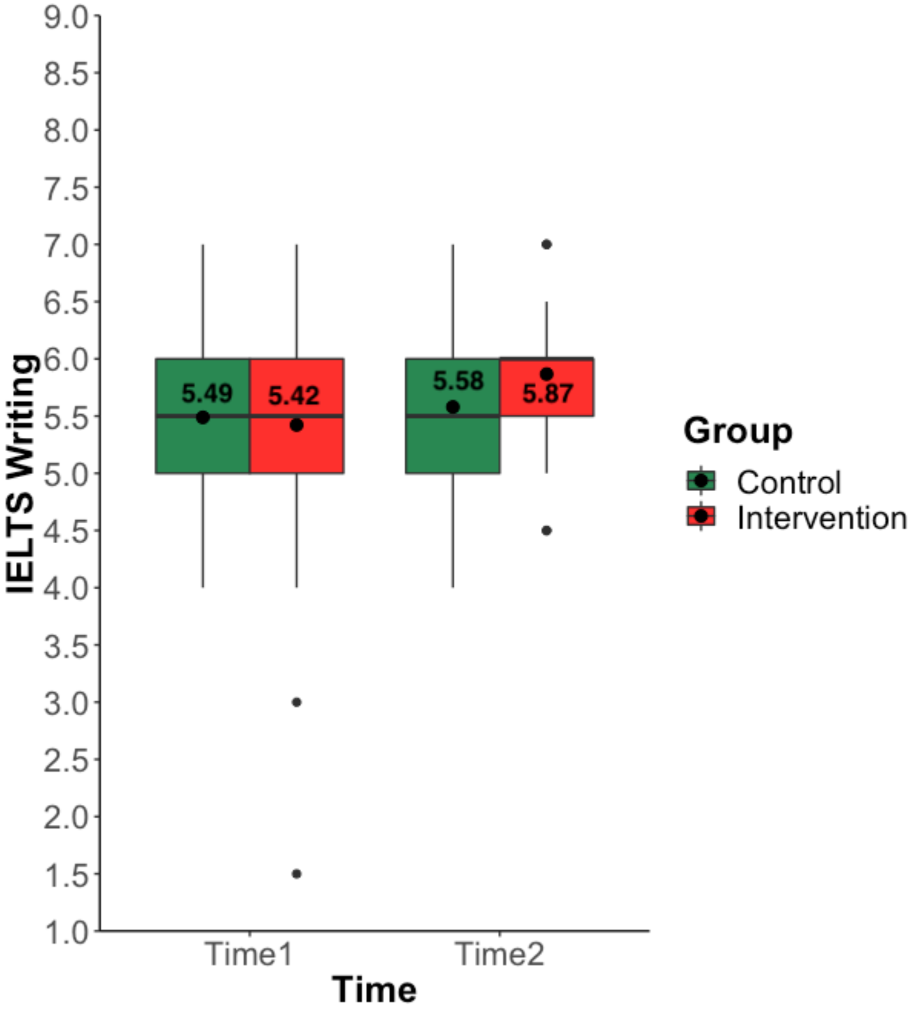
Moreover, the percentage of participants achieving high reading scores (7.5 and above) was eightfold (N=16, 36%) than that in control (N=2, 5%). Furthermore, the large score disparity observed at T1 had also shrunken at T2 in the intervention group, indicating that the intervention under investigation had successfully reduced the ratio of extreme cases.

In terms of IELTS writing and the effect of test preparation, Table 3.2.4 and Figure 3.2.4 summarise and illustrate participants' performance on IELTS writing at T1 and T2.

Table 3.2.4 The intervention and control participants' T1 and T2 IELTS writing scores and between-group differences

Measures		Intervention (N=45)				Control (N=44)				*r*
		Mean	SD	Median	Range	Mean	SD	Median	Range	
IELTS writing	T1	5.42	.98	5.5	1.5-7.0	5.49	.67	5.5	4.0-7.0	.03
	T2	5.87	.62	6.0	4.5-7.0	5.58	.66	5.5	4.0-7.0	.24

Figure 3.2.4 The intervention and control participants' T1 and T2 IELTS writing score distribution, and both groups' mean scores at two time points (N=89)



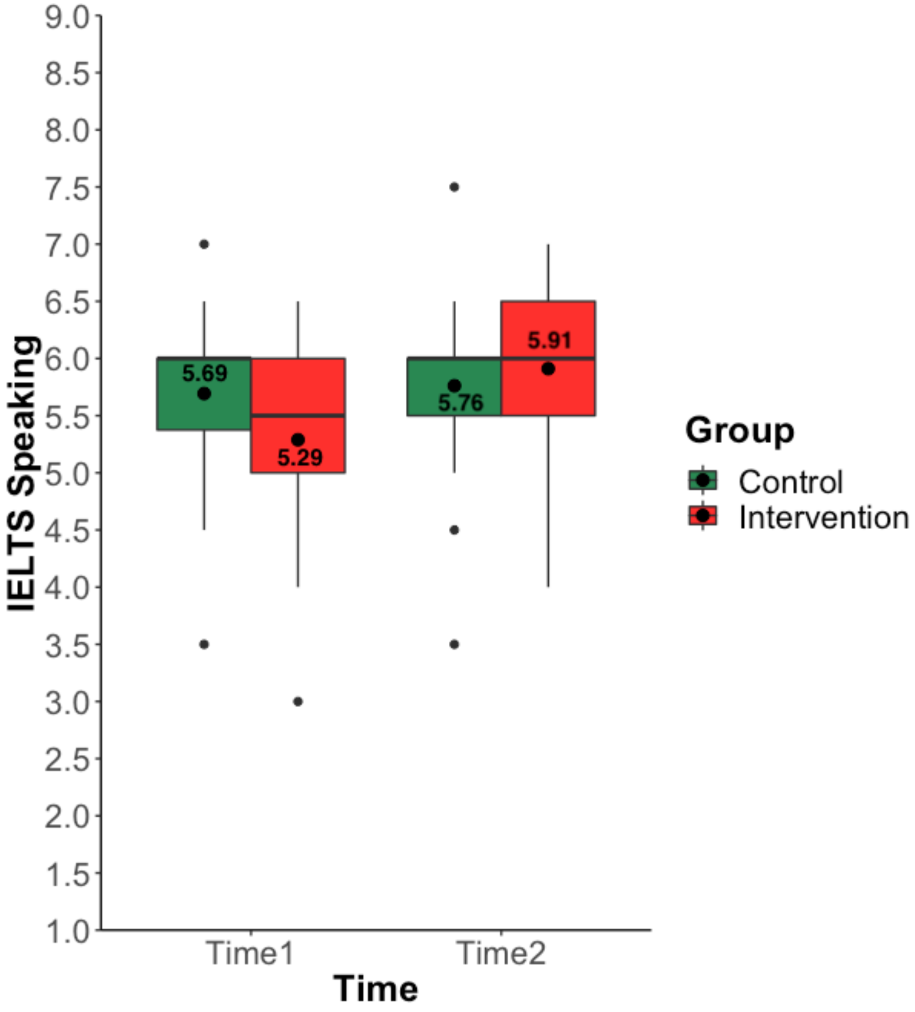
As shown, at T1, both groups' mean, median and score distributions were very similar to one another, *U*=1026, *p*>.05 but there was two participants scoring as low as 1.5 and 3 in the intervention. At T2, while there was very little change in control's overall performance (e.g. mean and standard deviation), the intervention participants managed to improve their performance noticeably. To begin with, the intervention's mean (5.87) and median(6.0) scores were significantly higher than that of the control, *U*=1255.50, *p*<.05. In addition, boxplot shows that there was much less variation in the intervention group's T2 writing scores in comparison with control as scores were more clustered towards the centre and there were relatively less extreme cases. This change in score variation could be interpreted as evidence indicating that the intervention, i.e. IELTS preparation course, had effectively centralised candidate's score range and reduced the proportion of low achieving scores.

In a similar fashion, Table 3.2.5 summarises both groups' performance on IELTS speaking at T1 and T2 and Figure 3.2.5 displays the mean scores as well as the distribution of scores at both times.

Table 3.2.5 The intervention and control participants' T1 and T2 IELTS speaking scores and between-group differences

Measures		Intervention (N=45)				Control (N=44)				*r*
		Mean	SD	Median	Range	Mean	SD	Median	Range	
IELTS speaking	T1	5.29	.71	5.5	3.0-6.5	5.69	.68	6.0	3.5-7.0	-.28
	T2	5.91	.57	6.0	4.0-7.0	5.76	.69	6.0	3.5-7.5	.13

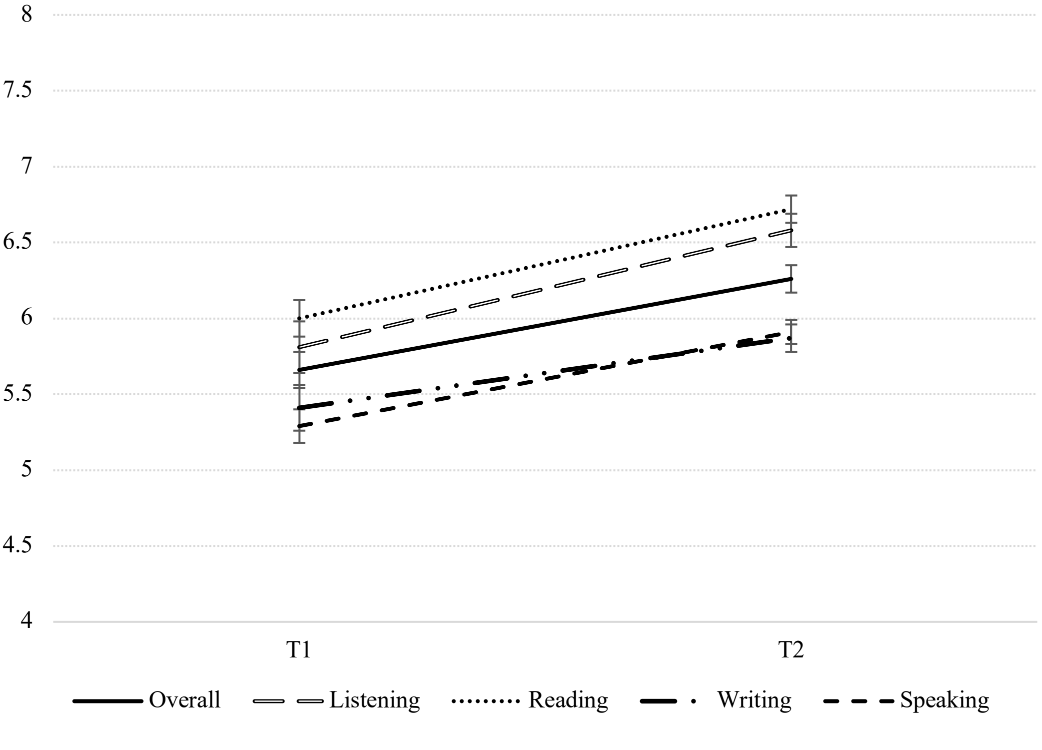
Figure 3.2.5 The intervention and control participants' T1 and T2 IELTS speaking score distribution and both groups' mean scores at two time points (N=89)



As shown, there was an obvious and significant between-group difference in speaking scores at T1, *U*=673.00, *p*<.01 with the intervention participants lagging behind with an effect size of .28. The above boxplot also showed that at T1, the intervention group's speaking score had a larger spread than that of the control, whose scores were more closely clustered towards the centre. At T2, while control's performance changed very little, the intervention participants managed to increase their scores to outperform the control. Although the difference in T2 speaking scores was not statistically significant, the closing of significant T1 gap indicates that the intervention, i.e. IELTS preparation, had significant effects on improving speaking scores.

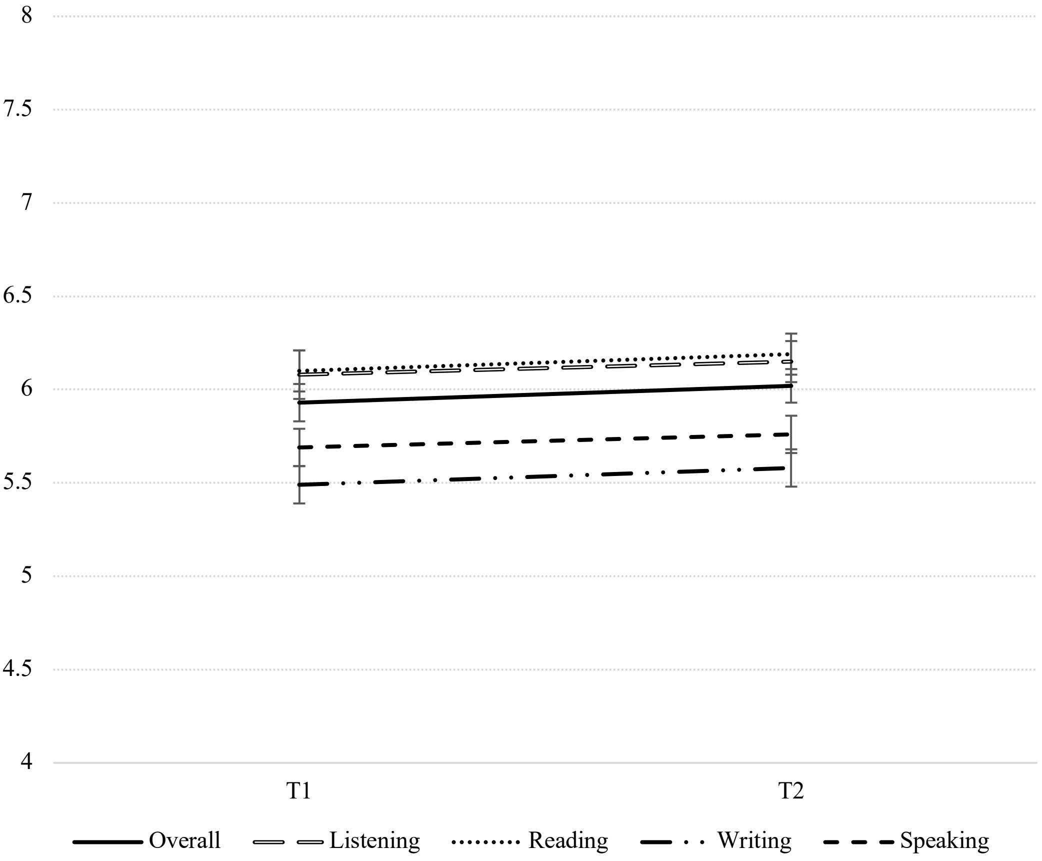
In terms of within-group score change, depicted in Figure 3.2.6, shows that there were clear upward trends in the intervention participants' IELTS performance, overall and by skill. The bigger score gains achieved, depicted by the two steeper lines, were found in listening (score gain M=.77, Median=1, Range=-.5-3.0) and reading (score gain M=.73, Median=.50, Range=-.5-2.5), i.e. the two receptive skills. In comparison, writing (score gain M=.46, Median=.5, Range=-.5-2.5) and speaking (score gain M=.62, Median=.5, Range=0-2.5), i.e. the two productive skills, were observed to be more challenging for a short-time score boost as the lines remained comparatively flatter.

Figure 3.2.6 The intervention participants' average IELTS score change from T1 to T2 (overall and by skill) (N=45) (Error bars ± 1 standard error)



By contrast, as shown in Figure 3.2.7, there was no noticeable change in control participants IELTS performance, which lends support to the reliability of the measures used. The slight increase observed in control's T2 IELTS performance could be attributed to the practice effect commonly found in quasi-experiments using a repeated-measures design as participants' performance becomes better due to the accumulated knowledge and experience of the tasks.

Figure 3.2.7 Control participants' average IELTS score change from T1 to T2 (overall and by skill) (N=44) (Error bars ± 1 standard error)



The effects of the intervention on IELTS scores were further examined using repeated measures ANOVA, results of which are summarised in Table 3.2.6. Here, time acted the within subject factor with two levels (T1, T2), group acted as the between subject factor (intervention, control), IELTS scores (overall and by skill) acted as the outcome.

Table 3.2.6 Comparison of The Intervention and Control participants' group means on IELTS (overall and by section), taken at T1 and T2 (N=89)

Measures		*F*-test statistics	*p* value
IELTS	Overall	$F_{time}(1,87)=66.30$	.000
		$F_{group}(1,87)=.02$	.877
		$F_{time*group}(1,87)=35.99$	.000
	Listening	$F_{time}(1,87)=42.03$	.000
		$F_{group}(1,87)=.19$	.663
		$F_{time*group}(1,87)=29.42$	.000
	Reading	$F_{time}(1,87)=45.04$	.000
		$F_{group}(1,87)=1.40$	.240
		$F_{time*group}(1,87)=27.15$	.000
	Writing	$F_{time}(1,87)=24.55$	.000
		$F_{group}(1,87)=.50$	.482
		$F_{time*group}(1,87)=10.93$	.001
	Speaking	$F_{time}(1,87)=56.35$	.000
		$F_{group}(1,87)=.92$	.342
		$F_{time*group}(1,87)=36.29$	.000

Overall, ANOVA found there was significant effect of time, indicating that all participants' performance at T2 (overall and by skill) was significantly different from that at T1 but the effect of group was not significant. More importantly, the effect of interaction (time*group) was found significant in both at the overall level and at each individual module level, indicating that the effect of time was different between group. Taking into consideration that there was little changes in control participants' performance as shown in Figure 3.2.7, it could be concluded that the main effect of time was mostly attributed to the score gains achieved by the intervention group, lending support to the afore-stated hypothesis that by attending test preparation (the intervention) candidates could boost their scores significantly within a short period of time.

3.2.2 Exploratory analysis: do candidates of lower proficiency achieve more IELTS score gains (overall and by skill) than candidates of higher proficiency through attending IELTS test preparation?

This question, based on the plateau effect noted in previous literature (e.g. Brown, 1998; Green, 2007), could be answered by correlating intervention participants' pre-course

IELTS scores (overall and by skills) with the score gains achieved through the intervention.

Descriptive statistics summarised in Table 3.2.7 showed that intervention participants who started with a low IELTS level achieved greater score gains both at an overall level and at a module level through attending the test preparation courses. In comparison, participants who started with a fairly high IELTS scores, the gains achieved were considerably smaller.

Table 3.2.7 IELTS score gains (overall and by skill) achieved by intervention participants' grouped according to participants' initial IELTS scores (N=45)

Score gains	Overall		Listening		Reading		Writing		Speaking	
Initial IELTS	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
2.0	2.50	-	3.00	-	2.00	-	3.00	-	1.00	-
4.0 (N=1)	2.00	-	2.50	-	2.00	-	2.50	-	2.50	-
4.5 (N=3)	.83	.58	1.00	.00	1.00	1.00	1.17	.76	.67	.76
5.0 (N=6)	.83	.41	1.00	1.05	1.17	.52	.42	.66	.83	.61
5.5 (N=8)	.63	.35	.81	.46	1.00	.85	.31	.26	.81	.53
6.0 (N=18)	.39	.32	.75	.67	.36	.61	.19	.35	.33	.34
6.5 (N=6)	.50	.00	.17	.26	.58	.38	.50	.32	.75	.27
7.0 (N=2)	. 00	.00	-.50	.00	.25	.35	.00	.00	.25	.35

Based on these descriptive statistics, it is reasonable to infer that the initial proficiency level (indicated by pre-course IELTS overall score) was associated with the achievable score gains. To test for this assumption, correlation analysis was run and the results indicated that the achieved score gains were indeed significantly correlated with intervention participants' pre-course IELTS level, overall gain *r*=-.733, listening gain *r*=-.629, reading gain *r*=-.502, writing gain *r*=-.673, speaking gain *r*=-.372, *p*<.05 in all cases. The negative correlation coefficient indicated that participants with higher initial proficiency level (as indicated by higher IELTS) were less likely to achieve greater score gains. This finding resonated with the regression to the mean phenomenon or the plateau effect noted in a number of previous test preparation studies (e.g. Brown, 1998; Green, 2007) and was in line with Elder & O'Loughlin's statement that "the language proficiency one had prior to the beginning of the course is the most constant indicator of how far one is likely to travel" (2003, p. 226).

3.2.4 To what extent does IELTS preparation course improve candidates' general proficiency, measured by linguistic measures they were not prepared for?

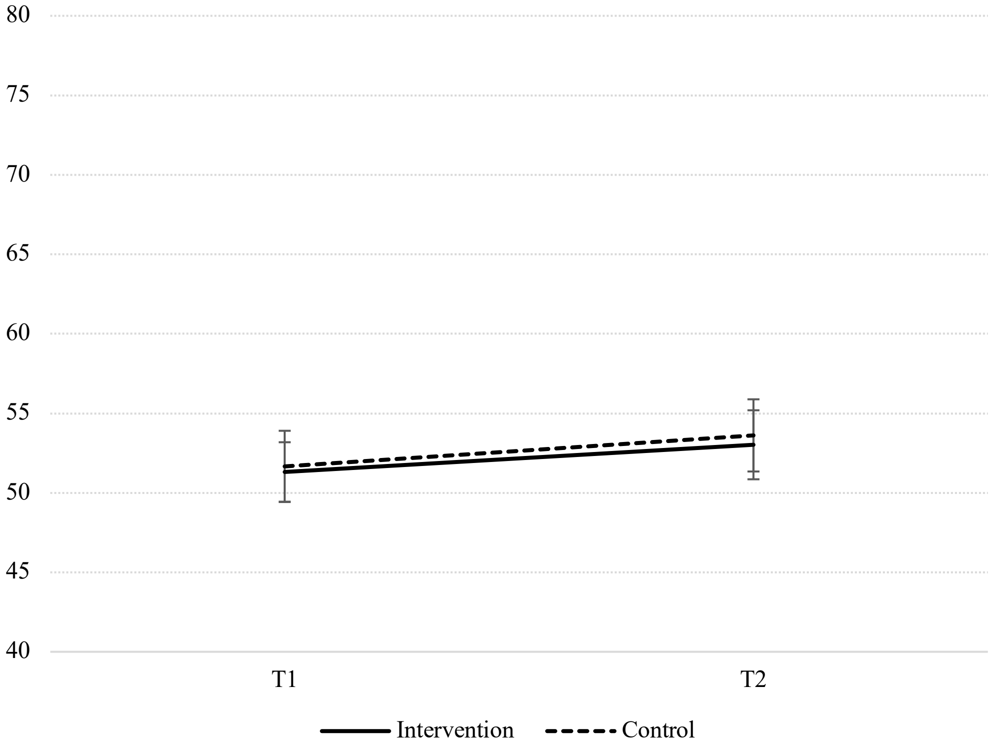
To answer this question, descriptive statistics were presented first to showcase participants' performance on general proficiency measures at T1 and T2, then Mann-Whitney test and repeated measures ANOVA were performed to examine the effect of the intervention, group, and the interaction.

Table 3.2.8 The intervention and control participants' T1 and T2 OOPT overall scores and between-group differences

OOPT		Intervention (N=45)				Control (N=44)				*r*
		Mean	SD	Median	Range	Mean	SD	Median	Range	
Overall	T1	51.31	12.56	54	21-73	51.66	14.86	49.5	17-91	.05
	T2	53.02	15.22	53	17-87	53.61	14.38	56	19-86	.04
Use of English	T1	58.89	16.67	60	24-94	60.70	17.00	63	25-95	-.05
	T2	61.78	19.02	67	6-92	62.05	16.71	66	29-86	.00
Listening	T1	43.60	13.75	46	1-62	42.68	16.33	41.5	9-91	.14
	T2	43.49	18.64	45	8-85	45.16	16.58	45.5	9-91	-.05

As shown, the between-group differences on OOPT (overall and by section) were less predominant than that observed in IELTS (Table 3.2.1-3.2.5). At T1, although the intervention participants scored lowered on both OOPT use of English and OOPT listening, Mann-Whitney U test revealed that the differences between groups were not significant, use of English *U*=936.00, *p*>.05, listening *U*=830.00, *p*>.05. There was no significant difference between group in OOPT overall scores either, *U*=932.50, *p*>.05. This insignificant differences in OOPT performance further illustrated that the two groups under investigation were well-matched in terms of their general proficiency at T1, setting the playground even for the later examination of the effects of the intervention.

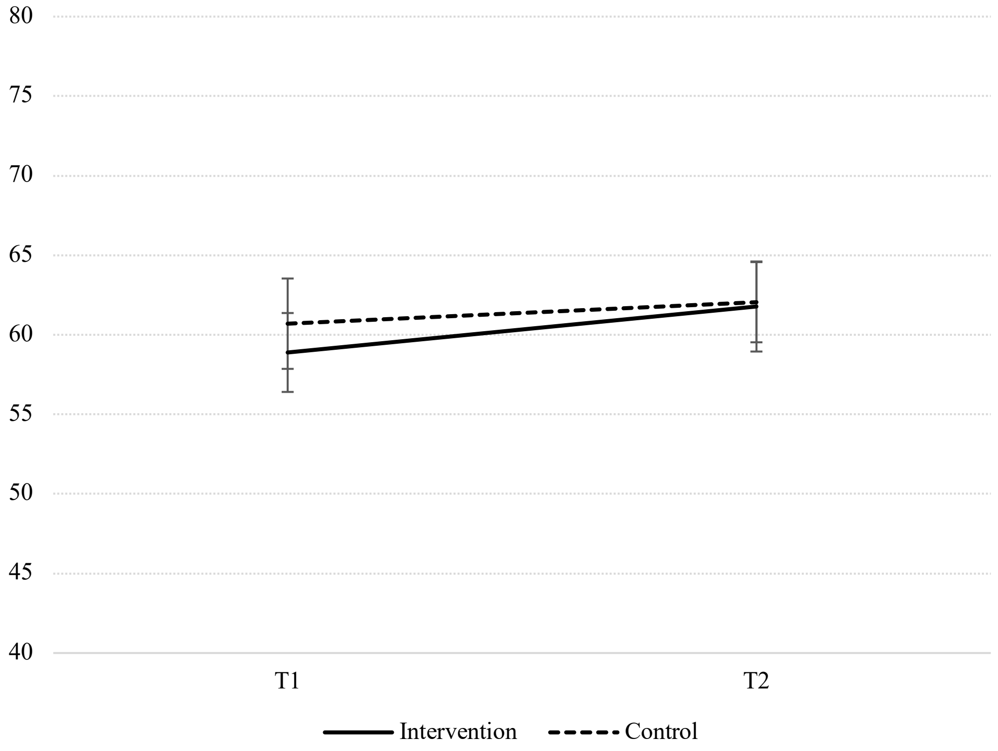
Figure 3.2.8 The intervention and control participants' T1 and T2 OOPT overall scores (N=89) (Error bars ± 1 standard error)



Unlike the steep upward trends in the intervention participants' IELTS scores observed earlier (Figure 3.2.6), there was little change in their OOPT performance from T1 to T2. As shown in Figure 3.2.10, the solid lines (representing the intervention) remained relatively flat over time.

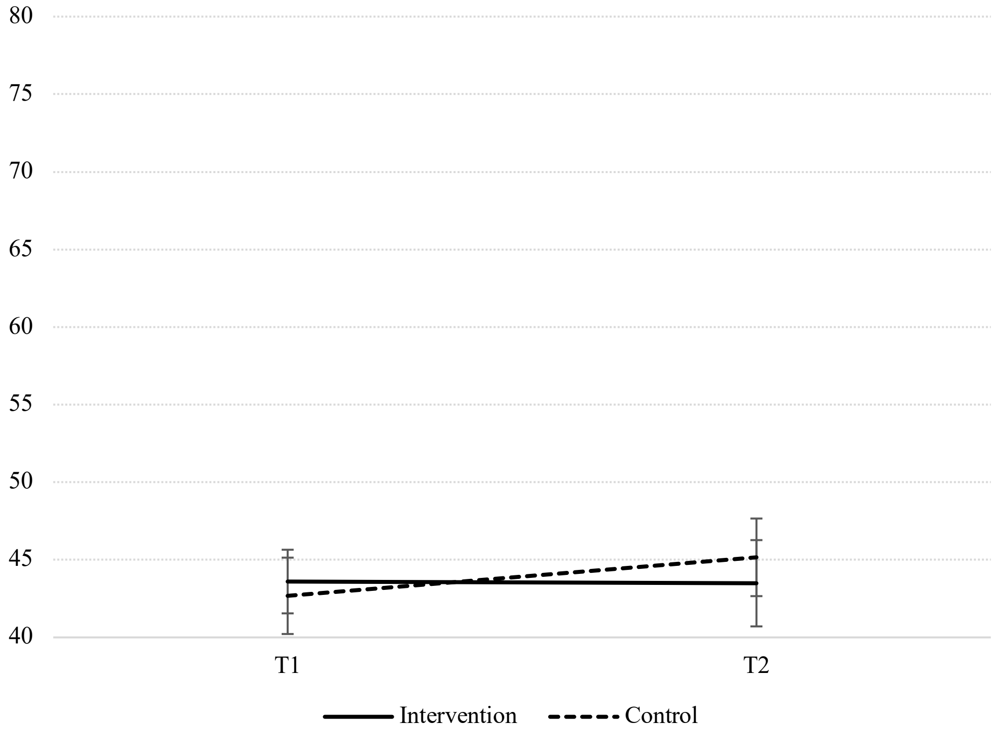
For OOPT overall, there were slight increases in both groups' scores but the between-group differences at T2 were not significant, *U*=949.50, *p*>.05. Moreover, taking into consideration that these increases observed in both lines were of very similar slopes, it is most likely that they were attributed to the practise effects common in repeated measures design.

Figure 3.2.9 The intervention and control participants' T1 and T2 OOPT use of English scores (N=89) (Error bars ± 1 standard error)



For OOPT use of language, similar patterns were shown; although the intervention participants achieved slightly more gains than the control, the between-group differences at T2 were not significant either, *U*=989.00, *p*>.05.

Figure 3.2.10 The intervention and control participants' T1 and T2 OOPT listening scores (N=89) (Error bars ± 1 standard error)



For OOPT listening, trends observed presented a somewhat confusing picture at first glance; unlike OOPT overall and OOPT use of English, here, control participants managed to achieve more gains while the intervention's score remained flat from T1 to T2, as shown by the interaction between the lines in Figure 3.2.12. Nonetheless, U-test revealed that at T2 OOPT listening, there were no significant differences either *U*=.929.50, *p*>.05. This forms stark contrast to the score change observed in IELTS listening (Table 3.2.2 and Figure 3.2.2). As OOPT listening and IELTS listening assess the same construct (i.e. English listening proficiency), albeit that OOPT listening focuses on both candidates' linguistic knowledge as well as their pragmatic knowledge, while IELTS listening focuses primarily on assessing candidates' linguistic knowledge, one would expect enhancement in listening skills would translate to score improvements on both tests. Therefore, the lack of score gains on OOPT listening indicates that the intervention did not improve candidates' listening proficiency.

Together, OOPT results revealed that there was little change in participants' general proficiency, regardless of whether they had received the intervention. The contradiction between obvious and significant score gains in IELTS and the little progress in their general proficiency suggested that knowledge and skills learned through IELTS preparation courses were mainly applicable to the IELTS taking, not the improvement of general proficiency. From this, one can also say that the boost in IELTS scores did not correspond to the equivalent improvement of general proficiency.

Two additional measures known from previous research to be tightly linked to the construct of language proficiency—lexical knowledge, sentence processing accuracy and speed—were used to further scrutinize the effects of IELTS preparation on candidates' general proficiency. Descriptive statistics on participants' performance on these two measures were summarised in Table 3.2.10.

Table 3.2.9 The intervention and control participants' T1 and T2 vocabulary and sentence processing accuracy scores and between-group differences

	Intervention (N=45)		Control (N=44)		
Measures	Mean	SD	Mean	SD	Cohen's d
Vocabulary T1	32.30	4.60	33.51	7.57	0.19
Vocabulary T2	32.97	4.57	34.52	6.87	0.27
Sentence processing accuracy T1	73.71	9.28	76.86	8.64	0.34
Sentence processing accuracy T2	73.84	10.06	77.56	9.11	0.39

As shown, scores on vocabulary test and sentence processing accuracy test presented a similar pattern to that on OOPT. At T1, both groups scored very similarly on the vocabulary tests, *t*(87)=.915, *p*>.05, although the intervention group lagged behind by 0.19 SD. Similarly, in terms of participants' sentence processing accuracy, both groups achieved very close scores, *t*(87)=1.657, *p*>.05 although the intervention was slightly outperformed by the control by .34 SD.

The score change in lexical knowledge and sentence processing accuracy, depicted in Figure 3.2.13-3.2.14, also resembled the trends observed in OOPT (Figure 3.2.10-3.2.12). Although upward trends were present in both groups' vocabulary and sentence processing accuracy scores from T1 to T2, little narrowing-of-the-gap could be observed as both lines remained relatively parallel to each other. This indicates that the intervention, i.e. test preparation, did not make a huge difference on vocabulary and sentence processing accuracy. The slight upward trends shown in Figure 3.2.13 and 3.2.14 were most likely due to practice efforts as afore-discussed, and t-test revealed that at T2, there were no significant between-group differences in vocabulary *t*(87)=1.262, *p*>.05, or in sentence processing accuracy *t*(87)=1.823, *p*>.05. In other words, participants did not significantly improve their lexical knowledge or the accuracy of their sentence processing, even with the experience of attending IELTS preparation.

Figure 3.2.11 The intervention and control participants' T1 and T2 Vocabulary test scores (N=89) (Error bars ± 1 standard error)

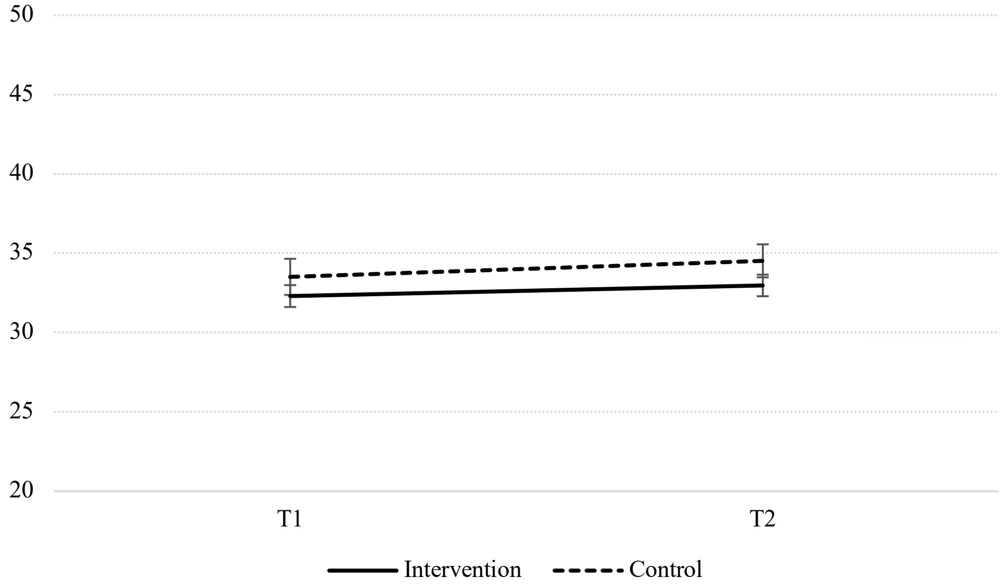
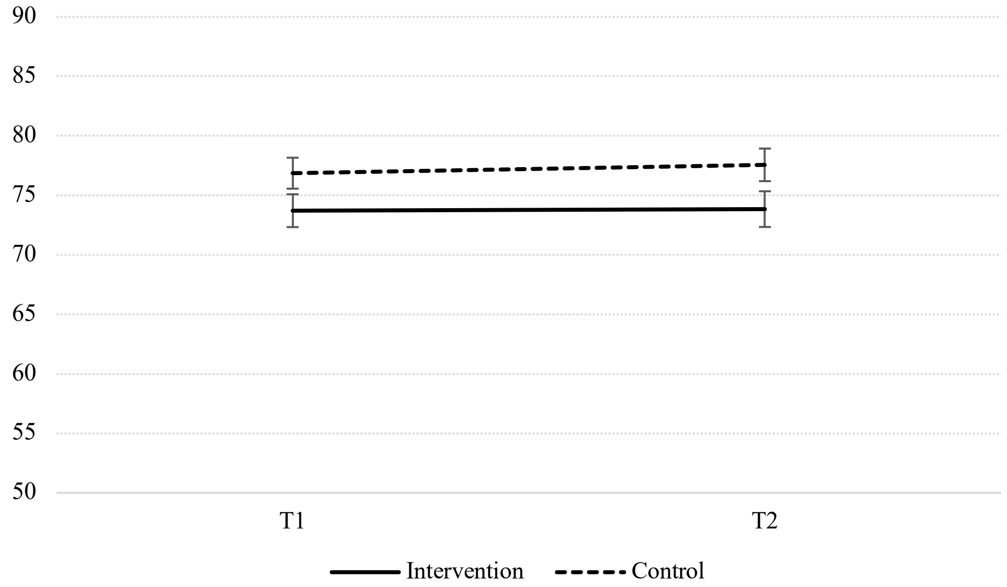


Figure 3.2.12 The intervention and control participants' T1 and T2 sentence processing accuracy scores (N=89) (Error bars ± 1 standard error)



Linking the trends observed in sentence processing accuracy back to that observed in IELTS reading (Table 3.2.3 and Figure 3.2.3), clear differences are found. Although the intervention did result in significant increases in participants' IELTS reading scores, the absence of improvement in participants' accuracy of processing sentences, an important measure of one's reading ability as well as general proficiency, indicates that the intervention did not effectively enhance candidates' reading and general proficiency. On the basis of this, one may speculate that knowledge and skills learned in test preparation courses could be testwiseness-oriented.

The effects of the intervention on general proficiency (indexed through OOPT, vocabulary, sentence processing accuracy) were further examined using repeated measures ANOVA, results of which were summarised in Table 3.2.11. Similar to the previous ANOVA, time acted the within subject factor with two levels (T1, T2), group acted as the between subject factor (intervention, control), OOPT scores, vocabulary test scores, sentence processing accuracy scores acted as the outcome.

Table 3.2.10 Comparison of the Intervention and Control participants' group means on OOPT (overall and by section), vocabulary test, reading comprehension accuracysentence processing accuracy test, taken at T1 and T2 (N=89)

Measures		*F*-test statistics	*p* value
OOPT	Overall	$F_{time}(1,87)=1.878$	.174
		$F_{group}(1,87)=.030$	.863
		$F_{time*group}(1,87)=.008$	.928
	Use of English	$F_{time}(1,87)=1.188$	.279
		$F_{group}(1,87)=.111$	.740
		$F_{time*group}(1,87)=.159$	.691
	Listening	$F_{time}(1,87)=.379$	.540
		$F_{group}(1,87)=.017$	.897
		$F_{time*group}(1,87)=.453$	.503
Vocabulary		$F_{time}(1,87)=7.136$	.009
		$F_{group}(1,87)=1.244$	.268
		$F_{time*group}(1,87)=.301$	.585
Sentence processing accuracy		$F_{time}(1,87)=1.020$	.315
		$F_{group}(1,87)=3.171$	.078
		$F_{time*group}(1,87)=.468$	.496

Unlike the effects observed in the previous ANOVA, this time, there was no significant effect of time or group, or the interaction of time*group on all measures but vocabulary, where a significant effect of time was observed, suggesting that performance on vocabulary test at T2 differed significant from that at T1. However, as there was no significant effect of group or interaction on vocabulary scores, it is more likely that this significant effect of time was attributed to the practice effect, not the intervention.

In addition to processing accuracy, participants' processing speed was also measured and results were summarised in Table 3.2.12.

Table 3.2.11 The intervention and control participants' T1 and T2 time taken(measured in seconds)

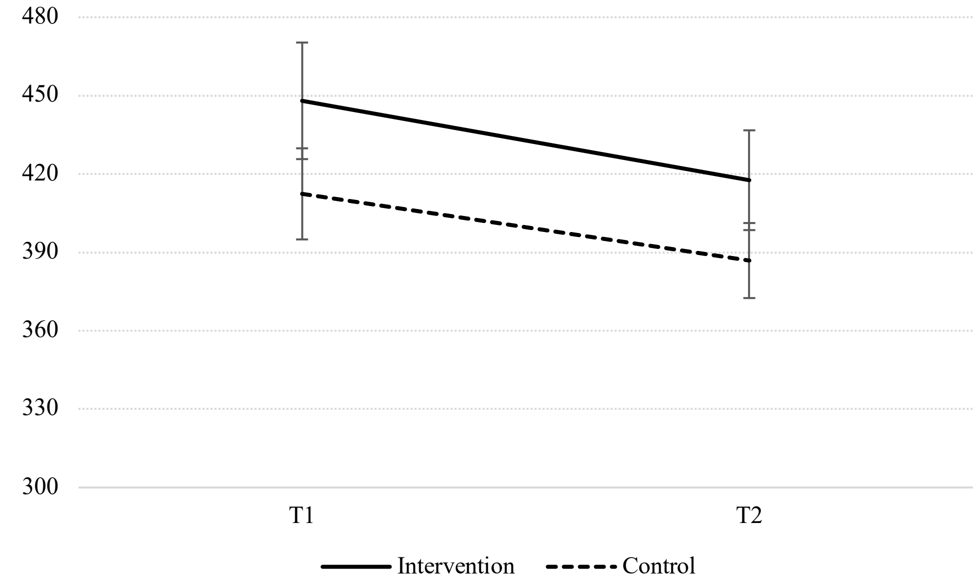
	Intervention (N=45)		Control (N=44)		
Sentence processing	Mean	SD	Mean	SD	Cohen's d
Speed T1	448.00	149.68	412.39	115.68	.27

Speed T2	417.62	128.07	386.86	95.25	.27

As shown, control participants were considerably faster than the intervention at both times, but t-test revealed this 0.27 SD difference was not statistically significant, T1 *t*(82.627)=-1.258, *p*>.05, T2 *t*(81.255)=-1.288, *p*>.05.

The change in processing speed from T1 to T2 for both groups were depicted in Figure 3.2.15.

Figure 3.2.13 The intervention and control participants' T1 and T2 time taken measured in seconds (N=89) (Error bars ± 1 standard error)



Different from previous figures, here, the downward trend represents to the lessening of time needed to read; thus it can be concluded that both groups needed less time and read considerably faster at T2 than they did at T1. However, as both lines remained parallel to each other and decreased with similar slopes, it seems that whether or not participants attended the IELTS preparation course did not have an significant effect on the improvement of reading processing speed as there was little interaction or the narrowing of gap between groups. It is more likely that participants read faster at T2 because they became familiar with the task at hand, i.e. practice effect of repeated measures.

Repeated measures ANOVA was also performed to examine the effects of time, group, time*group on participants' reading processing speed. Results, summarised in Table 3.2.14 revealed that there was indeed a significant effect of time, similar to what was found on vocabulary scores (Table 3.2.11), indicating that participants' reading speed at T2 differed significantly from T1. Nevertheless, there was no significant effect of group or time*group; thus the change in reading speed was less likely to be induced by the intervention, i.e. IELTS preparation courses, but rather the practice effect.

Table 3.2.12 Comparison of the Intervention and Control participants' group means processing speed test, taken at T1 and T2 (N=89)

Measures	*F*-test statistics	*p* value
Reading processing speed	$F_{time}(1,87)=28.514$	.000
	$F_{group}(1,87)=1.661$	.201
	$F_{time*group}(1,87)=.215$	.644

To sum up, analyses on participants' performance on general proficiency measures, i.e. OOPT, the vocabulary test and the sentence processing test, revealed that while IELTS test preparation asserted significant impact on the increasing of IELTS scores, there were little evidence suggesting that it also improved participants' overall general proficiency. Discussions relevant to findings of this quasi-experiment and its implication on IELTS validity are presented in the following sections.

3.3 Discussion

To achieve the primary goal of this quasi-experiment, i.e. empirically examine the hypothesis whether dedicated IELTS preparation course could boost candidate' IELTS score (overall and by skill) beyond the level of general proficiency, a pre-test/intervention/post-test research design was adopted, involving two groups of Chinese EFL learners (N=89) and a battery of linguistic measures. Participants were tested on all measures at two time points. For the intervention (N=45), they were tested once before the preparation course started and once after its completion; for the control (N=44), they were tested around the same timeframe of the intervention. In this section, results emerged through descriptive and statistical analysis are discussed at length in relation to previous literature on test preparation, test validity, and test washback.

3.3.1 To what extent does dedicated IELTS preparation course boost IELTS scores?

Starting with the IELTS overall score; comparison between the intervention participants' pre-test IELTS overall with their post-test IELTS overall showed that significant progresses were made. On average, a 0.60 band increase was achieved within 4 weeks; although this increase seems small in absolute terms, considering that IELTS only has 9 score bands, such short-term boost increase can be vital.

This finding closely resembles the score gains observed in Elder & O'Loughlin's study (2003) conducted in 4 different language learning centres in Australia and New Zealand (N=122). Elder & O'Loughlin's study reported that through 10-12 weeks' EAP study, participants improved their overall IELTS by an average of .60 (SD=.545), almost the same as the effects observed in this quasi-experiment.

Despite this similarity in research outcome, it should be highlighted that the effects observed in the present quasi-experiment was of greater magnitude on the following grounds. Firstly, as stated, Elder & O'Loughlin's study was conducted in two countries where English is used as a native language, which means that the observed in score gain could be attributed to other non-preparation related factors, such as daily exposure to the target language, i.e. English. By contrast, this quasi-experiment was set in Shanghai, China, where English is used as a foreign language and in general, little exposure to English was available to participants during the intervention apart from attending preparation courses. Therefore, there is good reason to attribute the effects observed on test scores to the dedicated learning in test preparation courses.

Secondly, the length of preparation course in Elder & O'Loughlin's study (10-12 weeks) was more than twice the length of the course examined in this quasi-experiment. Thus, if the length of courses were taken into consideration while interpreting the effects of preparation on score gains, it is probable that participants in this quasi-experiment could double their score gains through 10-12 week dedicated IELTS training. Together, it is reasonable to argue that although overall IELTS score gains observed in Elder & O'Loughlin's study and this was very similar, the test preparation courses under investigation here were indeed more effective in inducing score boosts.

On a module level, this quasi-experiment found that the intervention participants achieved the biggest score boost in listening, where a 0.77 mean increase was observed. This boost in listening score was much bigger compared to that reported by Elder & O'Loughlin (2003), an average of .40 (SD=.729). The second biggest score boost reported in this quasi-experiment occurred in IELTS reading, where the intervention participants achieved .73 band increase on average, in line with that reported by Elder & O'Loughlin (2003). This boost in listening scores could also be compared to Bagheri and Karami's (2014) test preparation study (N=40) set in EFL context (i.e. Iran). Bagheri and Karami's claimed that, by attending a 3-month IELTS preparation course, the intervention participants could boost their scores by 2.5 band from 5.5 to 8. Although score boosts were observed in both Bagheri and Karami's and this quasi-experiment, differences in the magnitude of score change are easily noticeable. The comparatively less score gains reported in this quasi-experiment could be attributed to the difference in the length of the intervention, 3 months in Bagheri and Karami's study and 1 month here. Understandably, with more time and longer learning period, participants could develop a deeper understanding of the skills, strategies and techniques taught during the course and internalise such knowledge for later use during the real exam. This echoes the aforementioned assumption that the length of test preparation may be positively related to the reported score gains and provides a good ground to assume that longer test preparation may results in more score gains.

Compared to the score gains observed in the two receptive modules, score gains on productive modules, though significant, were relatively less predominant. An average of .62 score boost was found in speaking, higher than that reported by Elder & O'Loughlin (2003), M=.50 (SD=.93). Speaking score gains reported in this quasi-experiment could be also compared to Issitt's study (2008) that also examined the effects of EAP and IELTS dedicated learning on improving participants' IELTS speaking in a UK university. Issitt claimed that through a total of 233 hours of EAP and IELTS learning, 7.5 of which was specifically targeted at IELTS speaking, participants managed to improve their speaking scores by .57, very similar to what was found in this quasi-experiment and slightly higher than that found by Elder & O'Loughlin (2003).

This difference in findings, although slight at first glance, becomes particularly interesting when the research contexts for all three studies are taken into consideration. As reviewed in the literature review chapter, both Elder & O'Loughlin (2003) and Issitt (2008) conducted their experiment in a

native speaking context, where, presumably, most participants "had" to communicate using English in their everyday life. Additionally, both Elder & O'Loughlin's (2003) and Issitt's (2008) studied involved a much longer test preparation or EAP learning period than the intervention in this quasi-experiment. Hence it would be more reasonable if more score gains were reported by Elder & O'Loughlin and Issitt in comparison to that here, not the other way around.
To account for this somewhat unexpected findings, the following hypotheses were put forward, drawing reference from Mickan and Motteram's (2008) test preparation classroom pedagogy observation research, Ma & Cheng's (2015) research on perceived value of TOEFL iBT preparation course, and Ma's (2014) qualitative research on international Chinese students' academic experience in a US university.

As discussed earlier in literature review, IELTS speaking consisted of 3 sections, the first focusing primarily on daily conversation with routine topics while the second and the third focusing mainly on academic topics (IELTS, 2018) . Most of these topics have been documented in an online IELTS speaking topic bank (*Ji Jing,* i.e. test taking experience in Chinese) created by Chinese IELTS candidates; when a candidate takes IELTS, he/she often posts what topics were asked during the exam for other candidates' reference (Ma, 2014). As typical IELTS speaking preparation involves a substantial amount of IELTS topic practising (Mickan & Motteram, 2008), it is likely that this IELTS speaking topic bank was incorporated as part of classroom practices in the present quasi-experiment. Because such online resources are written in Chinese, it is less likely that tutors involved in Elder & O'Loughlin's or Issitt's research, both situated in English speaking countries, were aware of, had access to and/or were unable to understand such resources. This could be the reason why participants involved in this study, a EFL context, managed to achieve more gains in speaking scores than those attending IELTS preparation in an English speaking context. In other words, the preparatory methods that were used in professional preparation centre were highly targeted and more effective in inducing score gains

The module that presented the most difficulty for short-time score boost for Chinese IELTS candidate involved in this study was IELTS writing, which is also the module where the majority of the limited test preparation research was focused on. In the present quasi-experiment, on average, IELTS preparation boosted the intervention participants' writing score by .46 band, lower than that reported by Elder & O'Loughlin(2003) (M=.55), which suggests that improvement of IELTS writing scores may require more time and effort from the candidates that a 4-week course does not seem to offer. However, this average .46 writing score gain was much higher when compared to Green's research (2007) on the effects of a 8-week IELTS preparation course. Situated in the UK, Green (2007) revealed that an average of .19 band writing score increase was achieved by 85 IELTS candidates. This discrepancy in research findings, particularly between this quasi-experiment and that of Green (2007) could be attributed to the make-up of the research sample. While the sample involved in this quasi-experiment was homogenous, the sample involved in Green's research was much more diverse, including 50 different nationalities. This diversity in cultural and first language means that IELTS preparation investigated in Green's study was taught using English, but here in this quasi-quasi-experiment, Chinese, first language of all participants was used. It is reasonable to speculate that the use of L1 would be more effective in conveying test taking relevant techniques, skills and strategies, thus leading to more score gains.

Score gains in writing from the present quasi-experiment can also be compared with that found by Brown (1998), who reported the biggest score gain in writing among existing IELTS test preparation literature. In Brown's research, an average of .94 score gain in writing was found, almost doubling the gains from this quasi-experiment. However, the generalisability of Brown's finding is worth questioning because only 9 participants were involved. Considering the small sample size, it is possible that Brown's study did not accurately portray the effects of IELTS preparation on IELTS writing scores; in other words, what was found in Brown's study might be the exception, not the norm. Therefore, reasons contributing to this discrepancy between Brown's and this quasi-experiment's findings are not further elaborated.

Further exploration of the data provided valuable insight regarding whether candidates' ability to achieve score gains can be predetermined by their existing language skills, an notion frequently referred to as the plateau effect in existing test preparation literature (Brown, 1998; Green, 2007). It was argued that candidates of comparatively higher proficiency may make less progresses and achieve less score gains through attending test preparation than candidates of comparatively lower proficiency. In a similar vein, Elder & O'Loughlin's noted "the language proficiency one had prior to the beginning of the course is the most constant indicator of how far one is likely to travel" (2003, p. 226). This notion was sustained in this quasi-experiment as explorative analysis found significant differences in score gains between those who began the IELTS preparation with higher proficiency and those who began with lower proficiency. By attending the same intervention course, lower proficiency candidates managed to achieve significantly more IELTS score gains than higher proficiency candidates, in each individual module and overall.

In a nutshell, data from the present quasi-experiment confirmed hypothesis that IELTS test preparation courses provided by a typical Chinese training school were effective in terms of significantly boosting Chinese candidates' IELTS scores within a short period of time, particularly for candidates of low language proficiency. Comparison between score gains observed in this quasi-experiment with that from previous literature indicates that preparation courses taught in Chinese, candidates' first language, might be able to induce more score boosts than courses taught in English, the target language being assessed by IELTS. This might seem counterintuitive at first, because exposure to the target language has been found an important factor influencing learners' language acquisition (Perani & Abutalebi, 2005). However, the nature and the goal of many IELTS test preparation courses are not geared towards improving candidates' acquisition of English but to help candidates achieve their desired scores. Under such circumstances, the use of L1 is likely to be more effective for the teaching of test taking knowledge than the use of English. Furthermore, it has been shown in this quasi-experiment and in existing literature that candidates who attended IELTS preparation were of relatively low English proficiency. This is reasonable because advanced EFL/ESL leaners do not necessarily need help from test preparation to achieve their desired scores. Given this low proficiency of preparation course attendees, the use of English as the language of instrument might also present challenges, creating additional obstacles in their path of IELTS preparation. This could have also attributed to the differences in score gains observed in this quasi-experiment and that reported in previous studies situated in English as a native language context.

On further note, given the adopted pedagogy of the intervention involved in this quasi-experiment, one could attributed the observed score gains to the explicit test taking strategy instruction (Type C preparation practice) which occupied a substantial portion of the test preparation course' pedagogy. However, it is also probable that the observed score gains in IELTS were relevant to the test format familiarisation provided through the test preparation courses (Type A preparation practice), which reduces the candidates' anxiety and thus making the scores a better representation of candidates' proficiency. As this quasi-experiment did not involve classroom-observation to determine which practices are accountable for the increase in score gains, one could only speculate the role Type A and C practice play. More research is in need to provide more conclusive evidence.

3.3.2 To what extent does dedicated IELTS preparation course improve candidates' general proficiency?

In addition to examining the effects of IELTS preparation on score boost, this quasi-experiment also looked at whether IELTS preparation could improve candidates' general proficiency using a battery of linguistic measures, including the Oxford Online Placement test, a vocabulary test, and a sentence processing accuracy and speed test.

As presented in the results section, contradictory to the steep increases observed in the intervention participants' overall and module-level IELTS scores, no significant changes in their OOPT were reported in this quasi-experiment both at the overall score level and at section level. Although participants of both groups did score slightly higher on the posttest than the pretest, given that control participants also achieve such gains, it is most likely to attribute such insignificant score improvements to the practice effect of the repeated measures research design, not the intervention. This absence of significant score improvement on OOPT suggested that the effect of IELTS test preparation did not extend to another linguistic measure. In other words, what was learnt through the IELTS preparation courses remained in the realm in taking IELTS tests. It can be further argued that what was learnt during the intervention course did not have significant positive impact on candidates' general proficiency.

Collaborative evidence was found through analysing participants' lexical knowledge, as measured by Spot-the-word, the vocabulary test. There was no significant change in how many words the intervention participants recognised out of the 60 test items, indicating that their lexical knowledge, an important indicator of general proficiency, was not significantly enhanced through this 4-week IELTS preparation course. This finding is particularly interesting because one would naturally assume, through learning IELTS, candidates would encounter unknown words, leading to subsequent vocabulary acquisition, especially in IELTS reading where they were asked to read 3 long academic passages. This assumption, however, was not supported by results from this quasi-experiment. Although counterintuitive at first, as the intervention involved in this quasi-experiment was test-oriented, with the majority of the learning contents focused on IELTS test taking techniques, it is probable that lexis were not taught or learnt explicitly, or even implicitly.

Similarly, the intervention participants involved in this quasi-experiment also failed to significantly improve their sentence processing accuracy as their posttest scores were not significantly different from their pretest scores; neither was there significant change in their sentence processing speed. This formed a sharp contrast with their significant increases in IELTS reading scores. Under normal circumstances, it only seems logical to associate increase in reading test scores with improvement in relevant linguistic knowledge (e.g. lexis) and reading abilities (e.g. processing accuracy and speed). However, this association may no longer be true in the context of test preparation as significant gains on IELTS reading did not correspond to improvement of lexical knowledge, processing accuracy or processing speed. As both processing accuracy and speed were considered good indicators of general proficiency (e.g. Schoonen et al., 2003; Segalowitz et. al., 1991), results on these two measured corroborated results from OOPT and affirmed that attending IELTS preparation did not significantly enhance candidates' proficiency to a level that corresponds with their IELTS score gains.

There is no literature with which this finding on the IELTS test preparation and candidates' general proficiency can be compared, as no previous research has tapped into this area. Although a handful of washback studies have noted that teaching and learning in test preparation courses differed from that in conventional language learning classrooms, none has employed quantitive measures to empirically measure such effects. In this way, this quasi-experiment had made valuable contribution on existing literature and has significant theoretical and pedagogical implications.

3.4 Conclusion

Using a pretest/intervention/posttest design, this quasi-experiment examined the effect of dedicated IELTS test preparation on candidates' IELTS scores (overall and by skill) and on their general proficiency. Data were collected from two groups of Chinese EFL learners (N=89, 45 intervention, 44 control) using a battery of tests, including a IELTS mock test and three other linguistic measures. Results indicated that the typical 4 week intensive IELTS preparation courses under investigation in this quasi-experiment significantly boosted candidates' IELTS scores but not their general proficiency, confirming the hypothesis that IELTS preparation could result in score gains that is not proportionate to improvement in general proficiency. This effect was even more predominant among candidates of low proficiency.

Findings from this quasi-experiment also enriches the discussion regarding the reliability and validity of IELTS as a widely used test of English proficiency. This quasi-experiment highlights that the reliability has been undermined by test preparation; hence the interpretation of IELTS scores and the extrapolation links between scores and proficiency should be cautioned. In addition to reliability, findings from this quasi-experiment also provides the implications on the construct and predictive validity of IELTS as a measure of English proficiency, which are presented in Chapter 5, along with findings from the second quasi-experiment.

Chapter 4 Quasi-experiment 2: the effects of repeated test taking and test preparation on Chinese candidates' IELTS scores, general proficiency, and their academic attainment at a UK university

This chapter presents the methodology and methods relevant to the second quasi-experiment of the present study. Details regarding the research design, instruments, testing procedure, hypothesis and data analyses are provided along with information on participants and ethical considerations of experimental research.

4.1 Methodology and methods

4.1.1 Research Design

Before proceeding to discuss the design of quasi-experiment 2, first, the link between this quasi-experiment and the previous should be clarified as it underpins the design and the methods used in second quasi-experiment. As put forward at the end of the literature review chapter, this study hypothesized that test preparation and repeated test taking could enable candidates to achieve IELTS that are higher than their general proficiency and thus contribute to the inconsistent predictive validity of IELTS. To examine the individual and combined effects of test preparation and repeated test taking on IELTS scores, general proficiency and subsequent academic attainment, the initial design of this study was to recruit participants at a test preparation centre in Shanghai, China and follow them throughout their IELTS preparation journey and throughout their academic study in an English-speaking higher education institution. Given that IELTS is mostly used as a language requirement for English-speaking higher education admission, it is reasonable to assume that candidates who signed up for IELTS test preparation courses were also preparing for studying abroad.

However, this design was found unfeasible because of the difficulty of retaining and following candidates through a long period of time as some candidates started preparing for IELTS years before their study abroad. Moreover, candidates recruited at the test preparation centre could proceed to study at different institutions in different countries, which not only added to the difficulty of sample retainment but also complicated the interpretation of their academic attainment (e.g. GPA achieved at an Australian university may not be comparable to degrees achieved in the UK). To collect sufficient data and to control for differences in operationalising academic attainment, this quasi-experiment adopted a retrospective and observational approach and recruited international Chinese students who had been accepted to study at a taught postgraduate level at a UK university. Their IELTS scores and their number of IELTS attempts were collected and their general proficiency was measured using two linguistic measures at the beginning of their study. It should be pointed out that as this quasi-experiment was set in a UK university, the sample involved was likely to be truncated because students needed to meet the IELTS requirement in order to be admitted into this university.

Given the high percentage of test preparation among IELTS candidates (e.g. Hawkey, 2006) and popularity of repeated test taking among Chinese candidates (e.g. Wilson, 1987), it is reasonable to assume that a fair number of the recruited international Chinese students had been involved in test preparation and repeated test taking, which meant that the findings from quasi-experiment 2 could corroborate the findings from quasi-experiment 1 to a wider population. Unlike quasi-experiment 1, which focused on one typical IELTS preparation programme offered at one particular training centre in one city of China, here, the overall IELTS preparation industry in China was looked at on a more general and holistic level as participants involved came from various parts of China, attending different test preparation programmes of various lengths and contents at different time points, provided by a variety of training organisations.

Furthermore, the combined effect of test preparation and repeated test taking on the predictive validity of IELTS was examined. To control for factors known to affect one's academic achievement, participants' non-verbal intelligence and their working memory were also measures in this experiment. Participants' weighted academic grades at the end of the taught component of their masters were collected as index for their academic attainment.

4.1.2 Participants

A total of 153 international Chinese students who learned English as a foreign language participated in this quasi-experiment on a voluntary basis, recruited through posters, emails and other social media platforms. All participants came from Mainland China, spoke Mandarin Chinese as their first language and received their previous education mainly in Mandarin Chinese. None of the participants had more than 3-month overseas experience in any English speaking country. Their average length of English as a Foreign language learning was 14.61 years, SD=2.21. The average age of participants was 23.32 years old, SD=1.63.

As for participants' previous IELTS preparation experience, over half of the research sample (N=87) had taken part in IELTS preparation course and 20 participants took part in more than one IELTS preparation course. The average length of their total IELTS preparation was around 6 weeks (Min=1 week, Max=24 weeks, M=5.84 weeks).

All participants were studying at a taught masters' postgraduate level at a UK University where this study took place. As a cross-disciplinary study, participants were from different disciplines. Out of 153, 82 participants were from the Department of Education, accounting for 54%. 34 were from the Department of Management (22%). 8 were from the Department of Economics and Related Studies (5%) and 7 (5%) were from the Department of Music. The rest were from the Department of Computer Science, Engineering, Language and Linguistic Science, Politics, Psychology, Women's Study and Sociology.

Given that the sample was of cross-disciplinary nature, the linguistic demand set for participants' postgraduate programme varied from one to another. To account for this variation, participants were categorised into linguistically less/more demanding group based on each programme's admission requirement for international students' admission. Programmes that required candidates to achieve an overall IELTS of 7 were labelled as linguistically more demanding (N=80) while the rest were categorised as linguistically less demanding (N=73). Overall band 7 was set as the criterion because candidates who achieved 7 are referred to as "*Good user*" of English who has "operational command of the language, although with occasional inaccuracies, inappropriacies and misunderstandings in some situations" (IELTS, 2018) which indicates that disciplines that set overall 7 as an admission requirement pose high demand on candidates' language skills, reasonable to be regarded as linguistically more demanding disciplines. Using this criterion, 70 participants were labelled as linguistically les demanding while 80 were labelled as linguistically more demanding.

For participants who did not manage to meet the IELTS demands, either overall or componential, they could also attend the presessional courses offered at the university. In this sample, a total of 55 participants were enrolled in the presessional programmes; 19 attended the 4-week presessional, 23 attended the 8-week presessional, and the remaining 13 took the 10-week presessional programmes. Independent t-test revealed that, in comparison with non-presessional participants, presessional participants had significantly lower IELTS scores (overall and by skills, overall $t(151)=5.762$, listening $t(151)=5.915$, reading $t(151)=2.830$, writing $t(151)=2.830$, speaking $t(151)=3.431$, $p<.01$ in all cases) and significantly lower general proficiency (as measured by C-test $t(151)= 4.573$, and DET $t(151)=5.461$, $p<.001$ in both cases), which in line the nature of presessional courses attendance. Because this quasi-experiment did not intend to examine the role of presessional courses, no further analyses concerning presessional attendance were included in the following sections of this thesis.

4.1.3 Research Instrument

Questionnaire

As previously stated, the first goal of this quasi-experiment was to establish a relationship between number of IELTS attempts, candidates' IELTS scores and their general proficiency, an online questionnaire was designed (see Appendix 5) , consisting of three main sections: 1) IELTS test taking and preparation history, including their first and final IELTS scores, the number of attempts made, and their IELTS test preparation length; 2) their demographics and their current academic programme; 3) their language use and English learning history. These questions were essential for building comprehensive understanding of the sample involved in this quasi-experiment and they provide crucial information for subsequent analyses on the effects of test preparation and repeated test taking.

Duolingo English Test (DET)

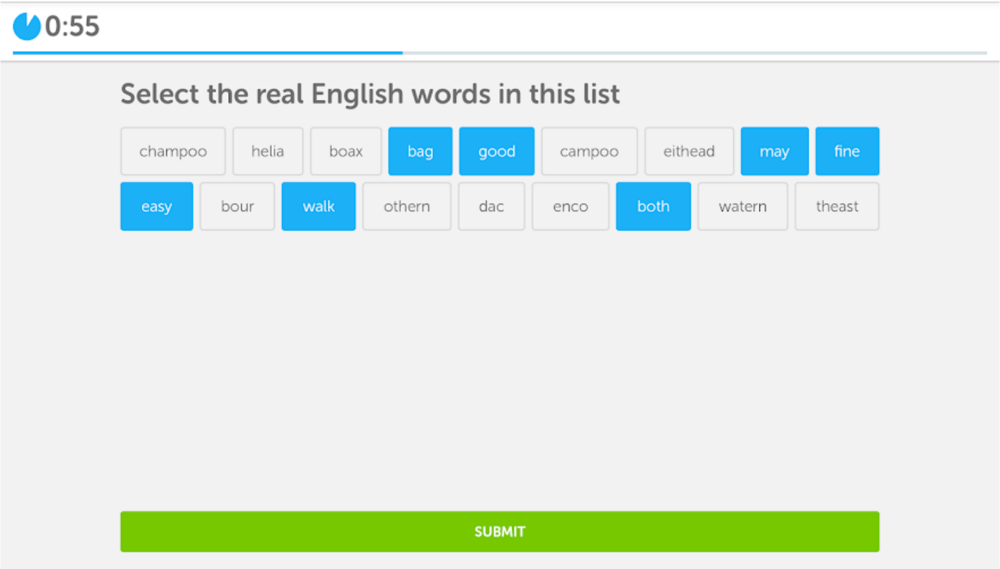
The DET test is a computer based, adaptive English test, developed by the Duolingo Language Learning programme (https://www.duolingo.com/). According to its homepage, the test is "designed to provide a precise and accurate assessment of real world language ability" and "measures real world usage" (Duolingo English Test, 2016). Test items that each candidate will see is decided by a computer algorithm, depending on how well she/he answered the previous items, similar to the Oxford Online Placement Test used in quasi-experiment 1. Because of the DET's adaptability, the number of test items one candidate takes might differ from another, so might the length of test (Wagner & Kunnan, 2015). According to the website, under normal circumstances, the DET usually takes about 30 minutes to complete (Duoling English Test, 2018).

Because of the recentness of DET, the amount of literature on the validity and the reliability of DET is limited; only 3 key research reports were found to have empirically examined on the reliability, validity, and the concordance of DET (Brenzel & Settles, 2017; Settles, 2017; Ye, 2014). To further illustrate the validity and reliability of DET as a competent measure of general proficiency, suitable for this quasi-experiment, analyses on specific DET task types are provided in the following.

DET consists of a written part and a verbal part (i.e. the Interview) and there are four types of tasks in the written part: listening tasks, speaking tasks, vocabulary tasks and completion tasks. Because it is an adaptive test, the number of tasks and test items taken and the sequence of which the following test items appear might vary from one candidate to another. Detailed illustrations of these four tasks were provided below, together with example test items and corresponding rationale on how these tasks relate to the measurement of general proficiency.

Vocabulary

Figure 4.1.1 DET Vocabulary Task by Brentel & Settles (2017)



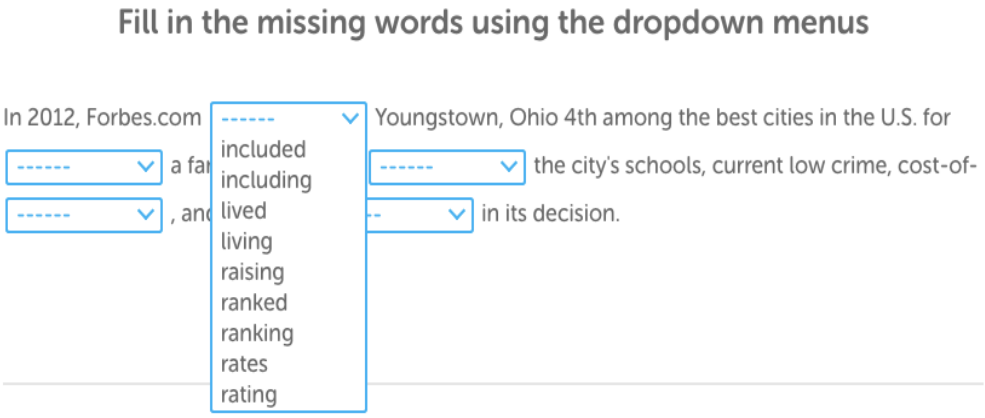
Select the real English words in this list

As shown in the Figure 4.1.1, DET vocabulary task asks candidates to discriminate the real English words (e.g. *bag, good*) from English-like pseudowords (e.g. *bour, dac*) within one minute (Brentel & Settles, 2017).

This vocabulary-identifying task taps into the similar cognitive process of (e.g., lexical and morphological activation) common in everyday reading, writing, and listening activities (Brentl & Settles, 2017) and assesses the size (or breath) of one's receptive/passive vocabulary knowledge, i.e. the words one recognises when he/she hears or sees the words. Vocabulary size, i.e. the number of words that language learners know at a particular level of language proficiency and an important component of one's vocabulary knowledge (Nation, 2001), has been found to relate closely with reading comprehension (Qian, 2002), the capability to obtain new details from texts among both native speakers and non-native speakers of English (Koda, 2005) and one's general proficiency (Laufer, 1992). By including a vocabulary session, one can say that DET as a test, has factored in lexical knowledge as an important indicator of proficiency, making it suitable for this quasi-experiment which looked at the effects of repeated test taking on general proficiency.

Completion

Figure 4.1.2 DET Completion Task by Ye (2014)



Fill in the missing words using the dropdown menus

In 2012, Forbes.com ------ Youngstown, Ohio 4th among the best cities in the U.S. for

------ a fa ------ the city's schools, current low crime, cost-of-

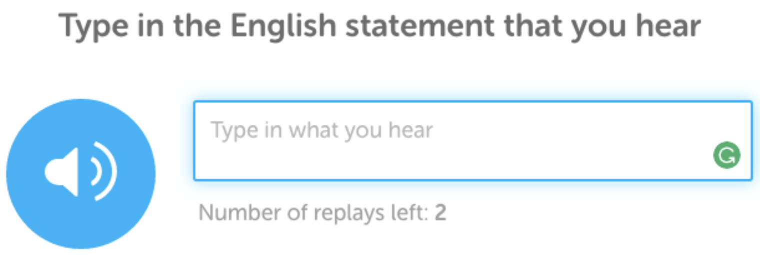
------ , an -- in its decision.

As illustrated, this task presents candidates with an incomplete short passage with missing words. Candidates need to fill in the missing words and complete this passage using the word options provided in the dropdown menus. Candidates have three minutes to complete this task.

This task resembles the rational cloze task (Wagner & Kunnan, 2015), a task developed from the traditional cloze task. The cloze test is a type of fill-in-the-blank test constructed through randomly deleting words from a prose passage, and replacing them with blanks. Cloze test is particularly useful for examining learners' micro skills such as reading comprehension (e.g. Chihara, Oller, Weaver & Chavez-Oller, 1977; Clarke, 1980; Bachman, 1982). As candidates need to utilise information outside the sentence or clause of the deleted item and often rely on the rest of the passage to fill in the blanks, cloze test is able to measure candidates' reading comprehension abilities. Cloze test is also seen a linguistically focused test because most of the time, candidates were able to deduce the answer from the immediate environment of the deleted item without utilizing information from the whole passage (e.g. Alderson, 1979; MacLean & d'Anglejan, 1986). High correlations between cloze test results and other standardised proficiency and placement tests outcomes, such as TOEFL (Darnell, 1968), UCLA E.S.A. Placement Examination (Oller, 1971), indicated that cloze test is a valid measure of proficiency, in line with DET's overall claim and suitable for the purpose of this quasi-experiment.

Listening

Figure 4.1.3 DET Listening Task by Ye (2014)



Type in the English statement that you hear

Type in what you hear

Number of replays left: 2

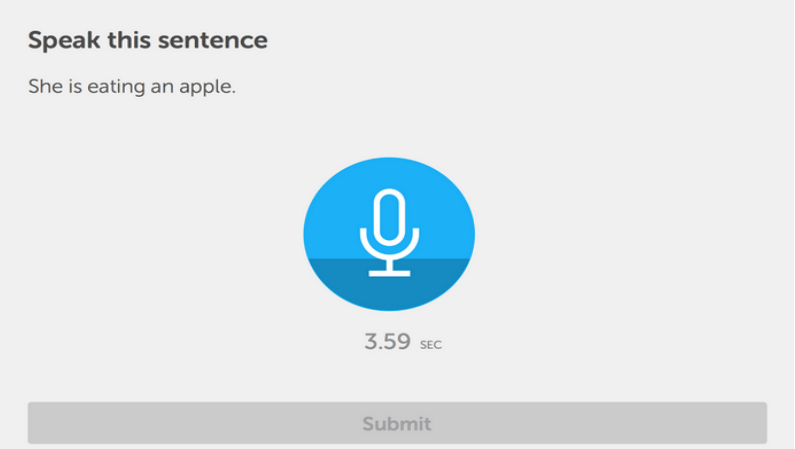
As shown, candidates are required to type in the English statement that they are going to hear into the box provided. The recording of this statement can be replayed twice, which means that all together candidates can listen to the statement for a total number of three times. Candidates have one minute to complete this task.

This task resembled an adapted version of the elicited imitation (EI) test. EI, traditionally designed to assess candidates' knowledge of grammatical items, has also been reported a good measure of interlanguage knowledge (Erlam, 2006). In the conventional EI tests, candidates hear a stimulus and are asked to repeat verbatim the stimulus exactly as it was read (Jensen & Vinther, 2003). Here, in DET Listening, candidates were presented with audio stimuli (i.e. recordings of English sentences) and instead of verbal repetition, candidates were to repeat the presented stimuli verbatim through typing down what they heard.

The rationale behind EI as a measure of interlanguage knowledge is that learners could only accurately reconstruct a stimulus that has been comprehended and parsed through their developing interlanguage system (Erlam, 2006). Research evidence indicates that by examining whether stimulus is repeated correctly, EI test could give an indication of candidates' implicit language knowledge (e.g. Erlam, 2006; Verhagen, 2011), listening comprehension (e.g. Jensen & Vinther, 2003), morphology (e.g. West, 2012) and oral proficiency (e.g. Naiman, 1974). Results from empirical studies suggested that, overall, EI test is an effective and efficient way to discriminate different levels of proficiency. Therefore the inclusion of an EI-based task in DET was considered in line with the test's overall aim.

Speaking

Figure 4.1.4 DET Speaking Task by Ye (2014)



Speak this sentence

She is eating an apple.

As shown, during the speaking task, candidates are asked to read the sentence presented on the screen out loud. They need to click on the micro icon to start the recording and click again when they finish recording. Candidates may record their sentence reading more than one time if they wish to. The algorism used for scoring for this speaking task is not clear. However, based on the design of this task, it is assumed that scores were based on common indicators of speaking proficiency, such as the accuracy of pronunciation, intonation and overall verbal fluency.

DET Reliability and concurrent validity

DET was reported to have high test-retest reliability with estimated reliability coefficient of .79 (Ye, 2016) and high internal reliability with a splithalf reliability coefficient of .96 (Settles, 2016). Regarding DET's concurrent validity, Ye's study also found that, among the 214 participants, performance on DET correlated substantially with overall TOEFL scores (r=.67, p<.05). In line with this, Ishikawa, Hall and Settles' study (2016) at an American university also reported that DET scores significantly correlated with on-campus faculty assessments of English ability for incoming international students.

To sum up, these above task analyses and research evidence indicate that DET is a good measure of general proficiency, thus suitable for inclusion in the present quasi-experiment.

C-test

Taking into consideration that DET was, at the time where this quasi-experiment took place, a relatively new and less widely used test of proficiency, to further ensure the accurate measurement of participants' general proficiency, C-test, a well-understood measure was incorporated.

C-test is a written test devised on the basis of the redundancy principle, which assumes that educated adult native English speakers are capable of drawing on the redundancy of language to restore the damaged messages using their language competence (Lei, 2008). Through examining candidates' ability to restore the damaged messages, consequently, C-test assesses candidates general language proficiency.

The validity, in particular the construct validity, of C-test has been a controversial topic for decades and a number of studies have been dedicated to such discussion. Although many have questioned what language construct C-test actually measures (e.g. Klein-Braley, 1984; Hastings, 2002), many have reported that C-test is indeed an indicator of general language competence (e.g. Dörnyei & Katona, 1992; Eckes & Grotjahn, 2006; Klein-Braley, 1994, 1997, Lei, 2008). For example, in Eckes & Grotjahn's study (2006), a total number of 843 participants took C-test as a part of a series of examinations in 16 different countries. Results supported the hypothesis that C-test was a valid measure of language proficiency. It was also found that high performance on C-test demanded candidates to show an integration of both skills and knowledge of the language being tested. Similar findings were reported in correlation studies examining the relationship between C-test and other standardized tests, such as the TOEFL (e.g. Dörnyei & Katona, 1992), the TOEIC (e.g. Daller & Phelan, 2006), the English Language Battery (e.g. Read & Chapelle, 2001). Because what language constructs DET aims to evaluate remains relatively uncertain, C-test was used to locally validate DET and act as another measure of participants' general English proficiency in this study.

C-test Format and design

Conventionally, a C-test consists of four to six authentic texts with a total of approximately 100 items. The texts are ordered according to the difficulty, from the lowest to the highest. The principle of *rule of two* is employed in the creating of a C-test, i.e. starting from the second word of the second sentence, the second half of every other word is deleted. If a word has an odd number of letters, the larger half is deleted. The first sentence of each text is usually left unchanged and names, numbers and one-letter word are undamaged. The deleted part of each word was indicated by a single underline of constant length (Klein-Braley & Raatz, 1984; Lei, 2008).

Below is the first passage of the C-test used in this quasi-experiment:

> Once upon a time, a child's bedroom had little more than a toy box, a bookshelf, and a few posters. Today i_______ looks mo_______ like miss_______ control a_________ Houston. Comp_______, mobile pho_______, televisions, DVD pla_______, game mach_______, and ot_______ 21st cen_______ toys fi_______ the ro_______, and of_______ make t_______ child's bed_______ the mo_______ expensive i_______ the ho_______. Britain's 8- to 16- year-olds have bedroom possessions worth an average of £3,300.

The C-test in this quasi-experiment was designed using five passages extracted from the New Headway English Course textbooks, published by Oxford University Press (Soars, Soars & Sayer, 2000, 2003). New Headways coursebook series were used here for the following reasons. First, they are well-recognised and widely used materials for English teaching and learning. Secondly, it incorporates authentic reading materials of different topics that are commonly accessible for the target participants of this experiment. Thirdly, it has different levels, which maps to the constitution of research sample's English proficiency in this study as some participants from linguistically less demanding programmes were assumed to have lower proficiency in comparison to those from linguistically more demanding programmes. Of these five passages, one was taken from the Pre-Intermediate level textbook, one was taken from the Intermediate level, and the rest were taken from the Upper-Intermediate level, arranged in increasing difficulty. The choice of passages also took into account the potential participants' language proficiency so as to ensure that the test was of sufficient difficulty to discriminate participants'' proficiency level. the A total number of 100 items were presented in this C-test.

C-test pilot

The designed C-test battery was piloted by 2 native-speakers of English and 5 Chinese English learners for feasibility and reliability. Piloting results indicated that the designed C-test battery was able to discriminate different level of English proficiency both between native-speakers of English and Chinese English learners and among Chinese English learners. Piloting also indicated that the average time needed to complete these five passages was approximately 20 minutes. When no time limit was given, participants tended to finish most of the items within 20 minutes and dwell on several items for an extra 5 minutes before turning the test paper in. Thus, a time limit of 20 minutes was set for the main data collection.

C-test reliability

To examine the internal consistency of the C-test devised for this quasi-experiment, Cronbach's alpha was calculated. As stated in the previous chapter, professionally developed high-stakes standardized tests should have internal consistency coefficients of at least .90 while lower-stakes standardized tests should have internal consistencies of at least .80 or .85 (Wells & Wollack, 2003). In this case, Cronbach's alpha provides a measure of the extent to which the items on the C-test provide consistent information with regard to students' mastery of the domain, i.e. their general English proficiency.

To calculate Cronbach's alpha, participants' C-test answers were entered into an excel spreadsheet item by item, using 1 indicating that the answer was correct and 0 incorrect and then exported into SPSS for statistically analysis. Reliability test results showed that for the 94 test items analysed[16], a relatively high degree of reliability was found. The average measure interclass correlation was .817 with a 95% confidence interval from .775 to .858, $F(152,14136)= 8.006, p<.001$, indicating that the C-test used for this quasi-experiment was internally reliable as an measure.

In addition to internal consistency, correlation analyses were also performed between C-test and DET so as to exam C-test's concurrent validity, i.e. to what extent scores on one test relate to the scores on another test of the same construct (section 2.5.5). A significant and moderate correlation was found, with a coefficient of .478 (p<.001). This further substantiated the validity of C-test used in this quasi-experiment.

WASI Matrix Reasoning and Digit Span

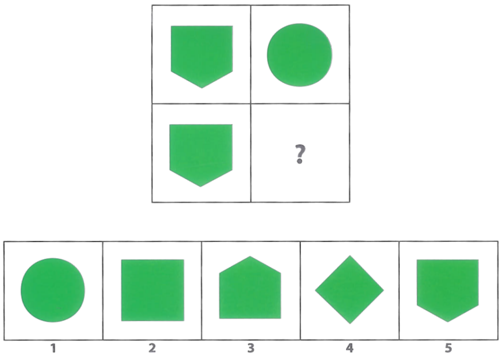
As the second goal of this quasi-experiment was to explore how repeated IELTS taking and test preparation contribute to IELTS's predictive power for international students'

[16] Although there were originally 100 items in C-test, 6 of which were not correctly answered by any of the 153 participants. Therefore only 94 items were analyzed for test reliability.

academic attainment at a UK university, in addition to the measurement of language proficiency, participants' non-verbal intelligence and their working memory were also taken into consideration, as literature indicates these two factors are closely related to one's academic performance.

Literature related to the prediction of academic performance showed that fluid intelligence, often referred to as reasoning and non-verbal problem-solving ability was one of the best predictors of academic performance (Cattell, 1978; Sternberg, Nokes, Geissler, Prince, Okatcha, Bundy & Grigorenko, 2001). Fluid intelligence is usually measured by tasks that have very little cultural content (Colom, Flores-Mendoza, Quiroga & Privado, 2005) and here The Wechsler Abbreviated Scale of Intelligence (WASI) II Matrix Reasoning subtest was used (Wechsler, 2011). WASI is an individually administrated, intelligence test which consists of four subtests: Vocabulary, Block Design, Similarities, and Matrix Reasoning. All subtests are shown to correlate strongly with general intellectual functioning (Saklofske, Caravan & Schwartz, 2000). In this quasi-experiment, WASI Matrix Reasoning subtest was utilised as a measure of fluid intelligence. An example of Matrix Reasoning is provided in the following Figure 4.1.5.

Figure 4.1.5 Example WASI Matrix Reasoning Item by (Wechsler, 2011)



In each WASI Matrix Reasoning test item, participants were shown a picture similar as above - with a question mark indicating a part missing. They were asked to complete the picture using the provided answers below within approximately 30 seconds. Scoring of WASI is provided in 4.1.6.

Research has also reported that working memory plays a crucial role when it comes to predicting academic performance as it is closely related to one's cognitive ability (Baddeley, 1992; Gignac & Weiss, 2015). Individual differences on working memory capacity had important consequences upon one's ability to acquire knowledge and new skills. Performance on working memory tasks was also found to predict reading achievement, phonological skills, math outcomes and computational skills (e.g. Colom, Abad, Rebello & Chun Shih, 2005; Colom, Flores-Mendoza, Quiroga, & Privado, 2005; Conway, Kane, Bunting, Hambrick, Wilhelm & Engle, 2005). Therefore, to control for the effect working memory might have on achieve academic success, forward Digit Span was incorporated in this quasi-experiment.

Digit Span is a task that measures working memory by asking candidates to recall a series of random single digits in the order with which they were read (Baddeley, 1992). As participants were all Chinese students who learnt English as a second language, two versions of Digit Span tasks were used, one in Chinese (participants' native language) and the other in English to measure their working memory in their both first and second languages. The sequence of digits varied in length, starting from three and gradually increased up to twelve for both languages. Scoring of digit span is provided in 4.1.6.

Academic attainment

In this quasi-experiment, participants' academic attainment was operationalised through their weighted average grades collected at the end of the taught component of their programme. Of the 153 participants, 1 failed to obtain sufficient credits to continue with her dissertation project; her term grades were included for later analyses nonetheless.

4.1.4 Ethics

Prior to the commencement of data collection, this quasi-experiment had gained approval from the Department of Education, University of York. An informed consent form that contained detailed information regarding the requirements and expectations in participating in this quasi-experiment was presented to each participant and I explicitly explained such information before the testing began. Participants were also made aware of the rewards they were entitled to as a token for their participation and they had the right to withdraw at any point of the data collection.

Anonymity was assured by replacing participants' real names with an assigned ID number. Participants were also made aware that their participation in this quasi-experiment would not in any means affect their academic study and their academic outcomes. The collected data would only be used for academic purposes and in future publication, only non-identifiable aggregated data would be presented. Before the testing began, participants were asked to sign the form, indicating that they were willing to comply with the requirements and took part in this study.

4.1.5 Procedure

Participants were recruited through emails and other social media platforms. It was specified that participation in this quasi-experiment required an electrical/paper proof of participants' IELTS results. Once they arrived at the quasi-experiment session and were confirmed to have met the criteria and brought along the needed proof of IELTS, they proceed to read and sign the informed consent form before the testing. They then proceed to answer a questionnaire, followed by Duoling English Test, WASI Matrix, Digit Span Chinese, Digit Span English and finally the C-test.

Questionnaire was administrated online using Qualtrics as the first step of data collection, which took around 3-5 minutes. When answering questions related to previous IELTS results, I checked the proof participants brought along (i.e. a scanned copy or a physical copy of their first and last IELTS reports) so as to make sure the validity of data collected. Once questionnaire was completed, participants proceeded to take the Duolingo English test using the provided computer with my presence, as to make sure there was no cheating during the test. Participants only completed the written part of DET as this was the only part contributing to their scores. DET took around 25-30 minutes.

This was followed by WASI Matrix reasoning, Digit Span Chinese and English, which took approximately 20 minutes in total. For WASI Matrix, rules and practice items were explained and shown to participants before the real testing. They were allowed approximately 30 seconds to answer each Matrix item. If they encountered an item that they did not know the answer of, they could answer "I do not know" and move onto the next item. When a participant failed to answer three continuous items correctly, the Matrix test was stopped and we moved onto the Digit Span.

Digit Span was carried out on a computer, using DMDX with pre-recorded soundtracks of the digits. Participants were first tested with Digit Span in their first language (i.e. Chinese) then in English. Two versions of Digit Span were used during this quasi-experiment and test papers were arranged in a counterbalance order. For example, participant 01 would be tested with Version A in Digit Span Chinese and Version B in Digit Span English, while participant 02 would be tested with Version B in Digit Span Chinese and Version A in Digit Span English.

For both versions of Chinese and English digit span tasks, a sequence of digits were presented with 650 ms between digits. After a digit was played, participants would see "*Please repeat now"* displayed on the computer screen and proceed to repeat the digit they just heard. To continue onto the next task item, they were instructed to press the spacebar. The length of sequence of digits increased by one every three sequences. Both Chinese and English version started with three digits and ended with a maximum of twelve digits in both English and Chinese.

All testing was conducted on a one-on-one base once, at the beginning of participants' postgraduate studies between week 1 and 7. All these afore-presented tests were completed in one session in a control office without having any breaks in-between. Under normal conditions, one testing session took approximately 65 to 70 minutes. All related instructions were given using English during the quasi-experiment. No feedback was given during or after the quasi-experiment.

4.1.6 Scoring

DET

As a computer based English test, DET scores are generated automatically and immediately on a scale of 100% at the end of the test. For the vocabulary, listening, and completion task, scoring was based on whether correct answers were entered. For the speaking task, candidates' responses were scored by the computer using a proprietary algorithm (Wagner & Kunnan, 2015). Results were sent to the candidate within 48 hours of the completion of DET. Although exact method/algorithm adopted by DET to generate such scores reminded unknown, the following table is provided as a reference for test-takers regarding how scores could be interpreted.

Table 4.1.1 Interpretation of DET Scores (Duolingo English Test, 2018)

Score range	Level	Abilities	Example
0% - 16%	Beginner	Can only understand very basic words or phrases in the language.	Can read public road signs, ask for basic directions, and fill out a simple form.
17% - 35%	Elementary	Can deal with simple, straightforward information and express themselves in familiar contexts.	Can have a short, coherent dialogue on topics of interest, but not extended conversations.
36% - 55%	Intermediate	Can understand the main points of concrete speech or writing on routine matters such as work and school. Can handle most situations that would come up while traveling where the language is spoken. Can describe experiences, ambitions, opinions, and plans, although with some awkwardness or hesitation.	Can open a bank account, if the procedure is fairly straightforward.
56% - 71%	Advanced	Can fulfil most communication goals, even on unfamiliar topics. Can understand the main ideas of both concrete and abstract writing, and interact with native speakers fairly painlessly.	Can show visitors around and lead a detailed guided tour of a place.
72% - 92%	Proficient	Can understand a variety of demanding texts and conversations, also grasping implicit or figurative meaning that is hidden. Can use language flexibly and effectively for most social, academic, and professional purposes.	Can get and hold onto his or her turn to speak at a party, or respond to interrogating questions with little or no hesitation.
93% - 100%	Expert	Can understand virtually anything heard or read, even intellectually demanding material such as an academic lecture or a book on philosophy. Can use the language fluently and spontaneously in a way that can even be more advanced than an average native speaker.	Can scan long texts for relevant information, and differentiate finer shades of meaning in complex social and professional situations.

WASI Matrix

WASI Matrix was scored manually by myself using the provided answer sheets. 1 point was given when participants answered the test item correctly. All participants in this study started from item 4 in WASI Matrix because if they could answer item 4 correctly, it was assumed that they could answer item 1,2,3 correctly as well. The minimum and maximum possible scores for WASI Matrix were therefore, 0 and 30.

Digit Span Chinese/English

Both Chinese and English digit spans were scored manually by the myself using the provided answer sheets. 1 point was given when participants recalled the number in the exact same order as was read to them. A final score was generated by averaging the length of last three digits participants recalled correctly in the exact same order as were read to them. The maximum score for both Chinese and English digit span was an average of 12.

C-test

C-test was scored and double-checked by myself. 1 point was given when participant completed the mutilated word correctly. An overall score was generated by adding up all the points participants achieved out of 100 test items. There were 101 items in this C-test but there were two items in the same clause sharing the same answer; therefore, 1 point was given to participants who answered this item (either or both) correctly.

4.1.7 Research questions and Hypothesis

RQ1: To what extent does repeated IELTS taking affect Chinese candidates' IELTS scores and their general proficiency?

For this question, it was hypothesised that through repeatedly taking IELTS, candidates would achieve scores that were beyond their general proficiency. This hypothesis was based on the assumed commonality in motives behind repeated test taking and engagement in test preparation courses, i.e. to obtain better scores. As was found in quasi-experiment 1, attending test preparation allowed candidates to achieve significant score gains within a short period of time, without corresponding improvement in their proficiency. Given this, it is logical to hypothesize that repeated test taking might have similar effect as well, i.e. boosting scores onto a level that is beyond candidates' proficiency, most likely through the cumulation of testwiseness.

Alternatively, it is also possible that through repeatedly taking the test, candidates developed higher proficiency given the time and effort they put in preparing for the test, similar to the argument made in Wilson's TOEFL repeater study (1987). In this case, one would expect participants who repeated IELTS more frequently to have higher proficiency.

RQ2: To what extent does IELTS predict international Chinese students' academic attainment at a UK university?

This research question re-examines the predictive validity of IELTS on academic attainment. It was hypothesised that overall, IELTS would be a significant predictor, contributing unique variance to the predictive model, based on evidence from Trenkic & Warmington (2018), Daller & Phalen (2013) and other IELTS predictive research reviewed in Chapter 2, section 2.5.7. Further, because this quasi-experiment involved an cross-disciplinary research sample, with participants studying programmes that assert varying linguistic demand, it was assumed that the predictive validity of IELTS would differ depending on discipline, based on research evidence from Feast (2002), Yen & Kuzma (2009) and Trenkic & Warmington (2018). IELTS was expected to account for more variance in academic attainment for those studying linguistically more demanding disciplines, e.g. language and related science, and account for less for those studying linguistically less demanding disciplines, e.g. computer engineering.

RQ3: Have test preparation and repeated test taking eroded the validity of IELTS as a predictor for academic attainment?

If the hypothesise for RQ1 is confirmed, together with effects observed from quasi-experiment 1 (i.e. test preparation resulted in significant IELTS score gains without corresponding improvement in candidates' proficiency), the hypothesised answer for this question would be that inclusion of attendance at test preparation and number of IELTS attempts would improve the predictive validity of IELTS based on the following ground. First, were hypothesised effects of test preparation and repeated test taking to be confirmed among this population, it means that the validity of IELTS can be undermined because construct irrelevant variance, i.e. testwiseness, has been introduced into the IELTS score. Consequently, this would also interfere with the predictive validity of IELTS, rendering IELTS a less powerful predictor for academic attainment. Thus, by adding the attendance at test preparation and number of IELTS attempts back into the predictive model, is hypothesized to compensate the loss of IELTS's predictive validity, thereby improving the overall model fit.

4.1.8 Analysis

Both SPSS 24 and R Version 3.5.0 were employed for statistical analysis. Preliminary normality check using Shapiro Wilk revealed that apart from IELTS scores, the rest of the data were normally distributed; hence, a combination of parametric and non-parametric tests were used for later analysis.

Descriptive statistics on all linguistic measures were presented first to show the IELTS and general proficiency level of the research sample involved in this study. Following this, prior to answer the research questions set out for this quasi-experiment, a transitional analysis on the effect of test preparation, candidates' IELTS scores and general proficiency was presented to corroborate findings from this quasi-experiment from that of the previous quasi-experiment. For this purpose, descriptive statistics were provided first, followed by Mann-Whitney U test to examine the difference between test preparation and non-test preparation candidates' general proficiency and their IELTS scores (overall and by skills).

To answer the RQ1 (i.e. to what extent does repeated IELTS taking affect Chinese candidates' IELTS scores and their general proficiency?), descriptive statistics were presented and correlation analyses were performed to examine whether there existed a relationship between number of IELTS attempts and participants' general proficiency.

To answer RQ2 (i.e. to what extent does IELTS predict international Chinese students' academic attainment at a UK university?), regression analysis was performed for the whole sample, using IELTS as the main predictor, controlling for non-verbal intelligence and working memory. Following this, participants were regrouped into *the linguistically more demanding* or *the linguistically less demanding* according to their discipline and the same regression analysis was performed again.

To answer RQ3 (i.e. to what extent does the number of IELTS attempts add to the predictive validity of IELTS?), a regression model using IELTS as the prime predictor, number of attempts as the additional predictor was built and tested for significance.

Descriptive statistics and results from the analyses were reported in the next section in relation to the proposed research questions, starting with confirmatory analyses that examined the afore-stated hypotheses, followed by exploratory analyses that revealed some unexpected findings.

4.2 Results and Analysis

Data collected for quasi-experiment 2 were analyses in this section and results from descriptive and statistical analyses were presented. To better depict the trends observed in this quasi-experiment, data were visualised using expectancy charts as used in Cho and Bridgeman's study (2012). The organisation of this section follows the sequence of the proposed research questions; descriptive statistics were presented first to provide an overview of the data, setting the basis for subsequent statistical analyses.

4.2.1 Descriptive statistics for all linguistic measures

Table 4.2.1 summarised the descriptive statistics on all linguistic measures used in quasi-experiment 2.

Table 4.2.1 Descriptive statistics of participants' IELTS scores (overall and by skill) and their performance on Duolingo and C-test

Measures		N	Min	Max	Mean	SD	Median
Number of IELTS attempts		153	1	8	3.17	1.49	3
IELTS Overall	First	153	4.50	8.00	6.27	0.61	6.50
	Final	153	5.50	8.00	6.73	0.48	7.00
IELTS Listening	First	152[17]	4.50	8.50	6.60	0.93	6.50
	Final	153	5.50	9.00	7.10	0.90	7.00
IELTS Reading	First	152	5.00	9.00	6.80	.87	7.00
	Final	153	5.00	9.00	7.24	0.83	7.00
IELTS Writing	First	152	4.00	7.00	5.74	.47	5.75
	Final	153	5.00	7.50	6.05	0.36	6.00
IELTS Speaking	First	152	4.00	7.00	5.82	.57	6.00
	Final	153	5.00	7.50	6.11	0.46	6.00
Duolingo English test (DET)		153	10.00	93.00	56.31	15.51	58.00
C-test		153	15.00	68.00	44.79	10.75	46.00
WASI		153	8.00	23.00	16.95	3.11	17.00
Digit span Chinese		153	4.33	12.00	8.80	1.32	8.67
Digit span English		153	3.67	9.00	5.47	.89	5.33

[17] Only 152 first attempt IELTS module scores were included because one participant made the first attempt so long ago that he could only remember the overall score.

As shown, for the research sample involved in this quasi-experiment, an average of IELTS 6.73 (SD=.48, Median=7.00) was achieved through 3.17 attempts (SD=1.49, Median=3). Their average scores on DET was at the Intermediate level (Table 4.1.4), which according to DET score interpretation corresponds to an overall of IELTS 6, considerably lower than the reported IELTS mean of this sample. Meanwhile, their average correct answer on C-test was less than half of the total items (100).

For participants who sat more than one IELTS (N=135), comparison between participants' first IELTS and final IELTS scores revealed that considerable progresses had been made, as Wilcoxon Signed-ranks test found significant differences between first and final IELTS scores, overall Z=-8.649, listening Z=-6.455, reading Z=-6.036, writing Z=-6.544, speaking Z=-6.071, p<.000 at all cases. These differences in first attempt final attempt IELTS scores could be attributed to many possibilities, such as IELTS test preparation, as 57% of the participants were indeed engaged in test preparation programmes (more detailed analysis regarding this is provided in section 4.2.2). Moreover, given the differences in time when the first IELTS was made, e.g. some were made as early as 2009 and some were made as recent as 2016[18], and the different psychological status when the first attempt was made, e.g. some candidates might not be fully committed in their first attempt and used it as an opportunity to "test the water", first IELTS scores may be unreliable. In comparison, final IELTS scores were considered accurate representation of participants' IELTS level at the time of data collection. Firstly, because the timing of final attempts was, in general, closer to when the quasi-experiment took place. Secondly, because participants needed IELTS results as part of their visa application and university admission, if multiple attempts were made, it is more likely that they were very committed to the test and made the most effort in the final attempt. Thus, from hereafter, the analyses were performed using participants' final IELTS results.

The average final IELTS overall achieved, which, according to IELTS round up rule (section 3.1.8), could be reported as an overall of band 7.0. This indicated that the sample involved in this quasi-experiment was at an advanced level, compared to the general Chinese IELTS candidature and their average overall score (overall 5.73 for 2017 Chinese IELTS candidature, section 2.8.4). Meanwhile, final module scores

[18] Data collection for experiment 2 took place in autumn, 2016.

revealed that these participants' receptive skills (listening and reading) were considerably higher than their productive skills (writing and speaking), which resembled the trends observed in the first quasi-experiment.

Regarding the number of IELTS attempts made by this sample, all participants took at least one IELTS tests before this study took place. As shown in the Table 4.2.2, only about one tenth of the sample took IELTS once (hereafter referred to as one-timers); the majority of this sample was IELTS repeaters. Over half of the participants took IELTS 2 to 4 times, before they either achieved the scores they desired or gave up, resorting to alternative means for admission, such as presessional courses. A very small portion of this sample made no less than 6 IELTS attempts, showcasing that the wide spread of repeated IELTS taking behaviour in this sample.

Table 4.2.2 Number of IELTS attempts by participants (N=153)

Number of IELTS attempts	Number of participants		IELTS overall score gains achieved			
	N	%	Mean	SD	Median	Range
1	18	11.8	-	-	-	-
2	37	24.2	.41	.50	.5	-.5-1.5
3	42	27.5	.39	.39	.5	-.5-1.0
4	31	20.3	.66	.40	.5	0-2.0
5	12	7.8	.67	.65	.75	0-1.5
6	9	5.9	.78	.26	1.0	.5-1.0
7	3	2.0	.50	.00	.5	.5-.5
8	1	.7	.50	-	-	-

With regard to score gains achieved with each increase in IELTS attempts, from Table 4.2.2, one can see that mean score gains achieved by four/five/six timers were considerably larger in comparison to that achieved by two/three timers, which could indicate that with more attempts, one can achieve bigger score IELTS score gains. This should be treated with caution as attempt groups in this sample were not evenly distributed and there was large disparity in terms of time spent in between attempts.

4.2.2 Effect of test preparation on Chinese candidates' IELTS scores (overall and by skill) and their general proficiency

This section corroborates the findings from quasi-experiment 1 regarding the effect of test preparation on IELTS scores and general proficiency and furthers findings from quasi-experiment 1 to a more general level using data collected from participants who attended different IELTS preparation courses of different length and contents, at different locations and different time points. In addition, it extends the findings to a different population: international Chinese students enrolled in UK HEIs, who had probably completed their test preparation courses months ago.

Summarised in Table 4.2.3 are participants' final IELTS overall and their scores on DET and C-test, two proficiency measures that these participants were unfamiliar with.

Table 4.2.3 Comparison of English proficiency scores between participants who attended IELTS-preparation programmes and those who did not (N=153)

	Test preparation (N=87)				Non-test preparation (N=66)				*r*
	Mean	SD	Median	Range	Mean	SD	Median	Range	
Final IELTS overall	6.66	.47	7.0	5.5-7.0	6.81	.48	7.0	5.5-8.0	.15
DET	53.90	15.05	52.0	20-91	59.49	15.64	59.5	10-93	.18
C-test	42.33	10.16	43.0	15-61	48.03	10.73	48.0	25-68	.26

On the one hand, both groups achieved very similar final IELTS scores (overall and by skills); on the other hand, there were clear discrepancies in participants' general proficiency measured through DET and C-test between group, with the test-preparation group lagging behind with an effect size of .18 and .26 respectively. Mann-Whitney U and independent samples t-test were performed to examine these between-group differences and results indicated that the differences in final IELTS overall was not significant, U=2391.50, p>.05, but the differences in DET and C-tests were, DET, t(151)=2.24, p<.05, C-test: t(151)=3.35, p<.001.

These results confirmed findings from quasi-experiment 1 and suggested that for international students admitted with very similar IELTS scores, their general proficiency varied. Those who attended test preparation programmes were likely to have lower general proficiency than those who did not attend such programmes even thought their final IELTS scores were at the same level. This further confirmed that through test preparation, candidates could achieve IELTS scores that were beyond their actual general proficiency.

4.2.3 To what extent does repeated IELTS taking affect Chinese candidates' IELTS scores and their general proficiency?

Given that Duolingo and C-test were used in this quasi-experiment to measure a very similar construct, i.e. general proficiency, and that the measures were moderately strongly correlated (*r*=.458, *p*=.000), a composite proficiency score was created by summing up *z* scores from both tests. Thus, hereafter in analyses, only the composite score was used.

To examine how number of IELTS attempts relate to participants' general proficiency and their IELTS scores, Table 4.2.4 was produced summarising the general proficiency score and final IELTS scores at each attempt level.

Table 4.2.4 Participants' general proficiency (composite score) and their final IELTS scores in relation to their number of IELTS attempts (N=153)

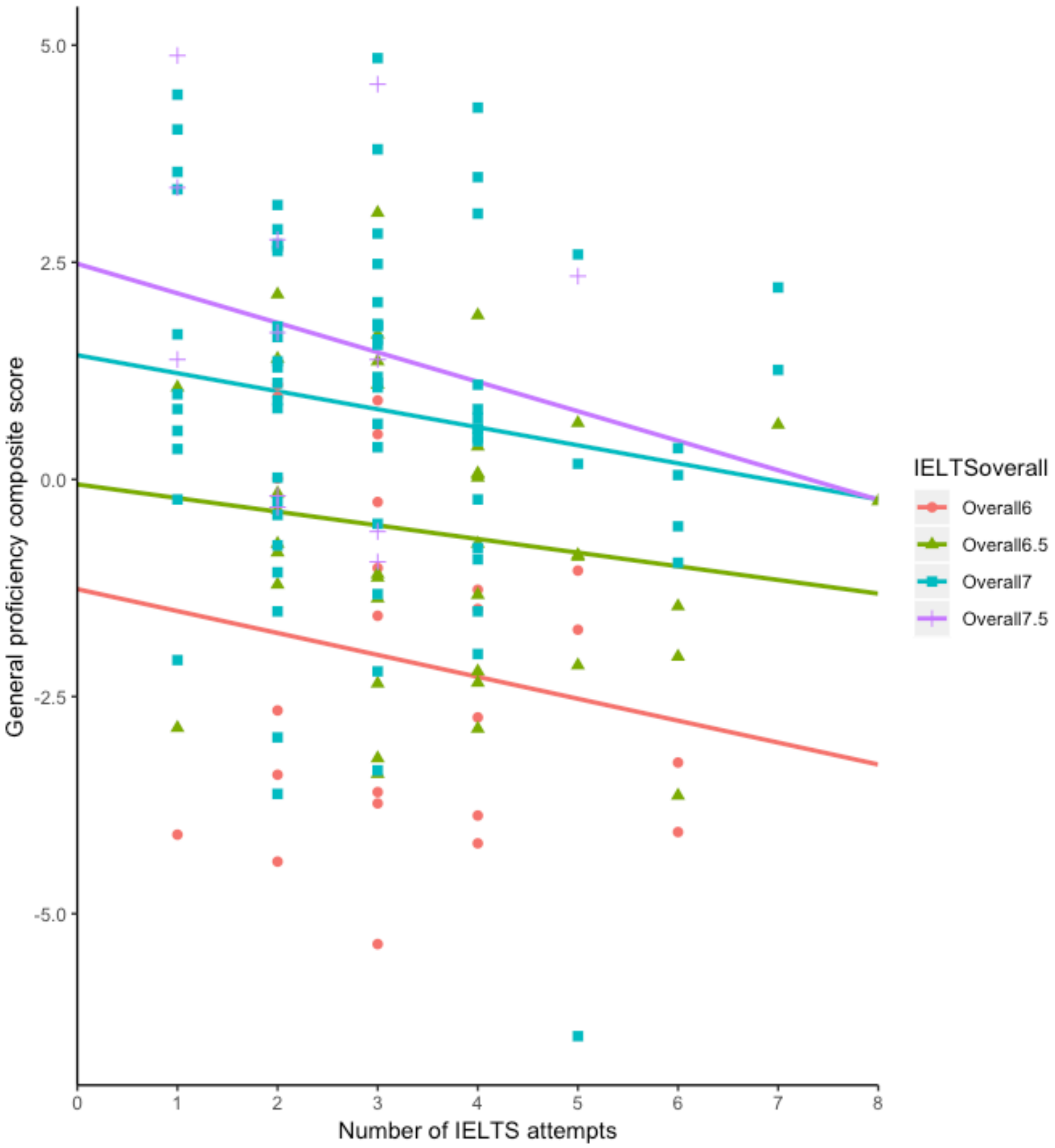
Number of IELTS attempts	N	General proficiency		Final IELTS									
				overall		listening		reading		writing		speaking	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
1	18	1.40	2.58	7.03	.44	7.53	.78	7.58	.81	6.17	.34	6.33	.49
2	37	.22	1.94	6.80	.45	7.24	.81	7.38	.88	6.08	.34	6.14	.52
3	42	.09	2.37	6.70	.49	7.01	1.00	7.21	.87	6.05	.35	6.15	.42
4	31	-.43	2.01	6.63	.43	7.03	.86	7.11	.73	6.00	.32	5.95	.35
5	12	-1.00	2.39	6.46	.62	6.79	.86	6.83	.78	5.96	.54	5.97	.46
6	9	-1.73	1.63	6.61	.42	6.56	.92	7.17	.79	5.94	.30	6.06	.46
7	3	1.37	.79	6.83	.29	7.67	.76	7.17	.76	6.33	.58	6.33	.29
8	1	-.25	-	6.50	-	7.00	-	6.50	-	6.00	-	5.50	-

As shown, IELTS one-timers had the highest proficiency composite score and the highest final IELTS overall. With each unit of increase in attempts, there was a corresponding decrease in participants' general proficiency composite score and IELTS scores, indicating that more attempts were made by participants who had relatively lower proficiency and less attempts were made by those with higher proficiency. This assumption was confirmed by correlation analysis, which found a significant yet negative relationship between number of attempts and general proficiency composite score, *r*=-.249, *p*<.01.

To understand whether repeated test taking could indeed boost IELTS scores onto a level beyond their general proficiency, participants were grouped according to their final IELTS overall[19], and their number of attempts were plotted against their general proficiency composite score in Figure 4.2.1.

[19] Participants who achieved a final IELTS overall 5.5 (N=4) and a final IELTS overall 8 (N=1) were excluded from Figure 4.2.1 due to the small number.

Figure 4.2.1 Relationship between number of IELTS attempts and general proficiency at difference final IELTS overall level (N=149)



In Figure 4.2.1, the pink round dots are participants who scored IELTS 6 on their final attempt; green triangles are those who scored 6.5; blue square are those who scored 7 and purple cross are those who scored 7.5. Each linear line is the regression line depicting the relationship between number of IELTS attempts and general proficiency composite score, coloured according to participants final IELTS scores.

As shown, at each final IELTS overall level (i.e. final IELTS overall being the same), with each increase in attempts, there was decrease in general proficiency. However, this is not to say the repeated test taking lowered participants' general proficiency. Instead, this indicates that participants who arrived the same final IELTS overall scores, their proficiency varied. Those who attempted IELTS more times to achieve a certain score were of a lower proficiency, while for those who attempted IELTS less but achieved the same score, their proficiency was likely to be higher. Together, figure 4.2.1 and results from correlation analyses lent support to the hypothesis that repeated IELTS taking could boost candidates' overall scores onto a level that's beyond their general proficiency.

Finally, to understand how IELTS test preparation and repeated test taking relate to participants' general proficiency in addition to what has been accounted for by final IELTS overall scores, a linear hierarchical regression analysis was fitted using composite score as the outcome variable, final IELTS overall, attendance of test preparation and number of IELTS attempts as the predictor variables, entered in the presented order. The results of this regression model were summarised in Table 4.2.5.

Table 4.2.5 Regression model using final IELTS overall score, attendance of IELTS-preparation programmes and number of test attempts to predict the English proficiency of participants on entry to university (N=153)

Model	B	Coefficients SE	ß	t	*p*	R^2	ΔR²	ΔR² sig.
1						.305	.305	.000
Constant	-17.71	2.18		-8.12	.000			
Final IELTS overall	2.64	0.32	.55	8.14	.000			
2						.329	.024	.022
Constant	-16.56	2.21		-7.50	.000			
Final IELTS overall	2.52	0.32	.53	7.18	.000			
Attendance of test preparation	-0.71	0.31	-.16	-2.31	.022			
3						.344	.015	.067
Constant	-15.04	2.34		-6.42	.000			
Final IELTS overall	2.39	0.33	.50	7.25	.000			
Attendance of test preparation	-0.64	0.31	-.14	-2.09	.038			
Number of attempts	-0.20	0.11	-.13	-1.85	.067			

As shown, in the first model, final IELTS score alone accounted 31% of the total variance in general proficiency. With the additional inclusion of attendance of test preparation, the model improved significantly by 2.4%. In the final model, inclusion of number of IELTS attempts improved the model further by 1.5% and the final model accounted for 34% of the total variance in general proficiency *F*(3,149)=26.00, *p*=.000. This, together with findings from quasi-experiment 1 and Figure 4.2.1, confirmed test preparation could significantly boost IELTS scores onto a level that is beyond candidates' general proficiency. With regards to the number of attempts, the model shows that the effect was less strong than the effect of test-preparation, and only marginally significant. For a regression model to reliably detect a small effect size, a much larger sample would be needed.

Meanwhile, it should also be noted that the inclusion of test preparation and number of attempts as the additional predictors only improved the model fit slightly. It is important to bare this in mind as this relates to the following analyses of using IELTS as a predictor for academic attainment.

4.2.4 To what extent does IELTS predict international Chinese students' academic attainment at a UK university?

The predictive validity of IELTS on international Chinese students' academic attainment (indexed through weighted average grades) was re-examined in this quasi-experiment, controlling for non-verbal fluid intelligence (measured through WASI) and working memory (measured through digit span Chinese/English).

On average, this group of international Chinese students achieved 59.92 in their first two terms assignments. The minimum grade needed for these students to progress to their dissertation project was set at 50 and 7% of the sample failed to meet this demand, meaning that they either had to resubmit assignments, resit exams, or to terminate their study. Note that the grades collected for analyses in this quasi-experiment were participants' *initial* academic grades, not their resit/resubmit grades.

Summarised in Table 4.2.6 are participants' weighted average academic grades, classified into Fail, Pass, Merit, and Distinction using the MA/MSc degree classification set by the university where this quasi-experiment was conducted, along with their final IELTS overall scores.

Table 4.2.6 Participants' weighted average academic grades classification IELTS (N=153)

	Grades					IELTS	
	N	Min	Max	M	SD	M	SD
Fail (<50)	11	31.17	49.75	45.24	5.29	6.50	.50
Pass (50-60)	66	50.00	59.80	55.63	2.92	6.74	.45
Merit (60-70)	61	60.20	68.92	63.96	2.52	6.74	.47
Distinction (>70)	15	70.00	81.27	73.08	3.17	6.77	.63

Intuitively, from Table 4.2.6, it could be seen that although participants varied considerably in terms of their academic attainment, their IELTS did not vary as much, indicating that as a whole, participants' ability to achieve academic success might not be closely related to their language proficiency. In line with this intuitive assumption, correlation analysis did not reveal a significant relationship between grades and IELTS, r=.111, p>.05. However, participants who failed to achieve enough grades to continue with their dissertation (i.e. 50) did have lower IELTS than those who succeeded.

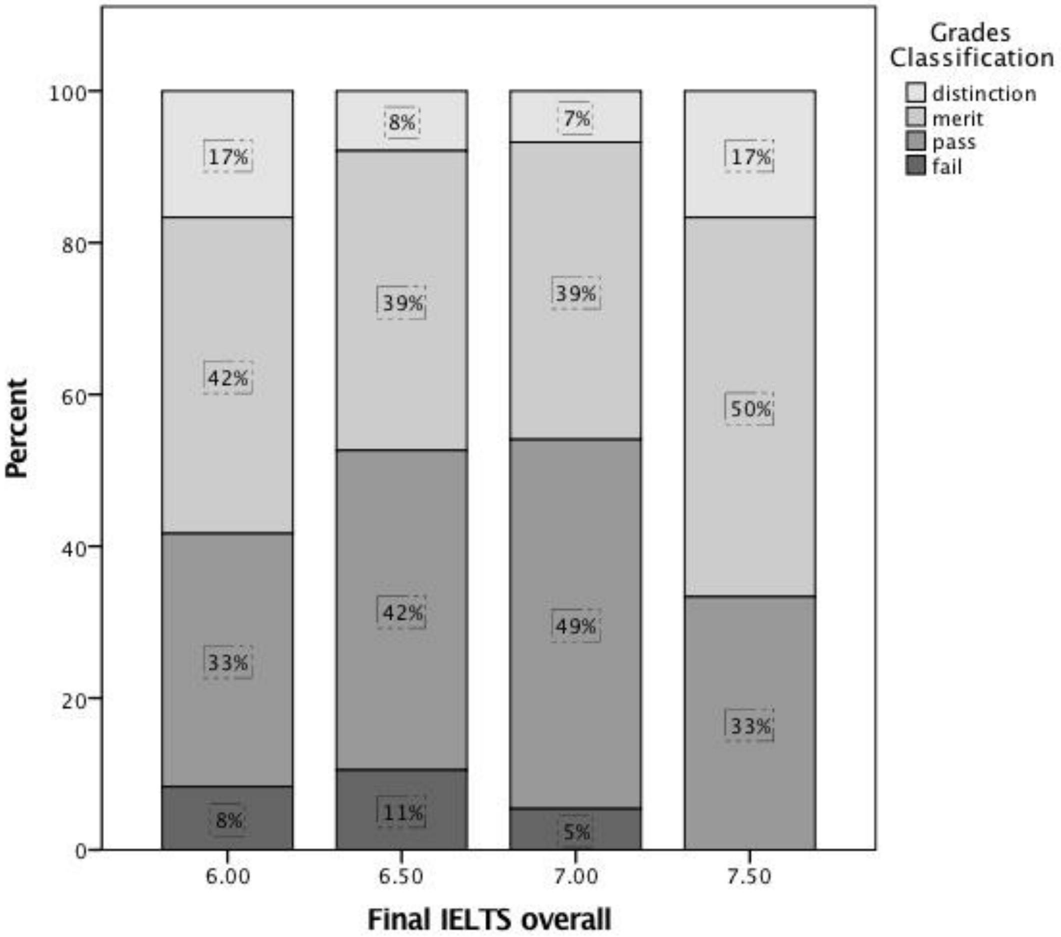
To test whether IELTS could predict academic attainment, linear regression was performed using weighted academic grades as the outcome variable and IELTS as the main predictor, controlling for non-verbal intelligence (i.e. WASI) and working memory (i.e. a composite score of digit span Chinese and digit span English). The focus of Table 4.2.7 was on model 3 where IELTS was included as a main predictor. The final model (model 3), summarised in Table 4.2.7, showed that IELTS was not a significant predictor overall, *F*(3,149)=1.486, *p*>.05.

Table 4.2.7 Regression model using IELTS overall to predict the academic grades for all participant, controlling for non-verbal intelligence and working memory (N=153)

Model	Unstandardized B	Coefficients SE	β	t	*p*	R^2	ΔR²	ΔR² sig.
1						.003	.003	.506
Constant	57.705	3.371		17.118	.000			
WASI	.130	.196	.054	.667	.506			
2						.018	.015	.133
Constant	57.662	3.357		17.177	.000			
WASI	.133	.195	.055	.682	.496			
Digit span	.537	.356	.122	1.510	.133			
3						.029	.011	.192
Constant	46.447	9.188		5.055	.000			
WASI	.138	.194	.057	.709	.480			
Digit span	.513	.356	.117	1.443	.151			
IELTS	1.655	1.263	.106	1.311	.192			

This regression outcome is in line with visualisation of trends observed in expectancy chart, as used in Cho and Bridgeman's study (2012), as shown in Figure 4.2.2.

Figure 4.2.2 Percentage of participants achieving fail, pass, merit, distinction grades by their final IELTS overall (N=148[20])



As the expectancy chart illustrates, with the increase of IELTS overall scores, there was no corresponding growth in the proportion of participants achieving grades at distinction/merit level, neither was there clear shrinkage in the proportion of participants achieving pass/fail grades. To be specific, the percentage of distinction participants was exactly the same in the 6.0 IELTS group and 7.5 IELTS group. Similarly, the percentage of participants achieving merits was the same when their IELTS was 6.5 and when their IELTS was 7.0. Together, visualisation of the data and the regression analyses outcome suggest that when all participants were looked at together, there was no clear relationship between proficiency as measured by IELTS and academic attainment.

[20] Participants who achieved a final IELTS overall of 5.5 (N=4) and 8 (N=1) were excluded from this chart as they were considered outliers

There are many factors that may have resulted in this model being insignificant. To start with, the data collected for this quasi-experiment were truncated because the context of research was set at a UK university with international Chinese students who had already met the language requirement. In other words, this sample did not include participants of low proficiency. In addition, this sample involved very few high IELTS scores with only one participant scoring as high as 8 and none above. This lack of high IELTS scores could be related to the high cost of taking IELTS, and/or the urgent need of IELTS scores for visa application purposes. In other words, candidates stop taking the test once the needed the scores were achieved, even although the score might underrepresent their true level of proficiency. Together, it is possible that this truncated nature of the sample failed to detect the relationship between proficiency and academic attainment.

Alternatively, it is also possible that correlation and regression failed to detect any significant relationship because this relationship was "masked" by the noise in the data due to the research sample's crossdisciplinary nature. As discussed in earlier literature review chapter (section 2.5.5-2.5.6), the relationship between proficiency measured by standardised tests (e.g. IELTS, TOEFL) could be affected by the linguistic demand of disciplines participants were enrolled in. For participants who were studying linguistically more demanding disciplines, a stronger relationship between proficiency and attainment may exist while for those studying linguistically less demanding disciplines, a weaker relationship between proficiency and attainment could be expected. Not taking into account this variation in linguistic demand of participants' field of discipline has been highlighted as one of the limitation of many previous predictive studies and hence, here, to further explore the predictive validity of IELTS, participants in the research sample were categorised into two subgroups.

Table 4.2.8 Weighted average academic grades and final IELTS overall of participants from less and more linguistically demanding disciplines (N=153)

	More demanding (N=80)					Less demanding (N=73)				
	Weighted grades			IELTS overall		Weighted grades			IELTS overall	
	N	M	SD	M	SD	N	M	SD	M	SD
Fail (<50)	6	45.93	2.81	6.75	.27	5	44.41	7.66	6.20	.52
Pass (50-60)	46	55.95	3.11	6.95	.26	20	54.91	2.35	6.28	.44
Merit (60-70)	24	63.65	2.60	7.00	.21	37	64.16	2.48	6.57	.52
Distinction (>70)	4	71.90	2.04	7.25	.50	11	73.51	3.47	6.59	.58

The regrouped data in Table 4.2.8 presents a somewhat different picture regarding the relationship between academic grades and IELTS. Firstly, it appears that both groups' academic grades were, in fact, closely associated with their IELTS scores, more demanding *r*=.377, *p*<.01, less demanding *r*=.235, *p*<.05. Although the correlation was slightly stronger in more demanding than that in less demanding, this difference in correlation strength was not significant, *z*=.95, *p*>.05. This indicated that for both groups, international Chinese students with higher IELTS scores were indeed associated with higher academic attainment.

The significant group-level correlation seems contradictory to the overall insignificant correlation as reported earlier, but this phenomenon could be explained by the Simpon's paradox (Wagner, 1982). Simpon's paradox is a statistical paradox that produce opposite results depending on how the data were divided. For example, in 1973, UC Berkely was sued for gender bias because the graduate school admission figures showed that the proportion of female applicants accepted (35% out of 4321) was lower than that of male (44% out of 8442). However, once the data were broken down and examined department by department, departmental admission data showed that out of the six departments examined, four were significantly biased in favour of female, which contradicts the overall admission rates in favour of male. This could be explained by that females tended to apply to the departments that were higher to get into while males tended to apply to the departments that were easier to get into (Fenton, Neil & Constatinou, 2015; Wagner, 1982). This statistical paradox highlights the importance of controlling for the variable that may affect the analytical outcome, which in this quasi-experiment means that the effect of disciplinary linguistic demand differences should be accounted for.

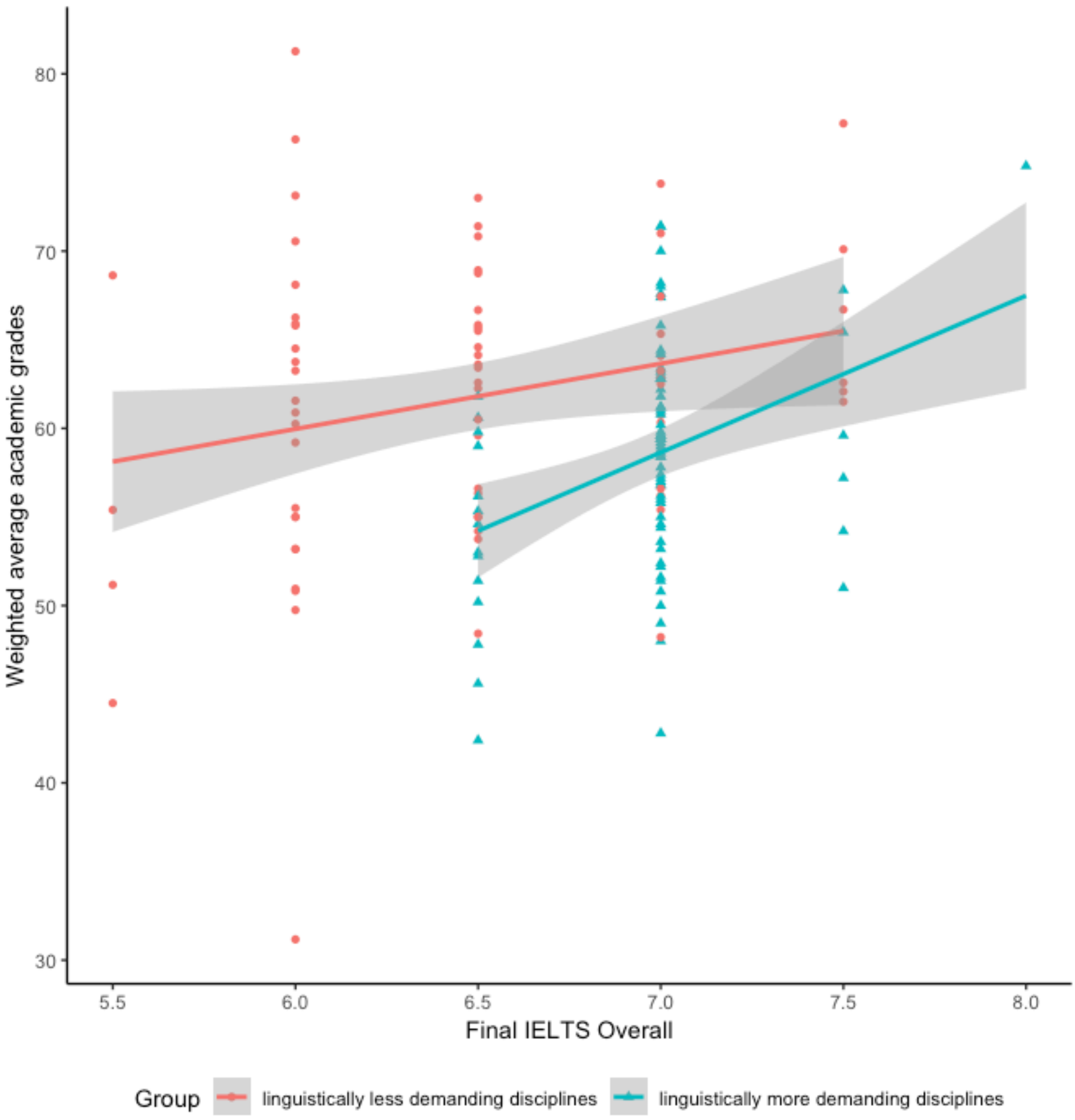
To account for the differences in disciplinary linguistic demand, participants from linguistically less demanding disciplines were labelled with 0 (N=73) while those from linguistically more demanding disciplines were labelled with 1 (N=80) and a linear regression model was fitted using disciplinary linguistic demand as a categorical predictor. Model 4 summarised in Table 4.2.9 showed that IELTS only became predictive when the difference of linguistic demand was accounted for. On the note of the interaction between IELTS and disciplinary linguistic demand difference, model 5 summarised in Table 4.2.9 showed that there was no significant interaction. This, coupled with visualisation from Figure 4.2.3, one could see that there was no significant interaction between the two lines (blue line representing the linguistically more demanding model, red line representing the linguistically less demanding model).

Table 4.2.9 Regression model using IELTS overall and disciplinary linguistic demand to predict the academic grades for participants, controlling for non-verbal intelligence and working memory (N=153)

Model	Unstandardized B	Coefficients SE	β	t	*p*	R^2	ΔR²	ΔR² sig.
1						.003	.003	.506
Constant	57.705	3.371		17.118	.000			
WASI	.130	.196	.054	.667	.506			
2						.018	.015	.133
Constant	57.662	3.357		17.177	.000			
WASI	.133	.195	.055	.682	.496			
Digit span	.537	.356	.122	1.510	.133			
3						.029	.011	.192
Constant	46.447	9.188		5.055	. 000			
WASI	.138	.194	.057	.709	.480			
Digit span	.513	.356	.117	1.443	.151			
IELTS	1.655	1.263	.106	1.311	.192			
4						.136	.107	.000
Constant	29.699	9.540		3.113	.002			
WASI	.072	.185	.030	.388	.698			
Digit span	.499	.337	.113	1.481	.141			
IELTS	4.758	1.399	.305	3.402	.001			
Linguistic demand	-5.732	1.341	-.383	-4.274	.000			
5						.147	.011	.166
Constant	36.701	10.761		3.411	.001			
WASI	.061	.184	.025	.329	.743			
Digit span	.447	.338	.102	1.323	.188			
IELTS	3.704	1.587	.237	2.334	.021			

Linguistic demand	-37.440	22.843	-2.504	-1.639	.103
Interaction	4.629	3.329	2.159	1.390	.166

Figure 4.2.3 Visualisation of the relationship between IELTS and academic grades by group (N=153)



Thus, taking this into consideration, two separate linear regression models were fitted to further examine the relationship between IELTS and academic grades, one for linguistically more demanding disciplines (summarised in Table 4.2.9), and the other for linguistically less demanding disciplines (summarised in Table 4.2.10).

Table 4.2.10 Regression model using IELTS overall to predict the academic grades for participants of linguistically more demanding disciplines, controlling for non-verbal intelligence and working memory (N=80)

Model	Unstandardized B	Coefficients SE	β	t	p	R^2	ΔR²	ΔR² sig.
1						.002	.002	.707
Constant	56.730	4.237		13.389	.000			
WASI	.094	.250	.043	.377	.707			
2						.050	.048	.052
Constant	56.072	4.174	.060	13.435	.000			
WASI	.132	.246	.220	.537	.593			
Digit span	.829	.419		1.978	.052			
3						.167	.117	.002
Constant	.332	17.515		.019	.985			
WASI	.065	.233	.029	.278	.782			
Digit span	.601	.401	.160	1.499	.138			
IELTS	8.169	2.501	.348	3.266	.002			

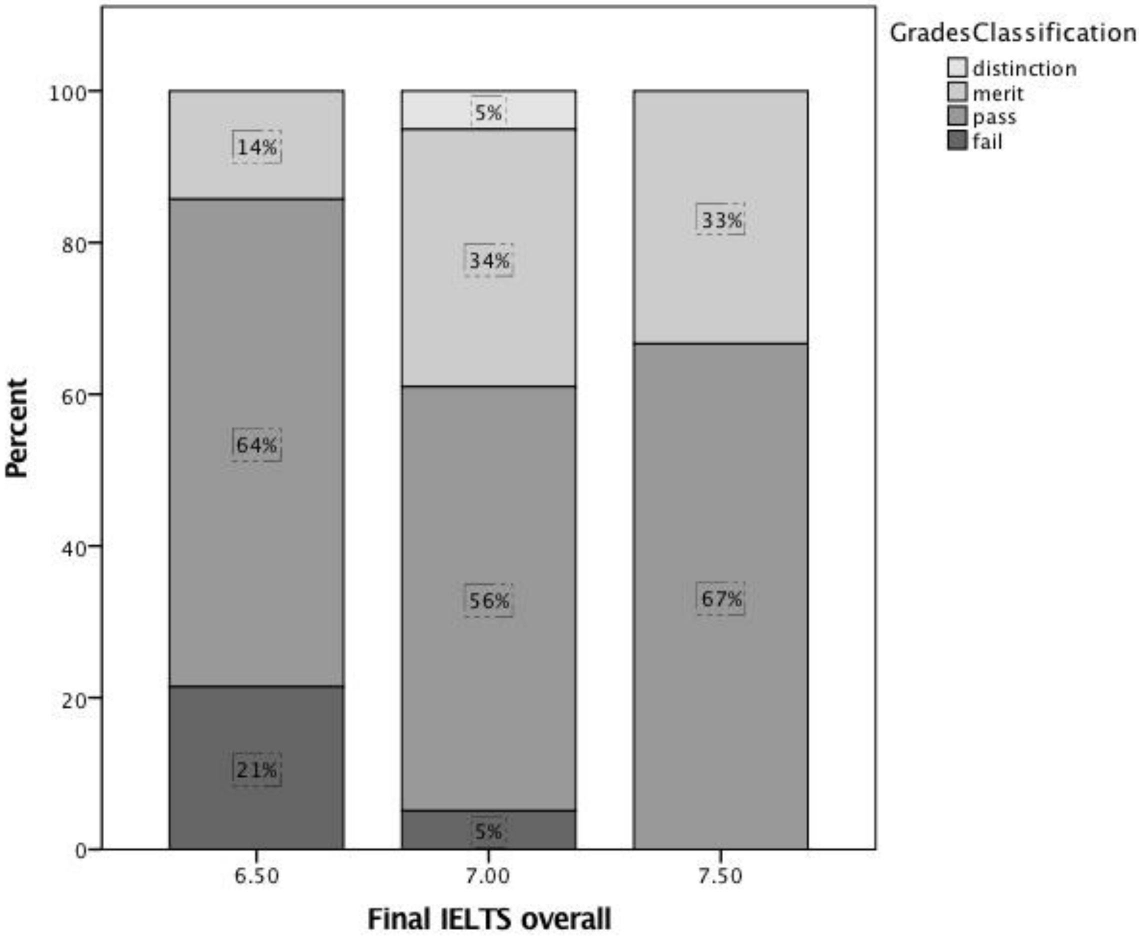
Table 4.2.11 Regression model using IELTS overall to predict the academic grades for participants of linguistically less demanding disciplines, controlling for non-verbal intelligence and working memory (N=73)

Model	Unstandardized B	Coefficients SE	β	t	p	R^2	ΔR²	ΔR² sig.
1						.001	.001	.784
Constant	60.296	5.129		11.756	.000			
WASI	.080	.293	.033	.275	.784			
2						.004	.003	.641
Constant	60.448	5.168		11.697	.000			
WASI	.072	.295	.029	.244	.808			
Digit span	.268	.572	.056	.469	.641			
3						.060	.055	.048
Constant	36.599	12.873		2.843	.006			
WASI	.069	.288	.028	.239	.812			
Digit span	.284	.560	.059	.507	.614			
IELTS	3.697	1.835	.235	2.015	.048			

Results showed that for the linguistically more demanding group, IELTS was indeed a good predictor (model 3 in Table 4.2.10), explaining unique variance in academic attainment, even after working memory and intelligence had been accounted for. Together, the final model accounted for 12% of the total variance in academic grades achieved by linguistically more demanding participants, *F*(3,76)=5.078, *p*<.01. With 1 bandcore increase in IELTS, weighted average grades increases by 2.5. In comparison, the predictive power of the final model with the same predictors was considerably weaker and not significant for linguistically less demanding participants (model 3 in Table 4.2.11), *F*(3,69)=1.455, *p*>.05, although the inclusion of IELTS did improved the model significantly. Overall, results showed that IELTS was indeed a good predictor for this sample's academic attainment when disciplinary linguistic demand was taken into consideration. This finding offers insights into the inconsistency regarding IELTS's predictive validity disputed in existing literature; future discussion on this is provided in section 4.3.

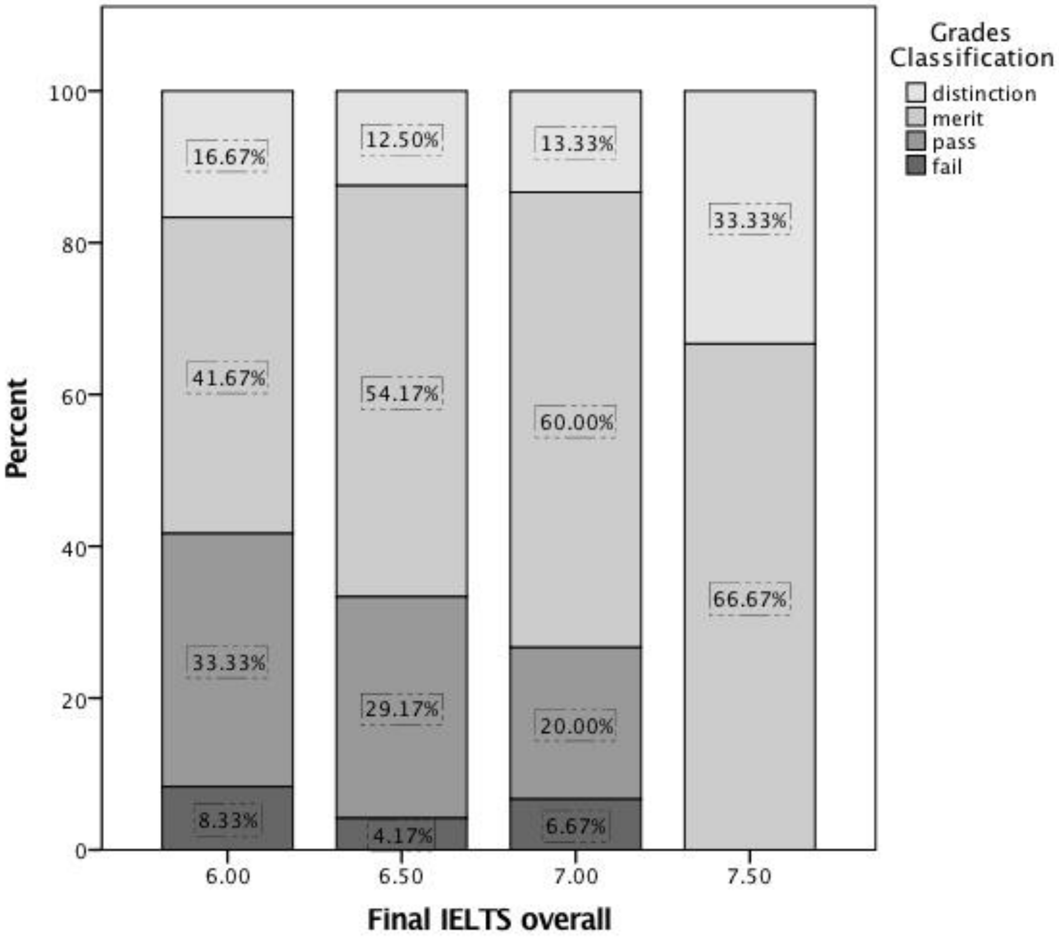
To visualise this relationship between proficiency measured by IELTS and academic grades, after accounting for the differences in disciplinary linguistic demand, expectancy charts (Figure 4.2.4 and 4.2.5) are produced.

Figure 4.2.4 Percentage of linguistically more demanding participants achieving fail, pass, merit, distinction grades by their final IELTS overall (N=79[21])



[21] Participant who achieved a final IELTS overall of 8 (N=1) was excluded from this chart as s/he was regarded as a outlier

Figure 4.2.5 Percentage of linguistically less demanding participants achieving fail, pass, merit, distinction grades by their final IELTS overall (N=69[22])



Contrast to the unclear trends in Figure 4.2.2, obvious patterns were shown in Figure 4.2.3 and 4.2.4. For the linguistically more demanding group, with the increase in IELTS, there was clear shrinking in percentage of participants achieving fail grades and rapid increase in the percentage of participants achieving pass/merit grades. Similarly, for the linguistically less demanding group, with the increase in IELTS, the proportion of fail participants decreased while the proportion of distinction and merit participants expanded considerably.

Together, regression results and the visualisation of data revealed that, when the variation of disciplinary linguistic differences was accounted for, proficiency as measured by IELTS had a clear and significant relationship with participants' academic attainment. This means that although participants were accepted onto their postgraduate

[22] Participants who achieved a final IELTS overall of 5.5 (N=4) were excluded from this chart as they were regarded as outliers

programmes on the premises that they had obtained the needed IELTS, their subsequent academic performance was, to various degrees, hindered by their proficiency.

4.2.5 Have test preparation and repeated test taking eroded the validity of IELTS as a predictor for academic attainment?

On the basis of the effects of test preparation and repeated test taking reported in quasi-experiment 1 and 2, this session explores whether the predictive validity of IELTS has been affected by the observed effects using correlation and regression analysis.

Summarised in Table 4.2.11 and 4.2.12 are participants' academic grades, final IELTS scores in relation to their number of IELTS attempts.

Table 4.2.12 Participants' weighted average academic grades, IELTS in relation to their attendance at test preparation courses

	N	Average academic grades		IELTS	
		Mean	SD	Mean	SD
Test preparation	66	59.08	7.89	6.66	.47
No test preparation	87	60.56	7.16	6.81	.48

As shown, there did not seem to be a large difference in grades between participants who attended IELTS preparation and those who did not. This was confirmed by independent t-test results, *t*(151)=-1.208, *p*>.05, which suggests that attendance at test preparation did not affect academic attainment to a statistically significant degree.

Table 4.2.13 Participants' weighted average academic grades, IELTS in relation to number of IELTS attempts made

	Academic grades					IELTS	
Attempts	N	Min	Max	Mean	SD	Mean	SD
1	18	48.22	74.80	61.72	7.94	7.03	.44
2	37	31.17	73.00	59.64	7.86	6.80	.45
3	42	44.50	76.30	61.13	6.67	6.70	.49
4	31	45.60	81.27	58.37	7.38	6.63	.43
5	12	42.40	77.20	57.69	8.92	6.46	.62
6	9	48.20	68.20	60.99	6.74	6.61	.42
7	3	51.40	56.20	53.33	2.53	6.83	.29
8	1	-	-	71.40	-	6.50	-

In a similar vein, there does not seem to be a straightforward linear relationship between number of attempts and participants' academic grades. Although participants who made only one IELTS attempt achieved the highest academic grades compared to the rest of the research sample (except the participant who repeated IELTS 8 times), the difference in grades was fairly small. Correlation analysis confirmed that there was no significant correlation between attempts and academic grades, r=-.088, p>.05, which suggests that attempts might not be a good predictor for academic grades, or contribute unique variances to the predictive validity of IELTS.

To test such assumption, regression analysis was conducted, using IELTS as the main predictor, test preparation attendance and number of attempts as the additional predictor (entered in the presented sequence), and academic grades as the outcome variable, controlling for non-verbal intelligence and working memory. Results of regression were summarised in Table 4.2.14.

Table 4.2.14 Regression model using final IELTS overall, test preparation attendance and number of attempts to predict the academic grades for all participants, controlling for non-verbal intelligence and working memory (N=153)

Model	Unstandardized B	Coefficients SE	β	t	*p*	R^2	ΔR²	ΔR² sig.
1						.003	.003	.506
Constant	57.705	3.371		17.118	.000			
WASI	.130	.196	.054	.667	.506			
2						.018	.015	.133
Constant	57.662	3.357		17.177	.000			
WASI	.133	.195	.055	.682	.496			
Digit span	.537	.356	.122	1.510	.133			
3						.029	.011	.192
Constant	46.447	9.188		5.055	.000			
WASI	.138	.194	.057	.709	.480			
Digit span	.513	.356	.117	1.443	.151			
IELTS	1.655	1.263	.106	1.311	.192			
4						.41	.012	.174
Constant	43.820	9.361		4.681	.000			
WASI	.128	.194	.053	.661	.510			
Digit span	.494	.355	.112	1.392	.166			
IELTS	1.928	1.275	.123	1.512	.133			
Test preparation	1.681	1.230	.111	1.367	.174			
5						.44	.003	.493
Constant	45.998	9.898		4.647	.000			
WASI	.129	.194	.053	.663	.509			
Digit span	.450	.361	.102	1.245	.215			
IELTS	1.733	1.308	.111	1.325	.187			
Test preparation	1.767	1.239	.117	1.427	.156			
Attempts	-.293	.426	-.058	.493				

As shown, the final predictive model with IELTS, test preparation and attempts was not significant, leading to the conclusion that when all participants were looked at together, IELTS, test preparation and attempts could not predict their academic attainment. However, as stated above, given the crossdisciplinary nature of this research sample and the differences in linguistic demand discussed in previous sections, more analyses were called for so as to further explore the predictive validity of IELTS and attempts.

To take into account the differences in linguistic demand, two more regression models (summarised in table 4.2.15 and 4.2.16) were fitted, one to predict the grades achieved by linguistically more demanding participants and the other for linguistically less demanding participants.

Table 4.2.15 Regression model using final IELTS overall, test preparation attendance and number of attempts to predict the academic grades for participants from linguistically more demanding disciplines, controlling for non-verbal intelligence and working memory (N=80)

Model	Unstandardized	Coefficients	β	t	*p*	R^2	ΔR²	ΔR² sig.
	B	SE						
1						.002	.002	.707
Constant	56.730	4.237		13.389	.000			
WASI	.094	.250	.043	.377	.707			
2						.050	.048	.052
Constant	56.072	4.174	.060	13.435	.000			
WASI	.132	.246	.220	.537	.593			
Digit span	.829	.419		1.978	.052			
3						.167	.117	.002
Constant	.332	17.515		.019	.985			
WASI	.065	.233	.029	.278	.782			
Digit span	.601	.401	.160	1.499	.138			
IELTS	8.169	2.501	.348	3.266	.002			
4						.178	.011	.313
Constant	-1.873	17.646		-.106	.916			
WASI	.057	.233	.026	.247	.806			
Digit span	.584	.402	.155	1.455	.150			
IELTS	8.398	2.511	.358	3.345	.001			
Test preparation	1.362	1.342	.107	1.015	.313			
5						.181	.003	.607
Constant	1.104	18.647		.059	.953			
WASI	.052	.234	.023	.220	.826			
Digit span	.532	.416	.141	1.280	.204			
IELTS	8.090	2.593	.345	3.120	.003			
Test preparation	1.454	1.361	.114	1.068	.289			
Attempts	-.248	.481	-.059	-.516	.607			

Table 4.2.16 Regression model using final IELTS overall, test preparation attendance and number of attempts to predict the academic grades for participants from linguistically less demanding disciplines, controlling for non-verbal intelligence and working memory (N=73)

Model	Unstandardized B	Coefficients SE	β	t	*p*	R^2	ΔR^2	ΔR^2 sig.
1						.001	.001	.784
Constant	60.296	5.129		11.756	.000			
WASI	.080	.293	.033	.275	.784			
2						.004	.003	.641
Constant	60.448	5.168		11.697	.000			
WASI	.072	.295	.029	.244	.808			
Digit span	.268	.572	.056	.469	.641			
3						.060	.055	.048
Constant	36.599	12.873		2.843	.006			
WASI	.069	.288	.028	.239	.812			
Digit span	.284	.560	.059	.507	.614			
IELTS	3.697	1.835	.235	2.015	.048			
4						.077	.018	.257
Constant	32.889	13.248		2.483	.016			
WASI	.054	.288	.022	.188	.851			
Digit span	.261	.559	.055	.467	.642			
IELTS	4.098	1.864	.261	2.199	.031			
Test preparation	2.265	1.980	.136	1.144	.257			
5						.080	.003	.668
Constant	30.576	14.369		2.128	.037			
WASI	.045	.291	.018	.155	.877			
Digit span	.286	.565	.060	.505	.615			
IELTS	4.340	1.957	.276	2.217	.030			
Test preparation	2.208	1.997	.132	1.106	.273			
Attempts	.297	.690	.053	.431	.668			

As shown, neither test preparation nor number of attempts made a significant contribution to the overall model fit for both subgroups, albeit IELTS remained a significant predictor for academic grades in both groups' final models. This indicates that the number of IELTS attempts made by participants to achieve the required scores did not significantly improve the predictive validity of IELTS or in other words, the predictive validity of IELTS was not eroded by test preparation and repeated test taking to a degree that was statistically significant.

4.2.6 Explorative analysis: IELTS repeater's profile

This section attempts to build a repeaters' profile for IELTS, using data collected through questionnaire. First, it looked at the relationship between admission IELTS requirement (summarised in Table 4.2.17) and the pattern of repeated IELTS taking. Then the relationship between age, length of English language learning and repeated IELTS taking was probed.

Table 4.2.17 Summary of participants' admission minimum IELTS requirement (overall and componential), number of attempts made (N=153)

Overall requirement	Componential requirement	Number of participants (%)	Average attempts made
6.0	a minimum of 5.5 in each component	16 (10%)	3.38
6.5	a minimum of 6.0 in each component	12 (8%)	3.25
6.5	a minimum of 5.5 in each component	3 (2%)	1.67
6.5	a minimum of 6 in each component	11 (7%)	3.09
6.5	with 6.0 in writing and a minimum of 5.5 in all other components	1 (1%)	5.00
6.5	a minimum of 6.5 in Writing and a minimum of 6.0 in all other components	27 (18%)	3.00
6.5	a minimum of 6.5 in each component	3 (2%)	4.67
7.0	a minimum of 6.0 in each component	71 (46%)	3.15
7.0	a minimum of 6.0 in Listening and Speaking, a minimum of 6.5 in Reading, and of 7.0 in Writing	1 (1%)	1.00
7.0	a minimum of 6.0 in Writing and no less than 5.5 in all other components	1 (1%)	6.00
7.0	a minimum of 7.0 in Writing and Speaking and no less than 5.5 in all other components	1 (1%)	2.00
7.0	a minimum of 6.0 in Writing	1 (1%)	3.00
7.0	a minimum of 6.0 in each component	5 (3%)	3.40

As shown in Table 4.2.17, in addition to the three main levels of minimum IELTS overall requirement: 6.0, 6.5, and 7.0, there were large variation in terms of minimum componential requirements. For example, within the same overall 6.5 category, some programmes only required "no less than 5.5 in each component" while others asked for "no less than 6.5 in each component"; the latter presents considerably more challenges than the former. Because of these specific requirements and that IELTS results taken at different times could not be merged, it is reasonable that many candidates resorted to taking IELTS repeatedly in order meet the demand, i.e. more attempts.

Patterns have emerged through looking at the number of attempts made and admission IELTS requirement among this research sample. Firstly, it seems that the achievability of IELTS componential-specific requirements are closely associated with the number of attempts needed. For example, for an overall of 6.5, it took participants almost twice as more attempts to achieve *a minimum of 6 in each component* than *a minimum of 5.5 in each component*. Secondly, it also appears that componential-specific requirements present more challenges to these participants rather than the overall IELTS score. When the overall requirement increased by one unit (e.g. from 6.5 to 7), it does not necessarily mean more attempts were needed but when the specific componential requirements increased by the same degree (e.g. from 6.0 to 6.5), more attempts were made indeed. For example, it took 3.15 attempts to achieve *7 with a minimum of 6.0 in each component* but 4.67 attempts were needed to achieve *6.5 with a minimum of 6.5 in each component*.

Regarding how age and length of English learning relate to repeated test taking, participants were grouped into the six groups on the basis of their number of IELTS attempts (Table 4.2.18). Note here that participants who made no less than 6 attempts were grouped together because the number of six-timers, seven-timers and eight-timers involved in this sample was very small.

Table 4.2.18 Summary of participants' age and length of English learning (N=153)

Group	Number of IELTS attempts	Amount of participants	Age (year)		Length of EFL learning (year)	
			Mean	SD	Mean	SD
1	1	18	24.27	2.03	15.72	3.07
2	2	37	23.06	1.32	13.98	2.09
3	3	42	23.32	1.63	14.49	1.78
4	4	31	23.31	1.55	14.99	1.90
5	5	12	23.63	2.00	15.38	2.67
6	6 and plus	13	22.53	1021	13.68	2.03

From Table 4.2.18, it could be seen that in general, one-timers were older than repeaters and had comparatively longer EFL learning. ANOVA was performed to examine whether there are statistical differences in participants' age and length of English as a foreign language learning between groups. Results showed that, in general, there was a marginally significant difference between groups, $F(5,147)=2.204, p=.057$. Planned contrasts revealed a significant difference in age between One-timers and Two-timers, $t(147)=2.634, p<.05$, between One-timer and Three-timers, $t(147)=2.118, p<.05$ and between One-timers and Four-timers, $t(147)=2.035, p<.05$, were also significant, meaning that those who only took IELTS once were significantly older than those who took IELTS twice, three-times and four-times.

ANOVA revealed similar results when examining the difference in participants' length of English as foreign language learning. In general, there was a significant difference between groups, *F*(5,147)=2.582, *p*<.05. Planned contrasts found a significant difference in EFL learning between One-timers and Two-timers, *t*(147)=2.799, *p*<.05, and between One-timers and Three-timers, *t*(147)=2.026, *p*<.05, suggesting that those who only sat one IELTS spent significantly more time learning English than those who sat two or three IELTS. This difference could be attributed to the fact that One-timers were significantly older than Repeaters, as shown in the previous analyses.

To sum up, explorative analyses came to the following two interesting findings. Firstly, the achievability of IELTS componential-specific requirements, instead of overall IELTS requirement, was closely associated with the number of attempts needed. In other words, the higher the IELTS componential-specific requirements, the more likely the repeated test taking was to take place. Secondly, it was also found that candidates who sat only one IELTS were significantly older and were likely to spent more time learning English as a foreign language than those who sat IELTS repeatedly, in particular, those who took IELTS twice, three-times and four-times.

4.2.7 Explorative analysis: Does the predictive validity of IELTS on academic attainment change over time?

To answer this explorative question, average academic grades were calculated from participants enrolled in the department of Education (N=82). The reason why only 82 data points were available was because the division of marks by term in which they were taken was available only for this subset of participants. For the following analysis on time, participants' grades were not weighted because all participants of this cohort earned 100 credits from 5 modules and each module had 20 credits.

On average, this cohort achieved an final IELTS overall of 6.93 (SD=.31), higher than the overall research sample, and a weighted academic average of 58.31 (SD=6.29), slightly lower than the sample. The average grades achieved by this cohort was 57.36 (SD=7.01) for term 1, and 59.72 (SD=7.27) for term 2. The assessment methods used to evaluate participants' term 1 academic attainment was one written exam and two essays while the methods used for term 2 was one written exam and one essay.

This explorative investigation was based on Chen & Sun's TOEFL study (2006) and Yen & Kuzma's (2009) IELTS study, where researchers found that predictive validity of TOEFL and IELTS was prone to change over time through correlational analyses. On the basis of this, similar analyses are performed here in this section.

Correlation analysis between participants' term 1 average academic grades and their final IELTS overall revealed a moderately significant relationship with an coefficient of .334, *p*=.002. On the contrary, correlation between IELTS and participants' term 2 average academic grades was not found significant, *r*=.169, *p*=.130.

This change in IELTS's relationship with academic grades lends support to the assumption that time may have an important role to play when examining the relationship between proficiency and academic attainment. The *wearing-off* of IELTS's significant correlation with academic attainment is interesting as it suggests that at the beginning of their academic study, participants' ability to perform well academically was significantly hindered by their lack of sufficient proficiency. However, as they furthered into their study, it seems that their proficiency had eventually caught up and no longer acted as a hinderance. This finding, combined with the afore-discussed effects of test preparation and repeated test taking, indicated that although test preparation and repeated test taking could create a discrepancy between scores and candidates' proficiency, over time, this discrepancy may diminish. On the basis of this, one may speculate that if sufficient time is provided and enough efforts are made, international students can eventually develop the needed proficiency for their academic study. However, as in this quasi-experiment, average term grades only consisted 2-3 modules scores, results may not be robust enough to draw a conclusive picture regarding how time affects the relationship between proficiency and academic attainment. More research is therefore called for to further explore this interaction.

4.2.8 Explorative analysis: to what extent does general proficiency (indicated by the composite score of DET and C-test scores) predict international Chinese students' academic attainment?

Given that this quasi-experiment has collected data concerning participants' general proficiency and academic attainment, examining the predictive power of general proficiency is considered worthy and appropriate. Hence, a regression modelling was fitted using the composite score of DET and C-test as the predictor, controlling for non-verbal intelligence (i.e. WASI) and working memory (i.e. a composite score of digit span Chinese and digit span English). The difference in disciplinary linguistic demand was also accounted as an additional predictor in the final model. The final model, summarised in Table 4.2.19, showed that general proficiency became a significant predictor overall, *F*(3,149)=1.223, *p*>.05 only after disciplinary linguistic demand was taken into consideration.

Table 4.2.19 Regression model using general proficiency to predict the academic grades for all participant, controlling for non-verbal intelligence and working memory (N=153)

Model	Unstandardized B	Coefficients SE	β	t	*p*	R^2	ΔR^2	ΔR^2 sig.
1						.003	.003	.506
Constant	57.705	3.371		17.118	.000			
WASI	.130	.196	.054	.667	.506			
2						.018	.015	.133
Constant	57.662	3.357		17.177	.000			
WASI	.133	.195	.055	.682	.496			
Digit span	.537	.356	.122	1.510	.133			
3						.024	.006	.333
Constant	57.870	3.364		17.201	.000			
WASI	.121	.195	.050	.618	.538			
Digit span	.489	.359	.111	1.361	.175			
General proficiency	.264	.272	.080	.971	.333			
4						.107	.083	.000
Constant	61.838	3.401		18.183	.000			
WASI	.035	.189	.015	.187	.852			
Digit span	.429	.345	.097	1.241	.217			
General proficiency	.740	.291	.223	2.546	.012			
Linguistic demand	-4.821	1.298	-.322	-3.714	.000			

The fact that the linguistic demand of the programme was a significant predictor is in line with the previous regression analyses results using IELTS as the main predictor (section 4.2.4). In other words, when all participants were looked at together without accounting for linguistic demand of the programme (Model 3), general proficiency was not predictive of academic outcomes. It is only when the linguistic demand was taken into account that the effect of general proficiency became apparent.

4.3 Discussion

The first goal of this quasi-experiment was to examine the effects of repeated test taking on candidates' IELTS scores and their general proficiency, aiming to answer the specific question: whether by repeatedly sitting the test, candidates could achieve IELTS scores that were beyond the level of their actual proficiency. In addition, this quasi-experiment also looked at the effect of dedicated IELTS preparation with the aim to corroborate the findings from the first quasi-experiment and generalise findings applicable to a different population. Furthermore, this quasi-experiment investigated the relationship between language proficiency (indexed by IELTS) and international Chinese postgraduate students' academic attainment at a UK university, taking into account the variation in disciplinary linguistic demand while controlling for non-verbal intelligence and working memory.

This quasi-experiment set out to answer three specific research questions, related to the effects of repeated test taking on proficiency and IELTS scores, predictive validity of IELTS, and the effect of repeated test taking and test preparation on the predictive validity of IELTS. Answers to these questions are presented and discussed in the following sections of this chapter, drawing reference from literature reviewed in Chapter 2, in particular section 2.5-2.6. Findings on test preparation from this quasi-experiment, proficiency and IELTS scores are not discussed here; instead they are included in the General Discussion (Chapter 5) where collaborative evidence from both quasi-experiments are synthesized.

4.3.1 To what extent does repeated IELTS taking affect Chinese candidates' IELTS scores and their general proficiency?

To answer this research question, this quasi-experiment adopted a correlational design. Participants' IELTS scores and their IELTS taking history were collected through questionnaire, and their general proficiency was measured using two linguistic measures: Duolingo English Test (DET) and C-test.

Results indicated that there was a significant and negative correlation between participants' overall general proficiency (indexed through the composite score of DET and C-test) and their number of IELTS attempts. This suggests that in general, participants who made more attempts were found to be of lower general proficiency and participants who made less attempts were found to be of higher proficiency. This confirms the first hypothesis proposed in section 4.1.7 (i.e. through multiple attempts, scores on IELTS can be boosted without corresponding improvement in proficiency), rejecting the notion that candidates who made more attempts could develop a higher proficiency owing to the effort and time needed for making multiple attempts. Further exploration of the data yielded that this negative correlation existed at every level of IELTS: for participants who reported the same final IELTS overall, those who made more attempts to achieve this score were of lower proficiency compared to those who achieve the same score through less attempts. This indicates that by repeatedly taking IELTS, candidates could boost their scores onto a level that is beyond their general proficiency. Because of the uniqueness of this quasi-experiment, this finding taps into an area that is rarely explored among previous research, apart from Wilson's TOEFL repeater study (1987) and Zhang (2008).

Wilson (1987) reported that for TOEFL, with each additional attempt, score gains could be achieved. He further hypothesised that such gains were attributed to longer learning time and more effort spared by the candidates. However, on the basis of this quasi-experiment's findings, assuming what was found in IELTS could also be applied to TOEFL, Wilson's hypothesis could be rejected. Evidence from this quasi-experiment highlighted that scores gains achieved on a proficiency test through repeated test taking did not necessarily translate to improvement in general proficiency. Instead, findings suggest that the observed score gains were more likely to be results of accumulated test-wiseness as candidates become more test-savvy through the repetition of test taking. Thus, the variation in general proficiency for participants who achieved the same IELTS scores could also be interpreted as variation in testwiseness, in the sense that those who attempted IELTS more might have a higher level of testwiseness than those who attempted IELTS less. This finding has important implication as to the validity of IELTS.

As a widely recognised test of English proficiency, the construct IELTS aims to measure is English language proficiency; to achieve this goal, IELTS assesses the four core language skills and produce an overall score as an index. If a candidate achieves a certain band on IELTS, one would associate the score with a particular corresponding level of proficiency based on the guidance of IELTS bandscore descriptor. If two candidates achieve the same IELTS score, one would naturally associate them with the same (or very similar) level of proficiency. However, this association may no longer be accurate because testwiseness, a construct irrelevant variance, has been introduced into IELTS scores through repeated test taking, thereby affecting inferences drawn from

IELTS scores. More detailed discussion on how repeated test taking affect the various validity dimensions of IELTS is provided in Chapter 5, along with the effects of test preparation.

4.3.2 To what extent does IELTS predict international Chinese students' academic attainment at a UK university?

Preliminary correlation analysis showed that participants' academic attainment (operationalised by the weighted average academic grades) was not significantly correlated with their final IELTS scores, which indicates that the ability to obtain academic success might be more closely related to other non-linguistic factors. This was in line with a number of IELTS predictive studies that reported an insignificant correlation between IELTS and international students' academic achievement measured by GPA, such as Cotton and Conrow (1998), Kerstjen and Nery (2000), Dooey and Oliver (2002). Although this finding indeed echoed with some existing studies, for reasons to be explained, this insignificant correlation could be argued as inaccurate.

Firstly and for most, this quasi-experiment involved a cross-disciplinary research sample as participants were heterogenous in terms of their fields of study. With this crossdisciplinarity comes variation in the linguistic demand of their academic programme. While some participants were enrolled in linguistically very demanding disciplines such as English literature and Education, where their language skills play an important role, half of the sample were from linguistically less demanding disciplines such as electric and electronics and chemistry. Given this variation, one may say that correlation between IELTS and academic grades at an overall level (i.e. participants were looked at together as one homogenous group) provides little insight into the "real" role proficiency plays. Hence, the insignificant correlation found in this quasi-experiment is believed an incorrect representation of the relationship between proficiency and academic attainment.

To account for the disciplinary variation in linguistic demand, participants were grouped as linguistically less or more demanding based on the language requirement set for their admission and correlation analyses were conducted respectively in each group. Contradictory to the overall insignificant correlation observed when participants were treated as one unified group, group-level correlation revealed that IELTS was, in fact, significantly correlated with participants' academic attainment both for the linguistically less demanding and for the linguistically more demanding. In addition to the significant correlation, it was also found that the correlation among the linguistically more demanding participants was stronger than that among the linguistically less demanding participants, confirming the hypothesis that the relationship between IELTS and academic attainment could be affected by the linguistic demand of the discipline candidates were enrolled in.

This change in correlation results, once discipline was factored in, highlighted the probability that, although there existed a significant relationship between proficiency and academic attainment, it can be "masked" by the noise from the data. This finding is important as it brings into questioning the findings from a number of predictive studies that did not find a significant relationship between scores on proficiency tests and international students' academic attainment, and thus went on to claim attainment was not affected by proficiency (e.g. Hwang & Dizney, 1970; Kerstjen & Nery, 2000; Krausz et al., 2005)

To further examine the predictive validity of IELTS, linear regression was performed, first at an overall level (i.e. participants were looked at together as one group) and later at a group-level (i.e. linguistically more/less demanding). Results revealed that IELTS was not a significant predictor for academic attainment at an overall level. However, and more importantly, when participants were grouped based on their disciplinary linguistic demand, IELTS was found to be a good predictor for academic attainment for participants studying linguistically more demanding disciplines. For those studying linguistically less demanding disciplines, the final model consisting of non-verbal intelligence, working memory, IELTS was not significant, albeit the inclusion of IELTS did improve the model fit significantly.

This significant predictive validity of IELTS for academic attainment among linguistically more demanding students reported in this quasi-experiment was consistent with that reported by Feast (2002) who examined the predictive validity of IELTS in a cross-disciplinary multi-nationality sample and Trenkic and Warmington (2018) who examined the relationship between IELTS and academic attainment in a homogenous sample. This finding highlights that it is crucial to consider disciplinary linguistic differences when working with a cross-disciplinary sample, which many previous predictive studies failed to do (e.g. Dooey & Oliver, 2002; Person, 2002).

4.4 Conclusion

This quasi-experiment illustrated that, although international students involved in this quasi-experiment were admitted on the pre-condition that they had obtained the required IELTS scores, their academic attainment was, to various extent, hindered by their proficiency level. Those who had better proficiency were more likely to achieve higher academic grades in comparison to those who had a lower IELTS scores. This creates a dilemma: on the one hand, receiving HEIs assume that international students arriving with the required IELTS scores should have obtained sufficient language proficiency that would put them on an even playground competing against home students, at least linguistically speaking. On the other hand, findings from the present study and other research (e.g. Daller & Xue, 2009; Daller & Phelan, 2013; Ho & Spinks, 1985) show that language is still an hinderance. From this, it is clear that the language requirements were not set at a correct level, allowing international students to fulfil their academic potential fully. Proficiency in the language of instrument does play a pivotal role in international students' pursue for academic success, especially for those who were enrolled in linguistically more demanding disciplines.

Chapter 5 General Discussion

This chapter first links the findings from both quasi-experiments and discusses the effects of IELTS test preparation and repeated test taking from two main aspects: test validity and academic attainment. Following this, the implication drawn from both quasi-experiments are presented.

5.1 Test preparation, repeated test taking and construct validity

Both quasi-experiments in this study illustrated that, scores on IELTS could be boosted onto a level that is beyond candidates' general proficiency by attending dedicated IELTS preparation programmes, and to a lesser extent, by taking IELTS repeatedly. Results from quasi-experiment 1 showed that through attending a typical 4-week intensive IELTS preparation programme offered at a standard Chinese test training centre, candidates were enabled to boost their overall IELTS scores significantly by 0.6 bandscore on average. This effect was even bigger for those who started the programme with a lower proficiency (mean overall score gain=.89). In a similar vein, quasi-experiment 2 found that candidates who arrived at the same final IELTS overall scores, their level of proficiency varied, with those who attended IELTS preparation and/or sat IELTS repeatedly being less proficient than those who did not. These findings are of significance as they relate to the validity of IELTS and the interpretation of IELTS scores.

To quickly recap, at its very essence, validity means whether a test measures what it claims to measure (Messick, 1989). In the context of this study, the validity of IELTS lies at the heart of whether IELTS can accurately measure the English language proficiency of its candidates through assessing candidates' proficiency in each language skill, i.e. listening, reading, writing and speaking.

Although both quasi-experiments were not designed to directly measure the impact of test preparation and repeated test taking on IELTS validity, inferences can still be drawn, particularly on IELTS's construct validity. As a widely recognised English proficiency test, naturally, the construct IELTS aims to measure is candidates' proficiency. Considering the difficulty of assessing the construct of proficiency as one unified concept, IELTS approaches the measure of proficiency via assessing the four language skills. For IELTS to have construct validity, the four modules included in

IELTS, i.e. IELTS listening, reading, writing and speaking, should assess candidates' listening, reading, writing and speaking skills.

As discussed in Chapter 2 (section 2.5.2), validity is not a property for the test per se, but rather, the score of the test and the inferences drawn from the test scores (Cronbach, 1971). Therefore, for IELTS scores to be construct-valid, one would expect the IELTS overall scores to be the correct representation of candidates' general proficiency and the IELTS listening, reading, writing and speaking scores to be in line with candidates' four language skills. However, findings from the present study indicate that because of the effects of test preparation and/or repeated IELTS taking, the construct validity of IELTS scores are undermined by construct irrelevance and construct underrepresentation.

In the context of IELTS as a proficiency test, construct irrelevance encompasses knowledge and skills that are not closely relevant to candidates' listening, reading, writing and speaking skills. One of the most common form of construct irrelevance is testwiseness, i.e. the rich experience of test-taking and the familiarity of test formats (Cohen, 2014). It is importantly to acknowledge although the nature of testwiseness is often not relevant to the construct, it does not necessarily mean that testwiseness *always* undermines a test's construct validity.

On the one hand, the familiarity of test formats can, in many cases, improve a test's construct validity in the sense that when candidates are familiar with the formats of the test, they are more likely to showcase their *real* ability through the test, in comparison to those who are unfamiliar. For example, in IELTS listening test, candidates familiar with the rule that IELTS listening recording is only played once are more likely to answer questions while listening. Without this awareness of the rule, candidates may wait till the end of the recording to answer the questions thinking that there is going to be a second play, and thus their scores may not be correct representation of their listening skills, weakening both the construct validity of the test and the scores. On the other hand, the rich experience of test-taking, often manifested through the use of test taking strategies, is considered detrimental to the construct validity of a test and its scores because this type of testwiseness allows candidates to achieve higher scores that are beyond the level of their "true" ability.

As shown by both quasi-experiments in this study, a gap between what candidates' IELTS scores entail and what their proficiency level is actually at has been created through attending IELTS preparation courses and at a lesser degree, by taking IELTS repeatedly. This suggests that, for candidates who attended test preparation and/or sat IELTS repeatedly, their IELTS scores may no longer correctly represent their proficiency, whereas for the uncoached candidates (i.e. control), IELTS scores are indeed accurate. This differences in the interpretation of IELTS scores for the coached and uncoached candidature, as a result of test preparation and repeated test taking, map to the weakening of extrapolation link discussed by Xie's research (2013).

As reported in quasi-experiment 1, based on the significant boost in IELTS scores (overall and by skill), the lack of improvement in general proficiency (indexed by scores on OOPT, lexical knowledge and sentence processing accuracy and speed) and the typical pedagogy of the test preparation course, it is reasonable to speculate that candidates were taught a considerable amount of test taking knowledge and test taking strategies, which made test taking easier for the coached candidates, i.e. construct-irrelevant easiness (Messick, 1981). This speculation is consistent with test-centred preparation practices reported by Mickan and Motteram's observational research (2008) and the effects of test-oriented practices, practice-and-drill, memorisation found in Green's (2007) and Xie's (2013) work. Since the design and the assessment goal of IELTS is to evaluate the proficiency level of learners of English, not how savvy candidates are at taking the test, there is good reason to conclude that test preparation and repeated test taking have eroded the construct validity of IELTS scores.

In addition, previous research has pointed out that construct validity could be affected by construct underrepresentation, i.e. the construct assessed in IELTS accurately portrays the demand of listening, reading, writing and speaking tasks candidates may encounter in an English speaking environment. For example, whether the listening tasks included in IELTS represent the listening tasks candidates face when they study or work in an English speaking university/organisation. As found through quasi-experiment 1, IELTS preparation significantly boosted candidates' IELTS scores with 4 weeks, if candidates were not aware that such boosts in scores did not map to improvement in their general, they might assume that improving general proficiency is a goal achievable in a short period of time, which contradicts the stand IELTS holds and evidence from a number of studies.

Cambridge Assessment, who owns IELTS jointly with Australia Idp, British Council, asserts that a minimum of 3 months learning (around 200 guided hours) is needed in order to achieve one unit of score band improvement (Desveaux, 2018). Green's research (2005) also reported that improvement in IELTS Academic writing is not a goal achievable without substantial time and efforts, as participants' involved in his study hardly made any progress in their writing scores after 15 months of EAP study. Similarly, Craven reported that a large proportion of students could not improve their IELTS Academic writing scores from 6.5 to 7.0 even after 3 years of studying at an Australian university. In a similar vein, Trenkic & Warmington's study highlighted that after a year's study in a UK university, international Chinese students did not significantly improve their proficiency after studying in the UK for 8 months. This stark contrast between the amount of time needed to improve proficiency (indexed by score gains) asserted by the IELTS developer and researchers and the actual time needed to achieve score gains observed in the present study serves as first evidence of underrepresentation.

Moreover, because the goal of test preparation courses investigated in quasi-experiment 1 was to achieve score boosts in a limited timeframe, the teaching and learning that took place during the courses were mostly centred on retired IELTS test papers (i.e. Cambridge IELTS book series). This narrow learning manifested in all modules, and as candidates' focuses were primarily on testwiseness learning; thereby, construct being assessed becomes simplified. For example, reading score gains observed in the present study, coupled with the fact that there was no corresponding improvement in participants' vocabulary knowledge or their accuracy and speed of sentence processing indicated that score gains were more closely associated with certain simplified testwiseness techniques, such as to look for clues through scrutinising the wording of the questions (Ma, 2014).

In a similar vein, construct underrepresentation could be also introduced through repeated test taking. For example, through repeated sitting the IELTS writing module, candidates may become increasing skilled at producing written work to cater for the specific marking criteria of IELTS so as to achieve a better score. By doing so, their learning becomes narrowed, and the construct that IELTS writing sets out to measure, i.e. academic writing skills, becomes underrepresented. Similarly, through repeated sitting IELTS speaking test, candidates have the chance to practise IELTS relevant questions and converse on IELTS related topics. Thereby, their scores are more closely related with their IELTS speaking specific knowledge, not their overall speaking skills in an academic context, which could be argued as construct representation.

Taken together, evidence from both quasi-quasi-experiments reveal that the construct validity of IELTS is eroded through attendance at test preparation programmes and/or to a lesser extent, by repeated test taking, which echoes Strathern’s claim that "when a measure becomes a target, it ceases to be a good measure" (1997, p.308).

As construct validity acts as the foundation upon which other dimensions of validity, such as predictive validity, are built, the next section explores whether test preparation and repeated test taking affected the predictive validity of IELTS for academic attainment for international students enrolled in a UK university.

5.2 Test preparation, repeated test taking and candidates’ proficiency

The fact that test preparation and repeated test taking were found to boost scores beyond candidates’ general proficiency not only among candidates preparing for IELTS tests but also for international students enrolled in a UK university indicated that there existed a discrepancy between IELTS scores and proficiency. This discrepancy can put international students at a disadvantage, particularly at the beginning of their academic journey, as a moderately significant relationship between IELTS scores and term 1 grades was found in quasi-experiment 2, but not between IELTS and term 2 grades. In other words, there is good reason to infer that many international students enrolled in English speaking higher education institutions do not have the same proficiency as what their IELTS scores entail.

This resonates with Du-Babcock’s study (2002) where international Chinese students were found to face linguistic struggles, particularly in the context of following lectures, even although they scores high on admission tests. Likewise, in Ma’s study (2014), international Chinese students noted that even when they scored way above the minimum requirement in TOEFL, once they arrived at an American university, they quickly released that their level of proficiency was not sufficient. They also experienced difficulty in gaining full understanding of the learning content (Edwards & Ran, 2006), and had problem participating in group discussion and voicing their opinions in front of their peers (Li, Duanmu & Chen, 2009). These struggles and difficulties international students experience could be related to the test preparation and repeated test taking because both allowed candidates to boost scores beyond the level of their proficiency, which means that it is possible that the admitted students are not as "linguistically ready" as what their scores may suggest. Because they may not be linguistically ready, coupled with the key role language proficiency plays in achieving academic success (e.g. Daller & Xue, 2009; Daller & Phelan, 2013; Trenkic & Warmingtom, 2018), it is not surprising that international students have been found to achieve less academic success in comparison to the home students (e.g. Crawford & Wang, 2014; Iannelli & Huang, 2013).

5.3 Test preparation, repeated test taking and academic attainment

The predictive validity of standardised proficiency tests such as IELTS on academic attainment has been heavily debated in the past two decades and mixed results have been reported (e.g. Cotton & Conrow, 2002; Dooey & Oliver, 2002; Feast, 2002; Kerstjens & Nery, 2000; Trenkic & Warmington, 2018). This study found that, albeit international students may have lower proficiency than what their IELTS scores entail, owing to the afore-discussed effects of test preparation and/or repeated test taking, IELTS was significantly correlated with academic grades, after disciplinary differences were accounted for. This is to say, despite the effects of test preparation and repeated test taking, IELTS was still sufficiently robust to differentiate abilities.

Regarding IELTS's inconsistent predictive validity, findings of this study highlights that it can be attributed to the variation in disciplinary linguistic demand if the research sample involved in the study is of cross-disciplinary nature. Therefore, the inconsistent predictive validity of IELTS reported by previous researchers might not be because IELTS is an unstable predictor, it could be that the predictive power of IELTS was subverted by the noise of the data. Meanwhile, based on the evidence this study found on the effects of test preparation and repeated test taking, the inconsistent predictive validity of IELTS could also be attributed to the sample consisting of participants who underwent test preparation and/or repeated test taking and those who did not. As stated earlier, IELTS scores may not correctly represent the proficiency level for candidates who attended IELTS preparation courses and/or sat IELTS repeatedly but for those who did not attend preparation and/or sat IELTS repeated, IELTS scores are believed to be a correct representation. Therefore when participants' test preparation and repeated test taking as well as disciplinary linguistic differences were not accounted for, it is reasonable that mixed findings have been reported.

5.4 Theoretical and practical implications

Given these findings regarding the effects of test preparation and repeated test taking on IELTS scores, general proficiency and international students' subsequent academic attainment at a UK university from the present study, the following theoretical and practical implications could be drawn.

On the theoretical front, the fact that IELTS did not significantly correlate with academic grades when participants were all looked at together but correlated significantly with grades when participants' disciplinary linguistic demand variation was accounted for, has great implication for when interpreting the relationship between proficiency and academic attainment. Although some claim that proficiency, as indicated by IELTS and TOEFL, does not affect the academic attainment for international students enrolled in English speaking HEIs based on their insignificant correlation analysis outcome (Hwang & Dizney, 1970; Kerstjen & Nery, 2000; Krausz et al., 2005), there is abundant literature (Ma, 2014; Robertson et. al. 2000; Sun & Chen, 1999; Wan, 2001) reporting that many international students clearly do not have sufficient language skills to maintain the demand from their academic study, especially for those enrolled in linguistically more demanding programmes such as literature or education. This discrepancy could be attributed to the failure to control for noise from the data such as disciplinary differences in terms of linguistic demand. Based on findings from this study, there is good reason to believe language proficiency does play a crucial role in international students' academic attainment but its importance can be masked when the research sample involves participants from a variety of disciplines.

On a practical front, findings from this study highlighted that academic attainment of international students, particularly those studying linguistically more demanding disciplines, is hindered by the lack of proficiency. On the basis of this, one can argue that the linguistic requirement for international students' admission is not set at a high enough level. According to IELTS Guide for for educational institutions, governments, professional bodies and commercial organisations (2015), for a non-native speaker student to perform effectively in linguistically demanding academic courses, an overall 7.5-9.0 is acceptable and an overall 7.0 is probably acceptable whereas an overall 7.0-9.0 is acceptable and an overall 6.5 is probably acceptable for studying in linguistically less demanding academic courses. From this one could infer that for international students studying linguistically more demanding disciplines, predictive power of IELTS would cease to exist for those who achieve an overall of 7.5 and for those who study linguistically less demanding disciplines, IELTS's predictive power would became insignificant once an overall 7.0 is achieved. However, the sample involved in this study (quasi-experiment 2) did not allow for the statistical testing of such assumptions, due to limited linguistically less demanding participants achieving 7.0 (N=20) and linguistically more demanding participants achieving 7.5 (N=7).

In addition, taking into account the effects of test preparation and repeated test taking, HEI admission should be aware that students arriving with the required IELTS scores may not have the corresponding proficiency that their IELTS may indicate. It is suggested that the receiving HEIs could take note of the test preparation practice candidates have engaged themselves with, for example, the type of courses they took, the intensity of the course, and the length of the course, in addition to candidates' final IELTS scores. By doing so, better insights could gained on the language proficiency of a student and further language support could be provided if necessary. Furthermore, the assessment of international students' academic attainment should also take into account that students who arrive with the required IELTS results may not have obtained the corresponding level of proficiency. More research is called for to determine what adjustments are needed and how such adjustments should be made to facilitate a fair evaluation of international students' academic attainment, be it extra time or access to a dictionary (Trenkic & Warmington, 2018).

From students' perspective, awareness should be risen on the fact that although test preparation and/or repeated test taking can lead to better scores, such score improvement does not correspond to a higher level of general language proficiency. Given the afore-discussed crucial role proficiency plays in their subsequent academic study, students are encouraged to invest more time and efforts in language learning at a more general and non-test specific level.

Because of the high stake tests such as IELTS hold, it is unrealistic to stop students from attending test preparation courses. However, it is possible for course developers and tutors to incorporate general language learning as well as test learning in the courses and expand the course content and practices to focus more than the sheer development of testwiseness.

Chapter 6 Conclusion

The present study involved two empirical quasi-experiments. The first quasi-experiment looked at the effects of test preparation on IELTS test scores and general proficiency using a pretest/intervention/posttest design among a total of 89 Chinese English learners. The second quasi-experiment examined the combined effects of test preparation and repeated test taking on IELTS scores, general proficiency and academic attainment among a group of 153 international Chinese students enrolled in a UK university, using a longitudinal correlational approach.

Results from both quasi-experiments show that test preparation and, to a lesser extent, repeated test taking allow candidates to achieve IELTS scores that are above their proficiency for both English learners in China and international Chinese students enrolled in an English speaking university. These findings shed light on the dilemma of why international Chinese students, admitted on the premises of their IELTS scores, are still linguistically challenged when studying in English-speaking higher education institutions (e.g. Angelova & Riatzatseva, 1999; Du-Babcock, 2002; Sun & Cheng, 1999; Trenkic & Warmington, 2018). Inferences drawn from the effects of test preparation and repeated test taking suggest that the construct validity and reliability of IELTS as a measure of English proficiency have been undermined, thus influencing the interpretation and extrapolation of IELTS scores.

This study also found, albeit the effects of test preparation and repeated test taking, international Chinese students' IELTS scores were indeed associated with their academic attainment, when students' fields of disciplines were taken into account. International Chinese students from both linguistically more or less demanding disciplines were more likely to achieve higher academic grades when their IELTS scores were higher, in comparison to those who had a lower IELTS scores. This positive and significant relationship highlights that language proficiency is indeed at the heart of academic attainment for this population.

Regarding the predictive validity of IELTS, this study found that, even when non-verbal intelligence were accounted for, IELTS still contributed unique variances for the explanation of academic grades, but only for those enrolled in linguistically more demanding disciplines. From a methodological viewpoint, results from this study highlighted the importance to account for the disciplinary linguistic differences when looking at the relationship between proficiency (as measured by IELTS) and academic attainment, especially when the research sample is of a crossdisplinary nature. This, to a certain extent, offers an plausible explanation as to why previous predictive research in the field of IELTS failed to reach an agreement.

Despite the efforts made, this study is still limited in the following respects. Firstly, data collected from quasi-experiment 1 were from one particular test preparation centre in Shanghai, China, which, though typical, might not be able to portray the general test preparation industry in China or in other countries. Thus, future test preparation research is advised to take into account the institutional differences into consideration and collect data from multiple preparation organisations so as to achieve more generalisability for research findings. Secondly, data collected for the second quasi-experiment were, to some extent, truncated, especially in terms of participants' IELTS scores. This is a common issue shared by many predictive research because most predictive study only had access to students who had achieved sufficient IELTS to be accepted by an English-speaking HEI, excluding those who achieved very low on IELTS. Future research could, therefore, involve more students at the lower end of the IELTS spectrum so as to further examine the role proficiency plays in the achievement of academic success at different proficiency level. Thirdly, although this study attempted to build a IELTS repeaters' profile following the guidance from Wilson's (1987) and Zhang's (2008) work, data collected here only allowed a descriptive and fairly superficial profile building for one particular subgroup of IELTS candidature (i.e. Chinese). Hence, to gain a fuller understanding of the repeated test taking behaviour, a cross-cultural large-scale survey is called for. Lastly, concerning the impact time asserts on the predictive validity of IELTS for academic attainment, the present study only had access to grades of two terms; thus, the analyses perfected might not be sufficiently robust. On the basis of this, further research is advised to follow students' academic progress for a longer period of time so as to provide more reliable findings.

Nevertheless, this study extends the literature on the effects of test preparation and repeated test taking on score gains, bridges the gap regarding how test preparation and repeated test taking relate to general proficiency, test reliability and validity, and reaffirms the crucial role proficiency plays in the achievement of academic success at a higher education level.

Appendix 1 Information page and consent form for Quasi-experiment 1 (Intervention group)

Name of Researchers

Ruolin HU

Title of Study

Chinese students' English proficiency

Brief Description of Study

The aim of this study is to investigate the impact of test preparation course on Chinese students' language development. The study will take part across several sessions. As a part of the study, you will be asked to take some language and memory tests in English. The first session will be conducted before your IELTS preparation course starts and the second session will be when your course ends. The subsequent session(s) will happen when you take an IELTS test. Each session will take about 3 hours to complete.

Due to the design of this study, the researcher will need access to your IELTS result report.

As a token of thanks for your participation, participants will be invited into a prize draw.

Where will the research session take place?

The session will take place in the training center where you signed up for your preparation course.

Who will run the research sessions?

The sessions are run by myself, Ruolin Hu.

Will all my details and the results be kept confidential?

Yes. All the information about participants in this study will be kept confidential and for no longer than 5 years. Data will be anonymised and stored securely. To ensure confidentiality each participant will be randomly assigned an ID number; this will be the only form of identification that will be included on any database and paper based tasks used in this study.

You are free to withdraw from the study at any point.

Consent form

Chinese students' English proficiency

Ruolin Hu

	Yes	No
I am a native speaker of Chinese	□	□
I am happy to take part in the above study	□	□
I am happy to take part in all sessions	□	□
I have signed up for IELTS test preparation courses	□	□
I am happy for to disclose my IELTS results to the researcher	□	□
I have read the statement concerning the research that I am being asked to take part in, and I	□	□
have had the opportunity to ask questions.	□	□
I understand that my data will be kept confidential	□	□
I understand that I may withdraw at any time	□	□

Name of participant ……………………………………………………………………

Name of the training course …………………………………………………………

Signature ……………………………………… Date …………………………………

Appendix 2 Information page and consent form for Quasi-experiment 1 (Control group)

Name of Researchers

Ruolin HU

Title of Study

Chinese students' English proficiency

Brief Description of Study

The aim of this study is to investigate the impact of test preparation course on Chinese students' language development. The study will take part across several sessions. As a part of the study, you will be asked to take some language and memory tests in English. The first session will be conducted before your IELTS preparation course starts and the second session will be when your course ends. The subsequent session(s) will happen when you take an IELTS test. Each session will take about 3 hours to complete.

Due to the design of this study, the researcher will need access to your IELTS result report.

As a token of thanks for your participation, participants will paid 100 Yuan for completing the experiment.

Where will the research session take place?

The session will take place in a controlled classroom in one of New Oriental Shanghai Schools.

Who will run the research sessions?
The sessions are run by myself, Ruolin Hu.

Will all my details and the results be kept confidential?
Yes. All the information about participants in this study will be kept confidential and for no longer than 5 years. Data will be anonymised and stored securely. To ensure confidentiality each participant will be randomly assigned an ID number; this will be the only form of identification that will be included on any database and paper based tasks used in this study.

You are free to withdraw from the study at any point.

Consent form

Chinese students' English proficiency

Ruolin Hu

	Yes	No
I am a native speaker of Chinese	□	□
I am happy to take part in the above study	□	□
I am happy to take part in all sessions	□	□
I have NOT signed up for IELTS test preparation courses	□	□
I am happy for to disclose my IELTS results to the researcher	□	□
I have read the statement concerning the research that I am being asked to take part in, and I have had the opportunity to ask questions.	□ □	□ □
I understand that my data will be kept confidential	□	□
I understand that I may withdraw at any time	□	□

Name of participant ……………………………………………………………………

Name of the training course ……………………………………………………………

Signature ……………………………………… Date …………………………………

Appendix 3 Questionnaire for Quasi-experiment 1

Participant ID ______

1. Date of birth: ____(month)/_____(year)
2. What do you currently do?

   □ student □ employed □ unemployed

   □ others (specify _________)
3. Level of education:

   □ postgraduate (year ___)

   □ undergraduate (year ___)

   □ high school (year ___)

   Major of Study (if you are a undergraduate or a postgraduate):

   ____
4. What's your mother tongue? ___________________________
5. How many languages do you speak? ___________

   What are these languages?

   ____

   What language(s) do you speak at home?

   ____
6. Have you visited/lived in any English-speaking country before?

   □ Yes

   □ No

   What are these countries?

   ____
7. How often do you do the following in English?

   Read English books/journals

   □ everyday □ several times a week □ several times a month □ rarely □ never

   Read English emails

   □ everyday □ several times a week □ several times a month □ rarely □ never

   Write English emails

   □ everyday □ several times a week □ several times a month □ rarely □ never

   Listen to English radio

   □ everyday □ several times a week □ several times a month □ rarely □ never

   Listen to English songs

   □ everyday □ several times a week □ several times a month □ rarely □ never

   Browse English Internet

□ everyday □ several times a week □ several times a month □ rarely □ never

Use English social network sites (Facebook, twitter, etc.)

□ everyday □ several times a week □ several times a month □ rarely □ never

Watch English tv

□ everyday □ several times a week □ several times a month □ rarely □ never

Watch English film

□ everyday □ several times a week □ several times a month □ rarely □ never

8. Are there any English native speakers in your daily social life?

□ No

□ Yes (How many? _________)

9. How often do you interact with English native speakers?

□ everyday □ several times a week □ several times a month □ rarely □ never

10. How many hours per day do you use to learn English? _________Hours

11. How many IELTS tests have you signed up for? ______________

12. When is your next IELTS test date? ________________________

13. Have you taken IELTS before?

□ Yes □ No

If yes, how many IELTS have you taken? _______

What are the scores?

IELTS 1: overall _____

listening _____ reading _____ writing _____ speaking _____

IELTS 2: overall _____

listening _____ reading _____ writing _____ speaking _____

If there is more, please specify: _____________________________________

14. Have you attended any IELTS preparation courses before this course?

□ Yes □ No

If yes, how many preparation courses have you taken? _______

Course name _____________________________

Course duration _____month/_____days

Course 2: course name _____________________________

Course duration _____month/_____days

If more, please specify: ________________________________

15. Using IELTS band score (1 being the lowest, 9 being the highest)

How would you rank your listening skill at the moment? Band _________

How would you rank your reading skill at the moment? Band __________

How would you rank your writing skill at the moment? Band __________

How would you rank your speaking skill at the moment? Band _________

16. Have you been taking any extra-curriculum non-test oriented English course(s) outside classroom?

□ No

□ Yes

If Yes, what are these courses?

____

17. Have you taken other English proficiency tests?

□ College English test-4

□ College English test-6

□ TOEFL

□ others (please specify_____)

What are the score(s)?

CET-4: _________________

CET-6: _________________

TOEFL: ________________

Others: _________________

Thank you very much for your time!

Appendix 4 Information page and consent form for Quasi-experiment 2

Name of Researchers	**Industrial partner**
Ms Ruolin HU Dr Danijela Trenkic	Dr Burr Settles, Duolingo

Title of Study
Chinese students' English proficiency and academic performance

Brief Description of Study
The aim of this study is to investigate the relationship between Chinese students' English proficiency and academic performance.

What does this mean for me?
You will be asked to answer some questions about your language learning history and complete some language tests in English. The session will take about 60 minutes.

Due to the design of this study, we will need to know your IELTS results and all your module marks. You need to be happy to show us your IELTS reports and to allow the university to share your module marks with us. The information you provide will only be used for research purposes and can in no way impact on your academic outcomes. You should also know that only aggregated, collective results will be reported in any presentations and publications, which means that your data will not be identifiable.

One of the language tests you will be asked to complete is called Duolingo English Test (DET). The test consists of listening, writing, speaking and vocabulary exercises, and takes about half an hour to complete. It is a computer-based, on-line test. After completing the test, you will receive an official document that certifies your proficiency in English, worth US $49.

In order to issue you with a certificate, the company administering the test ('Duolingo') needs to be sure that you have not received help from anyone, and that you are who you say you are. For this reason, your performance on this test will be video recorded through the computer camera and you will be asked to verify your identity by showing your passport. You can still sit the test if you don't have your passport with you or do not wish to share it with Duolingo, but you will not receive a certificate.

Where will the research session take place?
It will take place on campus at the University of York.

Who will run the research sessions?
Ruolin Hu

What will happen to the information you give to Duolingo?
Duolingo will keep a copy of your ID for 90 days, in case of any disputes. After that time, it will be destroyed. This information will not be held by the Researchers (Ruolin Hu and Danijela Trenkic) at any point.

Duolingo will also keep a video record of the session ('Testing video') to ensure that you were not helped by another person, and to evaluate the usefulness, accuracy and other aspects of the test. Duolingo will not share your personal information with third-parties unless you ask them to, or it is required by law.

Duolingo will share the complete and accurate results of Duolingo English Tests with the Researchers. They will not share your result with anyone else without your direction, but may use anonymized results to improve the examination and for research and analysis.

Information associated with the Duolingo English Test, including examination results and your Testing Video, may be collectively deleted from your Duolingo account, but anonymized examination data, including your examination results and Testing Video, may be kept indefinitely by Duolingo to improve the examination and for research and analysis.

You can read Duolingo's Privacy policy in full at: https://www.duolingo.com/privacy

What will happen to the information you share with the Researchers?
The Researchers will keep all the information about participants in this study confidential and data will be anonymised. To ensure confidentiality you will be randomly assigned an ID number. When the Researchers receive the results from Duolingo, they will store them under this ID number, and it will also be the only form of identification included on any database and paper based tasks. The researchers will not keep the Testing Video.

Any information that identifies you will be stored separately from the data, in a password-protected file that only the researchers can access, and will be kept for five years after the publication of research. You can stop your participation at any point during data collection and are free to withdraw from the study at any point before June 2017 by writing to us (danijela.trenkic@york.ac.uk; rh783@york.ac.uk).

Should you pass on any information that causes concern about your wellbeing or the wellbeing of others, we might be obliged to disclose it.

If I have any queries, concerns or complaints, whom shall I contact?
If you have any questions about the study that you would like to ask before giving consent or after the data collection, please feel free to contact the researchers by email (rh783@york.ac.uk; danijela.trenkic@york.ac.uk) or the Chair of Ethics Committee via email (education-research-administrator@york.ac.uk).

*****Please keep this page for your information*****

Consent form

Title of the study: *Chinese students' English proficiency and academic performance*

Researchers: Ruolin Hu, Danijela Trenkic

Please tick each box if you are happy to take part in this research.

I am a native speaker of Mandarin Chinese.	
I have read the information concerning the research that I am being asked to take part in, and I have had the opportunity to ask questions.	
I am happy to disclose my IELTS results to Ruolin Hu and Danijela Trenkic.	
I agree that the university can disclose my module marks to Ruolin Hu and Danijela Trenkic.	
I am happy to take the Duolingo English Test. I understand that by doing so I will share my information with Duolingo under their privacy policy, which has been explained to me and I have had an opportunity t check.	
I understand that researchers will keep my data for 5 years after the research is published.	
I understand that data could be used for future analyses.	
I understand that only aggregated, collective results will be reported in any presentations and publications, which means that my data will not be identifiable.	
I am happy to take part in this study.	
I understand that I may withdraw at any time before June 2017 by writing to rh783@york.ac.uk or danijela.trenkic@york.ac.uk	

Name of participant ……………………………………………………………………

Department……………….. …………………………………………………………

University of York username (email address)………………………………………

Signature ……………………………………………………………………………

Date …..…………………………………………………………………………..

Appendix 5 Questionnaire for Quasi-experiment 2

Participate ID: _____

1. Year of birth: _____(year) /_____(month)
2. Gender: Female / Male
3. Are you an undergraduate or a postgraduate?

   □ undergraduate

   □ postgraduate by taught

   □ postgraduate by research
4. What is the subject of your study?

   ____
5. How old were you when you started learning English?

   ____
6. How many languages do you speak?

   ____
7. What are these languages?

   ____
8. What language(s) do you speak at home?

   ____
9. Have you studied in any other English-speaking countries (including UK) before?

   Yes / No

   If Yes, what are these countries?

   ____

   How long did you stay?

   ____
10. Did you take any pre-sessional English courses?

    Yes / No

    If yes, how long did your pre-sessional English course last? _____ (weeks)
11. How many IELTS tests did you take before your study in York?

    ____
12. What was your final IELTS result?

    Overall ____ Listening_____ Reading_____ Writing_____ Speaking_____

    When did you take it?

    ____(year) / ____(month)

13. What was your first IELTS result?

Overall ____ Listening_____ Reading_____ Writing_____ Speaking_____

When did you take it?

____(year) / ____(month)

14. Did you take any more IELTS in-between?

Yes / No

If Yes, what were the results?

Overall ____ Listening_____ Reading_____ Writing_____ Speaking_____

When did you take it?

____(year) / ____(month)

15. Did you attend any IELTS preparation course?

Yes or No

If Yes,

When did you attend this course? _____(year) / ______ (month)

Where did you attend this course? _____ (country) / ________ (city)

How long did this course last? ______ (weeks)

16. Did you attend more than one IELTS preparation course?

Yes or No

If yes,

When _____(year) / ______ (month)

Where _____ (country) / ________ ______ (weeks)

Appendix 6 C-test for Quasi-experiment 2

Participate ID: _____

1. Once upon a time, a child’s bedroom had little more than a toy box, a bookshelf, and a few posters. Today i_______ looks mo_______ like miss_______ control a_________ Houston. Comp_______, mobile pho_______, televisions, DVD pla_______, game mach_______, and ot_______ 21st cen_______ toys fi_______ the ro_______, and of_______ make t_______ child’s bed_______ the mo_______ expensive i_______ the ho_______. Britain’s 8- to 16- year-olds have bedroom possessions worth an average of £3,300.

2. It was the American sitcom that defined a generation — and introduced one of the world’s most famous haircuts. The six st_______ of *Friends*, am_______ the longest_______, most succe_______ series ev_______ to h_______ the sm_______ screen, we_______ their sepa_______ ways af_______ 237 epis_______ and a dec_______ together a_______ flatmates, sha_______ the tri_______ of th_______ lives, lo_______, and car_______ in a tre_______ New York apartment. The last episode was seen by an estimated world audience of over 100 million viewers.

3. At only 28, Jamie Oliver is now an extremely successful and well-known chef, with his own acclaimed restaurant in the centre of London. His ri_______ to fa_______ and for_______ came ea_______ and swi_______. By t_______ age o_______ eight, he h_______ already sta_______ cooking a_______ his par_______ pub. It w_______ an ea_______ way t_______ earn a b_______ of poc_______ money! Af_______ two ye_______ in cate_______ college, a_______ some ti_______ spent in France, he started working in restaurants. He worked under 3 famous chefs in London before he was spotted by a TV producer at 21 and his life changed.

4. The drought is Australia’s worst in a century. Economist Justin Smirk sa_______ that produ_______ of wh_______ and ri_______ might fa_______ by 20 per_______. He beli_______ the dro_______ will cha_______ agricultural prac_______ forever. ‘It wo_______ be ea_______ to gr_______ crops su_______ as ri_______ in t_______ future. W_______ might ha_______ to st_______ farming i_______ very d_______ areas.’ Despite the sign, the government refuses to blame the drought on climate change, but scientist Peter Cullen is more certain.

5. The first Diana conspiracy site appeared on the Internet in Australia only hours after her death. Since th_______ an esti_______ 36,000 Diana consp_______ websites ha_______ been s_______ up – breath_______ by any_______ standards. Hypot_______ range fr_______ pure James Bond ('it w_______ all a_______ MI6 pl_______ to pro_______ the mona_______') to fa_______ ('it w_______ a fien_______ murder pl_______ thought u_______ by t_______ world's flor_______ to se_______ lots of flowers'). And most popular of all, Diana, isn't dead after all - that terrible car crash in Paris was an elaborate hoax to enable the Princess and her boyfriend to fake their own deaths so that they could live in blissful isolation for the rest of their lives.

Reference

Al-Musawi, N. M., & Al-Ansari, S. H. (1999). Test of English as a Foreign Language and First Certificate of English tests as predictors of academic success for undergraduate students at the University of Bahrain. *System*, 27, 389–399.

Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. TESOL Quarterly, *13*(2), 219-227.

Alderson, J. C. & A. H. Urquhart. (1983). The effect of student background discipline on comprehension: A pilot study. In A. Hughes and D. Porter (Eds.): *Current Developments in Language Testing* (pp. 121-128). London: Academic Press.

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses A study of washback. *Language Testing*, *13*(3), 280-297.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, *14*(2), 115-129.

Anderson J., Clapham.C., Wall. D (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Andrade, M. S. (2006). International students in English-speaking universities Adjustment factors. *Journal of Research in International Education*, *5*(2), 131-154.

Andringa, S., Olsthoorn, N., & van Beuningen, C. (2012). Determinants of success in native and non-native listening comprehension: An individual differences approach. *Language*. *62*(2). 49-78.

Angelova, M., & Riatzatseva, A. (1999). If you don't tell me, how can I know: A case study of four international students learning to write the U.S. way. *Written Communication, 16,* 491-525.

Arcuino, C. L. T. (2013). *The relationship between TOEFL, IELTS and academic success of international MA students* (Unpublish doctoral thesis). Colorado State University, USA.

August, D., Carlo, M., Dressler, C., & Snow, C. (2005). The Critical Role of Vocabulary Development for English Language Learners. *Learning Disabilities Research & Practice: A Publication of the Division for Learning Disabilities, Council for Exceptional Children*, *20*(1), 50–57.

Ayers, J. B., & Peters, R. M. (1977). Predictive Validity of the Test of English as a Foreign Language for Asian Graduate Students in Engineering, Chemistry, or Mathematics. *Educational and Psychological Measurement*, 37(2), 461–463.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford Uniersity Press.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *Tesol Quarterly*. *16*(4). 449-465.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.

Baddeley, A. D., Emslie, H., & Nimmo-Smith, I. (1992). *The speed and capacity of language processing (SCOLP) test*. Bury St Edmunds: Thames Valley Test Company.

Bagheri, M., & Karami, S. (2014). The Effect of Explicit Teaching of Listening Strategies and Gender on EFL Learners' IELTS Performance. *Journal of Language Teaching and Research*, 5(6). 1387-1392.

Baker, E. (1991). Alternative assessment and national policy. In *National Research Symposium on Limited English Proficient Students' Issues: Focus on Evaluation and Measurement,* Washington, DC.

Barnett, A. G., Van Der Pols, J. C., & Dobson, A. J. (2004). Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology*, *34*(1), 215–220.

Bangert-Drowns, R.L., Kalik, K.A. & Kalik, C.L.C. (1983). Coaching And Its Effects On Test Preparation, *Review Of Educational Research*, *53*(4), 571–588.

Bayliss, D., & Raymond, P. M. (2004). The Link Between Academic Success and L2 Proficiency in the Context of Two Professional Programs. *The Canadian Modern Language Review / La Revue Canadienne Des Langues Vivantes*, *61*(1), 29–51.

Beishline, M. J., & Holmes, C. B. (1997). Student preferences for various teaching styles. *Journal of Instructional Psychology, 24*(2), 95-99.

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 Listening Framework: A Working Paper*. Educational Testing Service Princeton, NJ.

Biber, B. & Gray, B. (2013). Discourse Characteristics of Writing and Speaking Task Types on the TOEFL iBT Test: A Lexico-Grammatical Analysis. Retrieved February 19, 2017 from https://www.ets.org/Media/Research/pdf/RR-13-04.pdf

Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. *Research Methodology in Second-Language Acquisition*, 245–261. edited by Elaine E. Tarone, Susan M. Gass, Andrew D. Cohen

Bool, H., Dunmore, D., Tonkyn, A., Schmitt, D., & Ward-Goodbody, M. (2003). *The BALEAP guidelines on English language proficiency levels for international applicants to UK universities.* London: British Association of Lecturers in English for Academic Purposes.

Brentzel, J. & Settles, B. (2017). The Duolingo English Test — Design, Validity, and Value. Retrieved October 18, 2016 from https://s3.amazonaws.com/duolingo-papers/other/DET_ShortPaper.pdf

British Council. (2013). *The English effect*. Retrieved September 17, 2017 from https://www.britishcouncil.org/sites/default/files/english-effect-report-v2.pdf

Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks. *ETS Research Report Series*, *2005*(1), i – 157.

Brown, J. D. (1998). Does IELTS preparation work? An application of the context-adaptive model of language program evaluation. *IELTS Research Reports*, 1, 20-37.

Brown, J. D. (2000). What is construct validity. Retrieved January 20, 2016 from http://hosted.jalt.org/test/bro_8.htm

Buck, G. (1988). Testing listening comprehension in Japanese university entrance examinations. *JALT Journal*, *10*(1), 15–42.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 Speaking Framework: A Working Paper*. Educational Testing Service Princeton, NJ.

Canale, M., & Swain, M. (1981). A Theoretical Framework for Communicative Competence. Retrieved April 11, 2015 from https://eric.ed.gov/?id=ED223105

Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In: *Testing the English proficiency of foreign students* (pp. 313– 320). Washington, DC: Centre for Applied Linguistics.

Cattell, R. B., & Horn, J. L. (1978). A check on the theory of fluid and crystallized intelligence with description of new subtest designs. *Journal of Educational Measurement*, *15*(3), 139–164.

Celestine, C., Cheah, S. M., & Others. (1999). The effect of background disciplines on IELTS scores. *IELTS Research Reports, 1999*(2), 36-61.

Chalmers, D., & Volet, S. (1997). Common Misconceptions about Students from South-East Asia Studying in Australia. *Higher Education Research & Development*, *16*(1), 87–99.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*, 254-272.

Charge, N., & Taylor, L. B. (1997). Recent developments in IELTS. *ELT Journal*, *51*(4), 374–380.

Chen, Y., & Sun, C. (2006). Language proficiency and academic performance. In Proceedings of *the 11th Conference of Pan-Pacific Association of Applied Linguistics*, South Korea: Kangwon National University.

Cheng, L. (1999). Changing assessment: Washback on teacher perceptions and actions. *Teaching and Teacher Education*, *15*(3), 253–271.

Cheng, L. (2008). The key to success- English language testing in China. *Language Testing*, *25*(1), 15–37.

Chihara, T., Oller, J., Weaver, K., & Chavez-Oller, M. A. (1977). Are cloze items sensitive to constraints across sentences? *Language Learning*, *27*(1), 63–70.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT scores to academic performance Some evidence from American universities. *Language Testing*, *29*(3), 421-442

Clarke, M. A. (1980). The “short-circuit” hypothesis of ESL reading—or when language competence interferes with reading performance. *Modern Language Journal*, *64*, 203–209.

Cohen, A. D. (2014). Strategies in Learning and Using a Second Language. Routledge.

Cohen, A. D., & Upton, T. A. (2006). Strategies In Responding To The New TOEFL Reading Tasks. *ETS Research Report Series*, *2006*(1), i – 162.

Colom, R., Abad, F. J., Rebollo, I., & Chun Shih, P. (2005). Memory span and general intelligence: A latent-variable approach. *Intelligence, 33,* 623–642.

Colom, R., Flores-Mendoza, C., Quiroga, M. Á., & Privado, J. (2005). Working memory and general intelligence: The role of short-term storage. *Personality and Individual Differences, 39,* 1005–1014.

Colom, R., Flores-Mendoza, C., Quiroga, M. Á., & Privado, J. (2005). Working memory and general intelligence: The role of short-term storage. *Personality and Individual Differences*, *39*(5), 1005–1014.

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12,* 769–786.

Cooper, B. J. (2004). The enigma of the Chinese learner. *Accounting Education*, *13*(3), 289–310.

Cotton, F., Conrow, F., & Others. (1998). An investigation of the predictive validity of IELTS amongst a group of international students studying at the University of Tasmania. *IELTS Research Reports, 1998* (*1*), 72-115.

Cranbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New York: Harper & ReM.

Craven, E. (2012). The quest for IELTS Band 7.0 Investigating English language proficiency development of international students at an Australian university. *IELTS Research Reports, 2012 (13), 1-61.*

Crawford, I., & Wang, Z. (2015). The effect of work placements on the academic performance of Chinese students in UK higher education. *Teaching in Higher Education, 20*(6), 569-586

Crawford, T. (2015). What students in China have taught me about U.S. college admissions. Retrieved July 29, 2016 from https://www.theatlantic.com/education/archive/2015/01/what-students-in-china-have-taught-me-about-us-college-admissions/384212/

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 443-507). Washington, OC: American Council on Education.

Crowell, T. A., Vanderploeg, R. D., Small, B. J., Graves, A. B., & Mortimer, J. A. (2002). Elderly norms for the Spot-the-Word test. *Archives of clinical neuropsychology, 17*(2), 123-130.

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing*, *21*(2), 107–145.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 Writing Framework: A Working Paper*. In *TOEFL Monograph Series Report No. 18*. Princeton, NJ: Educational Testing Service.

Daller, M. H., & Phelan, D. (2013). Predicting international student study success. *Applied Linguistics Review*, *4*(1), 173–193.

Daller, M. H., & Xue, H. (2009). Vocabulary knowledge and academic success: A study of Chinese students in UK higher education. In Richars, B. Daller, M.H., David, D. M., Meara, P., Milton, J. & Treffers-Daller, J. (Eds) *Vocabulary Studies in First and Second Language Acquisition* (pp. 179-193). Palgrave Macmillan UK

Darnell, D. K. (1968). The Development of an English Language Proficiency Test of Foreign Students, Using a Clozentropy Procedure. Final Report. Retrieved September 28, 2018 from https://eric.ed.gov/?id=ED024039

Desveaux, S. (2018). *Cambridge Assessment English: Guided learning hours.* Retrieved September 9, 2017, from https://support.cambridgeenglish.org/hc/en-gb/articles/202838506-Guided-learning-hours

Diamond, J.J. & Evans. W.J. (1972). An investigation of the cognitive correlates of test-wiseness, *Journal Of Educational Measurement*, *9*(2), 145–150.

Dooey, P. & Oliver, R. (2002). An investigation into the predictive validity of the IELTS Test as an indicator of future academic success. Retrieved August 9, 2018 from https://www.researchonline.mq.edu.au/vital/access/manager/Repository/mq:35533

Dörnyei, Z., & Katona, L. (1992). Validation of the C-test amongst Hungarian EFL learners. *Language Testing, 9*(2), 187-206.

Dörnyei, Z., & Taguchi, T. (2009). *Questionnaires in second language research: Construction, administration, and processing*. Routledge.

Douglas, D. (1997). Testing speaking ability in academic contexts: Theoretical considerations. *TOEFL Monograph Series Report No. 8*. Princeton, NJ: Educational Testing Service.

Du-Babcock, B. (2002). Teaching a large class in Hong Kong. *Business Communication Quarterly*. Retrieved August 11, 2017 from https://www.researchgate.net/profile/Bertha_Du-Babcock/publication/238335518_Teaching_a_Large_Class_in_Hong_Kong/links/53f738be0cf22be01c454e76.pdf

Duolingo English Test: Score interpretation. Retrieved September 19, 2018, from https://englishtest.duolingo.com/scores

Duolingo English test. (2018). Interpretation of DET Scores. Retrieved February 20, 2017 from https://englishtest.duolingo.com/scores

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing, 23*(3), 290-325.

Edwards, V., & Ran, A. (2006). Meeting the needs of Chinese students in British Higher Education. Retrieved March 20, 2016 from https://s3.amazonaws.com/academia.edu.documents/280512/MeetingTheNeeds.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1538328876&Signature=x8tolgZv5n7aV59LQ76zgwD%2FRV8%3D&response-content-

disposition=inline%3B%20filename%3DMeeting_the_needs_of_Chinese_students_in.pdf

Elder, C., & O'Loughlin, K. (2003). Investigating the relationship between intensive English language study and band score gain on IELTS. *IELTS research repor*ts, *4*(6), 207-254.

Elley, W. B. (1992). *How in the world do students read?* Hamburg, Germany: International Association for the Evaluation of Educational Achievement.

Enright, M., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 Reading Framework: A Working Paper*. Educational Testing Service Princeton, NJ.

Erfani, S. S. (2012). A comparative washback study of IELTS and TOEFL iBT on teaching and learning activities in preparation courses in the Iranian context. *English Language Teaching, 5*(8), 185-195.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge- An empirical validation study. *Applied linguistics, 27*(3), 464-491.

ETS. (2011). Validity evidence supporting interpretation and use of TOEFL iBT scores. *TOEFL iBT Insight Series 1 Volume 4*. Educational Testing Service Princeton, NJ.

Favreau, M., & Segalowitz, N. S. (1983). Automatic and controlled processes in the first- and second-language reading of fluent bilinguals. *Memory & Cognition*, *11*(6), 565–574.

Feast, V. (2002). *The impact of IELTS scores on performance at university International. Education Journal. 3*(4), *70-85.*

Feng, J. H. (1991). The adaptation of students from the People's Republic of China to an American academic culture. Retrieved August 19, 2017 from https://files.eric.ed.gov/fulltext/ED329833.pdf

Fenton, N., Neil, M., & Constantinou, A. (2015). Simpson's Paradox and the implications for medical trials. Retrieved February 12, 2019, from https://pdfs.semanticscholar.org/43c8/01351e3b17e84a745a0adc4462a84771ed00.pdf

Field, J. (2008). Bricks or Mortar: Which Parts of the Input Does a Second Language Listener Rely on? *TESOL Quarterly*, *42*(3), 411–432.

Frankel, S. (1983). *Study Of Test Burdens At The Elementary And Intermediate Schools.* Retrieved August 18, 2016 from https://files.eric.ed.gov/fulltext/ED251486.pdf

Frederiksen, J. R., & Collins, A. (1989). A Systems Approach to Educational Testing. *Educational Researcher*, *18*(9), 27–32.

Gan. Z. (2009). IELTS Preparation Course and Student IELTS Performance. *RELC journal, 40*(1), 23-41.

George, P. (1985). Coaching For Tests: A Critical Look At The Issues, *Curriculum Review*, *25*(1), 23–26.

Gignac, G. E., & Weiss, L. G. (2015). Digit Span is (mostly) related linearly to general intelligence- Every extra bit of span counts. *Psychological assessment, 27*(4), 1312-1323.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237–288.

Grabe, W. (1991). Current Developments in Second Language Reading Research. *TESOL Quarterly*, *25*(3), 375–406.

Green, A. (2005). EAP study recommendations and score gains on the IELTS Academic Writing test. *Assessing Writing*, 10(1), 44-60.

Green, A. (2006). Washback to the learner and teacher perspectives on IELTS preparation course expectations and outcomes. *Assessing writing*, *11*(2), 113-134.

Green, A. (2007). Washback to learning outcomes A comparative study of IELTS preparation and university presessional language courses. *Assessment in Education*, *14*(1), 75-97.

Gu, Q. (2009). Maturity and Interculturality: Chinese students' experiences in UK higher education. *European Journal of Education*, *44*(1), 37–52.

Gu, Q. & Brooks, J. (2008). Beyond the Accusation of Plagiarism. *System*, 36, 337–352.

Gu, Q., & Maley, A. (2008). Changing Places: A Study of Chinese Students in the UK. *Language and Intercultural Communication*, *8*(4), 224–245.

Hagedorn, L. S. (2012). International Graduate Students' Academic Performance- What Are the Influencing Factors? *Journal of International Students*, 2, 135–143.

Hamid, M. O. (2016). Policies of global English tests: test-takers' perspectives on the IELTS retake policy. *Discourse: Studies in the Cultural Politics of Education*, *37*(3), 472–487.

Hastings, A. J. (2002). In defense of C-Testing. In R. Grotjahn (Ed.). *Der C-Test: Theoretische Grundlagen und Praktische Anwundungen [The C-Test: Theoretical foundations and practical applications]*. Vol. 4. (pp. 11-25). Bochum: AKS- Verlag.

Hayes, B., & Read, J. (2004). IELTS test preparation in New Zealand: Preparing students for the IELTS Academic Module. In: L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 97–112). Mahwah, NJ: Lawrence Erlbaum.

He, Y., & Banham, H. (2009). International Student Academic Performance: Some Statistical Evidence And Its Implications. *American Journal of Business Education*, *2*(5), 89–100.

Hermann, F. (2003). Differential effects of reading and memorization of paired associates on vocabulary acquisition in adult learners of English as a second language. *TESL-EJ*, *7*(1), 1–16.

HESA (2018). Higher education statistics for the United Kingdom 2016–17. Higher Education Statistics Agency. https://www.hesa.ac.uk/

Higher Education Policy Institute & Kaplan International Pathways. (2018). The costs and benefits of international students by parliamentary constituency. Retrieved September 19, 2018 from https://www.hepi.ac.uk/wp-content/uploads/2018/01/Economic-benefits-of-international-students-by-constituency-Final-11-01-2018.pdf

Hill, K., Storch, N. &Lynch, B. (1999). A comparison of IELTS and TOEFL as predictors of academic success. *IELTS Research Reports,* 1999(2), 62-73.

Ho, D. Y. F., & Spinks, J. A. (1985). Multivariate Prediction of Academic Performance by Hong Kong University students. *Contemporary Educational Psychology*, *10*, 249–259.

Holmes, P. (2004). Negotiating differences in learning and intercultural communication ethnic Chinese students in a New Zealand university. *Business Communication Quarterly*, *67*(3), 294-307.

Hughes, A. (1989) *Testing for Language Teachers.* Cambridge: Cambridge University Press.

Hulstijn, J. H. (2015). *Language Proficiency in Native and Non-native Speakers: Theory and research*. Amsterdam: John Benjamins Publishing Company.

Hulstijn, J. H., Schoonen, R., De Jong, N. H., Steinel, M. P., & Florijn, A. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing*, *29*(2), 203–221.

Huntley, H. S. (1993). Adult international students: Problems of adjustment. Retrieved June 19, 2017 from https://files.eric.ed.gov/fulltext/ED355886.pdf

Hwang, K.Y., & Dizney, H. F. (1970). Predictive Validity of the Test of English as a Foreign Language for Chinese Graduate Students at an American University. *Educational and Psychological Measurement*, *30*(2), 475–477.

Hymes, D. (1972). On communicative competence. *Sociolinguistics*, *26*, 269–293.

Iannelli, C., & Huang, J. (2013). Trends in participation and attainment of Chinese students in UK higher education. *Studies in Higher Education*, 39(5), 805-822

IELTS Assessment criteria. (2018). https://takeielts.britishcouncil.org/find-out-about-results/ielts-assessment-criteria

IELTS Bandscore descriptor. (2018). https://takeielts.britishcouncil.org/find-out-about-results/understand-your-ielts-scores

IELTS Demographic data. (2018). https://www.ielts.org/teaching-and-research/demographic-data

IELTS Guide for educational institutions, governments, professional bodies and commercial organisations. (2015). https://www.ielts.org/-/media/publications/guide-for-institutions/ielts-guide-for-institutions-2015-uk.ashx?la=en

IELTS Home of the English Language Test (2018). https://www.ielts.org/

IELTS Scoring in detail. (2018). https://www.ielts.org/ielts-for-organisations/ielts-scoring-in-detail

IELTS test format. (2018). https://www.ielts.org/about-the-test/test-format

IELTS test performance. (2017). Retried July 12, 2018 from https://www.ielts.org/teaching-and-research/test-performance

IELTS test taker performance. (2018). https://www.ielts.org/teaching-and-research/test-taker-performance

IELTS. (2018). Understanding the Listening test. Retrieved September 25, 2018 from https://takeielts.britishcouncil.org/prepare-test/understand-test-format/listening-test

IELTS. (2018). Understanding the Reading test. Retrieved September 25, 2018 from https://takeielts.britishcouncil.org/prepare-test/understand-test-format/reading-test

IELTS. (2018). Understanding the Speaking test. Retrieved September 25, 2018 from https://takeielts.britishcouncil.org/prepare-test/understand-test-format/speaking-test

IELTS. (2018). Understanding the Writing test. Retrieved September 25, 2018 from https://takeielts.britishcouncil.org/prepare-test/understand-test-format/writing-test

Inhoff, A. W., Pollatsek, A., Posner, M. I., & Rayner, K. (1989). Covert attention and eye movements during reading. *The Quarterly Journal of Experimental Psychology Section A*, *41*(1), 63–89.

Ishikawa, H. & Settles, B. (2016). The Duolingo English and academic English. Retrieved January 17, 2017 from https://s3.amazonaws.com/duolingo-papers/reports/DRR-16-01.pdf

Issitt, S. (2008). Improving scores on the IELTS speaking test. *ELT journal*, *62*(2), 131-138.

Jakeman, V. & McDowell, C. (1996). *Cambridge practice tests for IELTS 1*. Cambridge: Cambridge University Press.

Jensen, E. D., & Vinther, T. (2003). Exact repetition as input enhancement in second language acquisition. *Language Learning*, *53*(3), 373–428.

Jochems, W., Snippe, J., Smid, H. J., & Verweij, A. (1996). The academic progress of foreign students- study achievement and study behaviour. *Higher Education*, *31*(3), 325-340.

Johnson, P. (1988). English Language Proficiency and Academic Performance of Undergraduate International Students. *TESOL Quarterly*, *22*(1), 164–168.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological Review*, *99*(1), 122–149.

Kember, D. (1996). The intention to both memorise and understand: Another approach to learning? *Higher Education*, *31*(3), 341–354.

Kerstjens, M., & Nery, C. (2000). Predictive validity in the IELTS test A study of the relationship between IELTS scores and students' subsequent academic performance. *IELTS research reports*, *2000*(3), 85-108.

Killian, L. (1992). A School District Perspective On Appropriate Test-Preparation Practices: A Reaction To Popham's Proposals, *Educational Measurement: Issues And Practice*, *11*(4), 13–15, 126.

Kim, E., Newton, F. B., Downey, R. G., & Benton, S. L. (2010). Personal factors impacting college student success: Constructing college learning effectiveness inventory (CLEI). *College Student Journal, 44*(1), 112-125.

Kirkpatrick, R., & Zang, Y. (2011). The Negative Influences of Exam-Oriented Education on Chinese High School Students: Backwash from Classroom to Child. *Language Testing in Asia*, *1*(3), 36.

Klein-Braley, C. (1994). *Language testing with the C-Test: A linguistic and statistical investigation into the strategies used by C-Test takers, and the prediction of C- Test difficulty*. Unpublished Higher Thesis, Department of Linguistics and Literature, University of Duisburg, Germany.

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal*. Language Testing, 14*(1), 47-84.

Klein-Braley, C. & Raatz, U. (1984). A survey of research on the C-test. *Language Testing, 1,* 134-146.

Knight, S. (1994). Dictionary Use While Reading: The Effects On Comprehension and Vocabulary Acquisition For Students Of Different Verbal Abilities. *The Modern Language Journal*, *78*(3), 285–299.

Koda, K. (1996). L2 word recognition research: A critical review. *The Modern Language Journal*, *80*, 450-460.

Koretz, D. (2008). *Measuring up*. Cambridge, MA: Harvard University Press.

Kuh, G.D., Cruce, T.M., Shoup, R., Kinzie, J., & Gonyea, R.M. (2008). Unmasking the effects of student engagement on first-year college grades and persistence. *The Journal of Higher Education, 79*(5), 540-563.

Kuo, J., Hagie, C., & Miller, M. T. (2004). Encouraging college student success: The instructional challenges, response strategies, and study skills of contemporary undergraduates. *Journal of Instructional Psychology*, *31*, 60-67.

Lado, R. (1961). Language Testing: The Construction and Use of Foreign Language Tests. A Teacher's Book. Retrieved August 8, 2015 from https://eric.ed.gov/?id=ED032799

Larmer, B. (2014). Inside a Chinese Test-Prep Factory. *The New York Times*. Retrieved August 8, 2015 from https://www.nytimes.com/2015/01/04/magazine/inside-a-chinese-test-prep-factory.html

Laufer, B. (1998). The Development of Passive and Active Vocabulary in a Second Language: Same or Different? *Applied Linguistics*, *19*(2), 255–271.

Laufer, B., & Goldstein, Z. (2004). Testing Vocabulary Knowledge: Size, Strength, and Computer Adaptiveness. *Language Learning*, *54*(3), 399–436.

Lebcir, R. M., Wells, H., & Bond, A. (2008). Factors affecting academic performance of international students in project management courses A case study from a British Post 92 University. *International Journal of Project Management*, *26*(3), 268–274.

Lei, L. (2008). Validation of the C-Test amongst Chinese ESL Learners. *The Journal of Asia TEFL, 5*(2), 117-140

Li, G., Chen, W., & Duanmu, J.-L. (2010). Determinants of International Students' Academic Performance. *Journal of Studies in International Education*, *14*(4), 389–405.

Li, L. (2013). 考生出国前要考3次雅思 6成考生期望雅思分数过7 [Chinese candidates need to sit 3 IELTS before studying abroad and 60% of them are expected to achieve an overall of at least 7]. Retrieved 8th February, 2015 from http://learning.sohu.com/20130209/n365924751.shtml

Ligon, G., & Jones, P. (1981). *Preparing Students For Standardized Testing: One District's Perspective.* Retrieved August 9, 2017, from https://eric.ed.gov/?id=ED218319

Liu, O. L. (2014). Investigating the Relationship Between Test Preparation and TOEFL iBT® Performance. *ETS Research Report Series*, *2014*(2), 1–13.

Ma, J., & Cheng, L. (2016). Chinese Students' Perceptions of the Value of Test Preparation Courses for the TOEFL iBT: Merit, Worth, and Signicance. *TESL Canada Journal*, *33*(1), 58–79.

Ma, W. (2014). Chinese international undergraduate students at a US university- A mixed methods study of first-year academic experiences and achievement. Unpublished doctoral dissertation. The University of Utah, Utah.

Maclean, M., & d'Anglejan, A. (1986). Rational Cloze And Retrospection: Insights into First and Second Language Reading Comprehension. *Canadian Modern Language Review*, *42*(4), 814–826.

Makepeace, E. & Baxter, A. (1990). *Overseas students and examination failure: a national study. Journal of International Education*, *1*(1), 36–48.

Matoush, M. M., & Fu, D. (2012). Tests of English Language as significant thresholds for college-bound Chinese and the washback of test-preparation. *Changing English*, *19*(1), 111-121.

McKenzie, K., & Schweitzer, R. (2001). Who succeeds at university? Factors predicting academic performance in first year Australian university students. *Higher Education Research & Development*, *20*(1), 21-33.

Meara, P. (2002). The rediscovery of vocabulary. *Second Language Research*, *18*(4), 393–407.

Mecartty, F. H. (2000). Lexical and grammatical knowledge in reading and listening comprehension by foreign language learners of Spanish. *Applied Language Learning*, *11*(2), 323–348.

Mehrens, W. A. (1989). *Preparing students to take standardized achievement tests*. ERIC Clearinghouse.

Messick, S. (1981). Issues of effectiveness and equity in the coaching controversy: implications for educational and testing practice. *ETS Research Report Series*. Retrieved August 8, 2016 from https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1981.tb01254.x

Messick, S. (1982). The values of ability testing: Implications of multiple perspectives about criteria and standards. *Educational Measurement: Issues and Practice*, *1*(3), 9–12.

Messick, S. (1987). Validity. *ETS Research Report Series*. Retrieved February 19, 2018 from http://onlinelibrary.wiley.com/doi/10.1002/j.2330-8516.1987.tb00244.x/full

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256.

Messick, S. (1998) Test Validity: A Matter of Consequence. *Social Indicators Research*, *45*(1), 35–44.

Mickan, P., & Motteram, J. (2009). The preparation practices of IELTS candidates Case studies. *IELTS Research Reports*, 10, 1-39.

Migration Advisory Committee. (2018). Impact of international students in the UK. London: Migration Advisory Committee.

Moore, T., Morton, J., & Price, S. (2012). 4 Construct validity in the IELTS Academic Reading test: A comparison of reading requirements in IELTS test items and in university study. *IELTS Collected Papers*, 2, 120–211.

Mori, S. C. (2000). Addressing the Mental Health Concerns of International Students. *Journal of Counseling & Development*, *78*(2), 137–144.

Morrison, J., Merrick, B., Higgs, S., & Métais, J. L. (2005). Researching the performance of international students in the UK. *Studies in Higher Education*, *30*(3), 327–337.

Naiman, N. (1974). *The use of elicited imitation in second language acquisition research*. Ontario Institute for Studies in Education.

Newton, P., & Shaw, S. (2014). *Validity in Educational and Psychological Assessment*. SAGE.

Oller, J. W. (1976). Evidence for a general language proficiency factor: An expectancy grammar. *Die Neueren Sprachen*, *75*(2), 165–174.

Oller, J.W. (1973). Cloze tests of second language proficiency and what they measure.

Oxford Economics. (2014). The impact of independent schools on the British community. Retrieved September 1, 2018 from https://www.isc.co.uk/media/2588/2014_economicimpact_report_isc.pdf

Paton, M. J. (2007). Why international students are at greater risk of failure. *International Journal of Diversity in Organizations, Communities & Nations*, *6*(6), 101–112.

Pelletier, C. (2003). The experiences of international students in UK higher education: A review of unpublished research. Retrieved October 11, 2016 rom http://www.ukcisa.org.uk/about/material_media/archive.php

Pennycook, A. (1996). Borrowing Others' Words: Text, Ownership, Memory, and Plagiarism. *TESOL Quarterly*, *30*(2), 201.

Perani, D., & Abutalebi, J. (2005). The neural basis of first and second language processing. *Current Opinion in Neurobiology, 15(*2), 202–206.

Person, N. E. (2002). Assessment of TOEFL Scores and ESL Classes as Criteria for Admission to Career & Technical Education and Other Selected Marshall University Graduate Programs. Retrieved August 8, 2018, from https://files.eric.ed.gov/fulltext/ED473756.pdf

Phillips, A. (1983). Test Taking Skills: Incorporating Them into the Curriculum. Retrieved June 19, 2016 from https://eric.ed.gov/?id=ED235200

Pollitt, A. (2010). The oxford online placement test: The meaning of OOPT scores Retrieved August 8, 2016 from https://www.oxfordenglishtesting.com/uploadedfiles/buy_tests/oopt_meaning.pdf

Pulido, D. (2003). Modeling the Role of Second Language Proficiency and Topic Familiarity in Second Language Incidental Vocabulary Acquisition Through Reading. *Language Learning*, *53*(2), 233–284.

Purpura, J. (2009). *The Oxford online placement test: What does it measure and how.* Retrieved September 25, 2018 from https://www.oxfordenglishtesting.com/uploadedfiles/6_New_Look_and_Feel/Content/oopt_measure.pdf

Qian, D. (1999). Assessing the Roles of Depth and Breadth of Vocabulary Knowledge in Reading Comprehension. *Canadian Modern Language Review*, *56*(2), 282–308.

Rajapaksa, S. & Dundes, L. (2002) ‘It’s a long way home: International student adjustment to living in the United States’. College Student Retention, *4*(1), 15–28.

Ramsay, S., Barker, M., & Jones, E. (1999). Academic Adjustment and Learning Processes: a comparison of international and local students in first-year university. *Higher Education Research & Development*, *18*(1), 129–144.

Raymond, P.M., & Parks, S. (2002). Transitions: Orienting to reading and writing assignments in EAP and MBA contexts. *The Canadian Modern Language Review,* 59, 152 180.

Read, J. (1988). Measuring the vocabulary knowledge of second langauge learners. *RELC Journal*, *19*(2), 12–25.

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing*, *18*(1), 1–32.

Ricketts, J. C., & Rudd, R.D. (2005). Critical thinking skills of selected youth leaders: The efficacy of critical thinking dispositions, leadership, and academic performance. *Journal of Agricultural Education*, *46*(1) 32-44.

Rienties, B., Beausaert, S., Grohnert, T., Niemantsverdriet, S., & Kommers, P. (2012). Understanding academic performance of international students the role of ethnicity, academic and social integration. *Higher Education*, *63*(6), 685–700.

Robb, T. N., & Ercanbrack, J. (1999). A Study of the Effect of Direct Test Preparation on the TOEIC Scores of Japanese University Students. *Teaching English as a second or Foreign Language*, *3*(4), 2-16.

Robertson, M., Line, M., Jones, S., & Thomas, S. (2000). International students, learning environments and perceptions A case study using the Delphi technique. *Higher education research and development*, *19*(1), 89-102.

Rosenfeld, M., Leung, S., & Oltman, P. (2001). *TOEFL monograph series: The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels.* Princeton, New Jersey: Educational Testing Service.

Russell, J., Rosenthal, D., & Thomson, G. (2010). The international student experience-three styles of adaptation. Higher Education, *60*(2), 235–249.

Saklofske, D. H., Caravan, G., & Schwartz, C. (2000). Concurrent validity of the Wechsler Abbreviated Scale of Intelligence (WASI) with a sample of Canadian children. *Canadian Journal of School Psychology, 16(*1), 87-94.

Saville-Troike, M. (1984). What Really Matters in Second Language Learning for Academic Achievement? *TESOL Quarterly*, *18*(2), 199.

Sawaki, Y., & Nissan, S. (2009). Criterion-related validity of the TOEFL iBT listening section. *ETS Research Report Series*, *2009*(1), i – 82.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88.

Schoonen, R., Gelderen, A. van, Glopper, K. de, Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language Learning*, *53*(1), 165–202.

Schoonen, R., Hulstijn, J., & Bossers, B. (1998). Metacognitive and language-specific knowledge in native and foreign language reading comprehension: An empirical study among Dutch students in grades 6, 8 and …. *Language Learning*. Retrieved August 23, 2015 from http://onlinelibrary.wiley.com/doi/10.1111/1467-9922.00033/full

Segalowitz, N., Poulsen, C., & Komoda, M. (1991). Lower level components or reading skill in higher level bilinguals: Implications for reading instruction. *AILA Review*, *8*, 15-30.

Senyshyn, R.M., Warford, M.K. & Zhan, J. (2000). Issues of adjustment to higher education: International students' perspectives. International Education, *30*(1), 17–35.

Settles, 2016. The Reliability of Duolingo English Test Scores. Retrieved January 17, 2017 from https://s3.amazonaws.com/duolingo-papers/reports/DRR-16-02.pdf

Sewell (2009). Analysis of the IELTS Test (Unpublished MA dissertation). Birmingham: University of Birmingham

Sharon, A. T. (1972). English Proficiency, Verbal Aptitude, and Foreign Student Success in American Graduate Schools. *Educational and Psychological Measurement*, *32*(2), 425–431.

Shei, C. (2005). Plagiarism, Chinese learners and Western convention. *Taiwan Journal of TESOL*, *2*(1), 97–113.

Shepard, L. A., & Dougherty, K. C. (1991). Effects of high stakes testing on instruction, paper presented at the Annual Meeting of the American Educational Research Association and the National Council on Measure- ment in Education. Retrieved August 9, 2018, from https://eric.ed.gov/?id=ED337468

Shu, H. (2008). *Sojourners in transition: Chinese women undergraduate students at an American university*. University of Arkansas.

Smith, P. J., & Smith, S. N. (1999). Differences between Chinese and Australian students: Some implications for distance educators. *Distance Education*, *20*, 64-80.

Smith, Y. M. & Eccles, T. (1993) *An investigation into the learning experience of overseas students: a pilot study* (School of Surveying Occasional Paper no. 3). London: Kingston University.

Soars, L. Soars, J. & Sayer, M. (2000). New Headway: Pre-Intermediate: Student's Book. Oxford: Oxford University Press.

Soars, L. Soars, J. & Sayer, M. (2003). New Headway: Intermediate: Student's Book. Oxford: Oxford University Press.

Soars, L. Soars, J. & Sayer, M. (2003). New Headway: Upper-Intermediate: Student's Book. Oxford: Oxford University Press.

South China Morning Post. (2015). English exam body withholds results from 350 Chinese students over violations. Retrieved March 29, 2018, from http://www.scmp.com/news/china/society/article/1867668/english-exam-body-withholds-results-350-chinese-students-over

Spencer-Oatey, H., & Xiong, Z. (2006). Chinese students' psychological and sociocultural adjustments to Britain- An empirical study. *Language, culture and curriculum*, *19*(1), 37-53

Sternberg, R. J., Nokes, C., Geissler, P. W., Prince, R., Okatcha, F., Bundy, D. A., & Grigorenko, E. L. (2001). The relationship between academic and practical intelligence: a case study in Kenya. *Intelligence*, *29*(5), 401–418.

Stover, A. D. (1983). *Effects of Language Admission Criteria on Academic Performance of Non-native English-speaking Students*. University Microfilms.

Strathern, M. (1997). "Improving ratings": audit in the British University system. *European Review* , *5*(3), 305–321.

Stupnisky, R. H., Renaud, R. D., Daniels, L. M., Haynes, T. L., & Perry, R. P. (2008). The interrelation of first-year college students' critical thinking disposition, perceived academic control, and academic achievement. *Research in Higher Education*, *49*(6), 513.

Summers, J. A., & Shobe, R. E. (1983). Improving Test-Taking Skills. Retrieved July 10, 2017 from https://eric.ed.gov/?id=ED230573

Sun, W., & Chen, G. M. (1999). Dimensions of difficulties mainland Chinese students encounter in the United States. *Intercultural Communication Studies*, *IX*(1), 19-30.

Swain, M., Huang, L.-S., Barkaoui, K., Brooks, L., & Lapkin, S. (2009). The Speaking Section Of The TOEFL IBT: Test-Takers'reported Strategic Behaviors. *ETS Research Report Series*, *2009*(2), i – 118.

Tang, W. (2010, April 27). Minhong Yu's talk on "happy learning and healthy grow-up" in Wuxi. The New Oriental. Retrieved November 21, 2017 from http://old.neworiental.org/publish/portal0/tab410/ info498224.htm

TOEFL Homepage. (2018). https://www.ets.org/toefl

TOEFL iBT: test content. (2018). https://www.ets.org/toefl/ibt/about/content/

Trenkic, D., & Warmington, M. (2018). Language and literacy skills of home and international university students: How different are they, and does it matter? *Bilingualism: Language and Cognition*, 1–17.

Tunks, J. (2001). The Effect of Training in Test Item Writing on Test Performance of Junior High Students. *Educational Studies*, *27*(2), 129–142.

Universities UK international. (2017). *International facts and figures*. Retrieved September 19, 2018, from https://www.universitiesuk.ac.uk/policy-and-analysis/reports/Documents/International/International_Facts_and_Figures_2017.pdf

Vandergrift, L. (2004). Listening to Learn or Learning to Listen? *Annual Review of Applied Linguistics,* 24, 3–25.

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, *40*(3), 191–210.

Verhagen, J. (2011). Verb placement in second language acquisition: Experimental evidence for the different behavior of auxiliary and lexical verbs. *Applied Psycholinguistics*, *32*(4), 821–858.

Verspoor, M., & Lowie, W. (2003). Making Sense of Polysemous Words. *Language Learning*, *53*(3), 547–586.

Wagner, E., & Kunnan, A. J. (2015). The Duolingo English test. *Language Assessment Quarterly*, *12*(3), 320-331.

Wagner, C. H. (1982). Simpson's Paradox in Real Life. *The American Statistician*, *36*(1), 46–48.

Wall, D. (2000) The impact of high-stakes testing on teaching and learning can this be predicted or controlled. System, *28*(4), 499-509.

Wan, G. F. (2001). The learning experience of Chinese students in American universities: A cross-cultural perspective. *College Student Journal*, 35, 28-44.

Wang, L., Eignor, D., & Enright, M. K. (2008). A final analysis. *Building a Validity Argument for the Test of English as a Foreign Language*, 259–318.

Wang, X. (2007). From restoration to mega-expansion: higher education reform in China: Higher education reform in China: A three decade review (1976-2006). Analytical Reports in International Education, *1*, 7-20.

Ward, C., & Masgoret, A. M. (2004). The experiences of international students in New Zealand: Report on the results of the national survey. Retrieved June 27, 2017 from http://www.educationcounts.govt.nz/__data/assets/pdf_file/0006/15288/040604- final-report-for-printers.pdf

Waters, A. (1996). *A review of research into needs in English for academic purposes of relevance to the North American higher education context.* TOEFL Monograph Series Report No. 6. Princeton, NJ: Educational Testing Service.

Wechsler, D. (2011). *Wechsler abbreviated scale of intelligence, 2nd edition* (WASI-II). Oxford: Pearson.

Weir, C., Hawkey, R., Green, A., Unaldi, A., Devi, S., & Others. (2009). The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university. *IELTS Research Reports, 2009*(9), 97-156.

Wells, C. S., & Wollack, J. A. (2003). An instructor's guide to understanding test reliability. *Testing & Evaluation Services. University of Wisconsin*. Retrieved August 19, 2016 from https://testing.wisc.edu/Reliability.pdf

West, D. E. (2012). Elicited imitation as a measure of morphemic accuracy: Evidence from L2 Spanish. *Language and Cognition*, *4*(3), 203–222.

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *The Journal of Animal Ecology*, *75*(5), 1182–1189.

Wilson, K. M. (1987). Patterns of test taking and score change for examinees who repeat the Test of English as a Foreign Language. *ETS Research Report Series*, *1987*(1), i-68.

Xie, Q. (2013). Does Test Preparation Work? Implications for Score Validity. *Language Assessment Quarterly*, *10*(2), 196–218.

Xu, M. (2007). How to get a slice of the action: The analysis of China's English training industry (Unpublished master's thesis). University of Nottingham, UK.

Yan, A. (2015). Test of credibility: How Chinese exam "cheats" threaten students' dreams of studying abroad. Retrieved March 29, 2018, from http://www.scmp.com/news/china/money-wealth/article/1874818/test-credibility-how-chinese-exam-cheats-threaten-students

Yang, J.(2015). Upper level SSAT scores cancelled in China administration Retrieved September 22, 2016, from http://www.chinadaily.com.cn/china/2015-10/22/content_22249699.htm

Ye, F. (2014). Validity, reliability, and concordance of the Duolingo English Test. Retrieved January 17, 2017 from https://s3.amazonaws.com/duolingo-papers/other/ye.testcenter14.pdf

Yen, D., & Kuzma, J. (2009). Higher IELTS Score, Higher Academic Performance? The Validity of IELTS in Predicting the Academic Performance of Chinese Students. *Worcester Journal of Learning and Teaching*, *3*, 1–7.

Yixin, W., & Daller, M. (2014). Predicting Chinese Students' academic achievement in the UK. In *Proceedings of the 47th Annual Meeting of the British Association for Applied Linguistics, Learning, Working and Communicating in a Global Context,* 217–227.

Yu, L., & Suen, H. K. (2005). Historical and contemporary exam-driven education fever in China. *KEDI Journal of Educational Policy*, *2*(1). 17-33.

Yu, Y. (2014). A study on Chinese learners' IELTS preparation efforts. *TESOL Quarterly,* 33(2), 329-337.

Yuan, W. L. (2011). Academic and cultural experiences of Chinese students at an American university: A qualitative study. *Intercultural Communication Studies, XX*(1), 141-157.

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship Between Lexical Competence and Language Proficiency: Variable Sensitivity. *Studies in Second Language Acquisition*, *27*(4), 567–595.

Zhang, Y. (2008). Repeater analyses for TOEFL iBT. Research Memorandum ETS RM-08–05. Princeton, NJ Educational Testing Service.

Zhao, C. M., Kuh, G. D., & Carini, R. M. (2005). A comparison of international student and American student engagement in effective educational practices. The Journal of Higher Education, *76*(2), 209-231.

Zheng, Y., & Cheng, L. (2008). Test review: College English Test (CET) in China. *Language Testing*, *25*(3), 408–417.

Zheng, Y., & De Jong. (2011). Concurrent Validity of Pearson Test of English Academic. Retrieved February 8, 2016 from https://pearsonpte.com/wp-content/uploads/2014/07/RN_EstablishingConstructAndConcurrentValidityOfPTEAcademic_2011.pdf

Zhou, Y., Jindal-Snape, D., Topping, K, & Todman, J. (2008). Theoretical models of culture shock and adaptation in international students in higher education. *Studies in Higher Education, 33*(1), 63-75.

Zi, M. (2004). New Oriental guilty of copyright violation. Retrieved September 22, 2017 from http://www.chinadaily.com.cn/english/doc/2004-12/29/content_404573.htm

Zou, Y. L. (2000). The voice of a Chinese immigrant in America: Reflections on research and self-identity. In E. T. Trueba & L. I. Bartolome (Eds.), *Immigrant voices: In search of educational equity* (pp. 187-203). Maryland: Rowman & Littlefield Publishers, Inc.