This is a repository copy of *The effects of coaching and repeated test-taking on Chinese candidates' IELTS scores, their English proficiency, and subsequent academic achievement*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/152768/

Version: Accepted Version

Article:
Hu, Ruolin (Rowling) and Trenkic, Danijela orcid.org/0000-0001-6340-6030 (2021) The effects of coaching and repeated test-taking on Chinese candidates' IELTS scores, their English proficiency, and subsequent academic achievement. International Journal of Bilingual Education and Bilingualism. pp. 1486-1501. ISSN 1367-0050

https://doi.org/10.1080/13670050.2019.1691498

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

The effects of coaching and repeated test-taking on Chinese candidates' IELTS scores, their English proficiency, and subsequent academic achievement

Abstract

Although most international students arrive with required language qualifications, many struggle with the linguistic demands of their programmes (Murray, 2010). This study explored whether the test-preparation industry undermines the qualifications with which students arrive.

English proficiency of 153 Chinese student in the UK was tested on the Duolingo English Test and a C-test. Students who attended IELTS-coaching programmes scored lower on both measures compared to students who met entry requirements without such help. Furthermore, the number of attempts to achieve a particular IELTS score was negatively correlated with the other English proficiency scores on arrival. The results confirm that coaching, and to some extent repeated test-taking, boost IELTS scores without generalising to other proficiency measures. The effects were, however, small so that despite the observed inflation, IELTS scores were a reliable predictor of academic success: the rise of one IELTS band resulted in the average mark increase of 9 and 4 points (out of 100) in linguistically more and less demanding disciplines, respectively.

The results underscore the important role that language plays in study success, and show that many international students get accepted with levels of English that limit their academic achievement. The test-preparation industry contributes in part to this.

Keywords: IELTS, test-preparation, coaching, repeated test-taking, academic success

Introduction

Literacy and proficiency in the language of instruction are cornerstones of success in any academic subject. Limited mastery of these skills diminishes the opportunity to learn and makes assessment more difficult. This has been long recognised in research on primary and secondary school-age populations, both monolingual and bilingual (August & Shanahan, 2006; Hakuta, Butler & Witt, 2000; Kieffer, 2008; Preevo, Malda, Mesman & van IJzendoorn, 2016). In tertiary education, where a lot of learning happens through reading and nearly all academic outcomes are assessed in writing, language and literacy remain critically important. Yet, because universities had traditionally been academically selective and linguistically homogenous in its intake, individual differences in these skills were not typically among the key factors influencing academic success (Abraham, Richardson & Bond, 2012). It is only now, in the era of widening participation and intense internationalization of higher education that we observe how individual differences in literacy and proficiency affect achievement in this context (see contributions in this volume).

Although linguistic abilities are predictive of academic success, this relationship appears asymptotic: below a certain threshold, the relationship is strong – the higher the skills, the better the prospects; beyond that threshold, however, language ceases to be predictive of academic success. This does not mean that everyone above the threshold is guaranteed the highest academic marks, but merely that language ceases to be a barrier for fulfilling one's academic potential, however big – or indeed, small – the actual potential may be.

In university contexts where English is the language of instruction, the notion of threshold and the importance of well-developed language and literacy skills are to some extent recognised in relation to international students who do not speak English as their first language. These students are required to demonstrate their readiness to study in English on one of the approved tests. The most commonly accepted at UK, New Zealand and Australian universities is the academic version of the International English Language Testing System (IELTS Academic; henceforth IELTS). The test consists of four parts (writing, reading, listening and speaking), and the results are expressed as an overall score on a scale from 1 to 9, and subscores for all four language skills. Requirements differ by institution and by programme, but scores between 6.0 and 7.0 are typically accepted for unconditional entry (Feast, 2002; Green, 2007).

Yet despite the minimum standards, set at the level which the receiving institutions consider adequate, many international students struggle with the linguistic demands of their programmes (Murray, 2010). Indeed, there is growing research evidence that language skills with which international students arrive constrain what they can achieve academically (Daller & Phelan, 2013; Daller & Xue, 2009; Elder, Bright & Bennett, 2007; Read & Hayes, 2003; Trenkic & Warmington, 2019). Although some international students do exceptionally well, as a group, they experience lower academic success than home students (Crawford & Wang; 2015; Iannelli & Huang, 2014; Morrison et al, 2005).

Several factors may contribute to the situation where international students arrive with a level of English that is below the threshold that would enable them to perform academically to their full potential. First, receiving institutions often set the language entry bar lower than test developers' recommendations. The IELTS test score guidance for educational institutions (IELTS, 2014) describe IELTS scores in relation to linguistically demanding courses as fully acceptable only if they are in the 7.5 – 9.0 range (i.e. higher than the range normally accepted); an IELTS score of 7.0 is described as 'probably acceptable', and 6.5 and lower as needing more work ('English study needed'). For linguistically less demanding courses, the recommendations are half a band lower: 7.0 is 'acceptable', 6.5 is 'probably acceptable', and a candidate scoring 6.0 should improve their English first. This means that many students arriving with the minimum score for unconditional entry start their studies with the level of English that the test-developers consider as falling below the needs and requirements for the discipline-relevant degree-level study.

Second, even applicants who fail to meet the minimum entry requirements may still be offered a place on condition that they attend a remedial English course with the receiving institution (or an approved subsidiary); satisfactory completion of the course leads to direct admission into the degree programme. Although such courses may well be useful for developing general academic skills, published research suggests that they rarely lead to an improvement in English proficiency (Green, 2005; 2007). The weaker English skills with which they arrive put these students at a disadvantage: research shows that those who gain entry through attending English remedial courses achieve lower academic grades compared to students who met the minimum requirements on one of the secure tests (Eddey & Baumann, 2011; Oliver, Vanderford & Grote, 2012), who in turn do not do as well as students who exceed the minimum language requirements (Trenkic & Warmington, 2019). This suggests that for students arriving with the minim language requirements, and even more so for those who miss them, language is a significant barrier to fulfilling academic potential.

There may be a further factor that aggravates the situation – the one that we investigate here – and it is that scores with which students apply for a university place may themselves be inflated. Because of its high-stake status and the growing demand for international education, IELTS has spawned a large and burgeoning test-preparation industry. The industry's principal aim is to help candidates attain a required score, often by any means necessary. Some of the practices have led to the annulment of student results, or to hefty fines imposed on test-preparation providers (Yan, 2015; Zi, 2004). Even when activities stay on the right side of the law, test-preparation institutions typically focus on repetitive practice of earlier tests, encouraging participants to memorise "canned answers to probable questions" (Matoush & Fu, 2012, p.113). Although these practices may introduce construct-irrelevant variance in the scores, there is at present limited understanding of just how much they impact on the test scores' validity. By focusing on the population of Chinese students in the UK, this study explored how IELTS-preparation industry – specifically coaching and repeated test-taking – affects the level of English with which international students arrive (the test's extrapolation validity), and whether this, in turn, affects the predictive validity of IELTS scores.

Previous research on test-preparation and repeated test taking on test scores validity

Coaching

Dedicated test-preparation programmes, also known as test coaching, are on offer to candidates who need to pass high-stakes tests. They can improve test scores through several distinct mechanisms (Messick, 1982). Some of these pose no threat to the validity of scores. For example, activities that familiarise students with the format of the test may reduce anxiety and so improve the validity by removing irrelevant factors. Other activities may improve scores by helping develop the underlying skill that the test is measuring. Some test-preparation activities, however, can undermine the validity of scores. Notably, they are activities that improve scores by narrowing the curriculum to focus only on the content and the type of questions that are likely to feature on the test. Such activities can lead to an improvement in scores without a corresponding improvement in the underlying skill, compromising the extrapolation validity of scores (the strength of inference from the tested to an untested behaviour).

Improving language proficiency takes long periods of study. Evidence suggests, however, that some improvement in test scores can be achieved relatively quickly through

curriculum-narrowing practices. In one of the most robust studies on the effects of preparation on standardised English tests, Xie (2013) explored a two-month score gain by 850 Chinese students preparing for the national College English Test (CET 4). The strongest predictor of the final performance was the score achieved two months earlier, showing a good internal validity of the test. Yet, some test-preparation activities also made a unique, positive contribution to the final score. Invariably, they were of the curriculum narrowing type: intensive practice of retired test papers and specific test items, taking the test repeatedly, and reciting vocabulary lists. Although the overall effect of these practices was small in absolute terms, it represented almost 1/3 of the overall effect. Thus while the internal validity of the test seemed preserved, the extrapolation validity of scores appeared compromised. The scores were effectively inflated.

Similar effects of the curriculum-narrowing practices were observed by Green (2007). The study measured IELTS writing score gains amongst 476 international students in the UK. They were attending either dedicated IELTS-preparation programmes or more general English for Academic Purposes (EAP) pre-sessional programmes, across 15 institutions. Although no discernible differences were found between different types of programmes, the only course parameter that made a unique, positive contribution to the score gains across all programmes were 'activities in the class similar to IELTS test' (Green, 2007, p.90).

These are exactly the types of activities that the thriving English test-preparation industry has honed on (Matoush & Fu, 2012). In a bid to prepare candidates 'for a game of probability' (Yan, 2015), test-preparation centres tend to focus on repetitive practice using retired exams and parallel test items. Students themselves report that their reason for attending such programmes is not to develop language skills but to develop skills to pass the test (Ma & Cheng 2015).

Following a group of 45 Chinese students through a 4-week preparation in one such centre, Trenkic and Hu (under review) observed a reliable increase in the IELTS scores of about half an IELTS band. At the same time, candidates' performance on a number of other tests indexing English proficiency – Online Oxford Placement test, vocabulary task, and written sentence comprehension – showed no corresponding improvement. Furthermore, their performance on these tests, both before and after the training, was very similar to a group of 44 control participants who were not engaged in any test preparation at the time. The only measure of proficiency on which the test-preparation group outperformed the control group after the training was the IELTS test itself. The results suggest that attending a 4-week coaching programme can boost IELTS scores by about half a band without this improvement

being reflected on other tests of English proficiency. In other words, the test-preparation activities appear to undermine the extrapolation validity of scores. Previous doubts regarding the power of dedicated test preparation courses 'to deliver anticipated yields' (Green, 2007, p.93) may not have taken into account the more intense end of the test coaching industry.

Repetitive test-taking

Most international students do not achieve a language score required for an unconditional university offer on the first attempt (Li, 2013). The great majority repeat the test at least once. On its website, IELTS advises test repeaters that their test scores are unlikely to increase on a further attempt unless they make a significant effort to improve their English. Published research supports this guidance. On both IELTS and TOEFL, scores do go up with the number of attempts, but the time between resits, or the length of the training that students undergo in the meantime, is a significant predictor of the gain (Green, 2005, 2007; Wilson, 1987).

Acknowledging that language development is slow and gradual – even when a lot of effort is put into it – IELTS used to have a 90-day resit rule in place. The rule was, however, removed in 2006, allowing candidates to take the test repeatedly and long before their proficiency has improved sufficiently to warrant a higher score. Hamid (2015) describes a case of a serial repeater who took IELTS 14 times in an 8 month period (including 3 attempts within a single month) in the failed quest for a particular score.

Although large gains on language tests are unlikely to result from repeated testing alone, there is at least some research to suggest that smaller improvements are possible. Analysing the data from around 12,000 candidates who repeated TOEFL within a single month, Zhang (2008) found evidence of small but reliable gains in scores (effect sizes of between .12 and .17 SD for the test components, and .17 SD for the test as a whole). Given the short interval within which the test was repeated, it is unlikely that the gains had resulted from an improvement in proficiency. The better performance on the repeat, however, could have resulted from the improved familiarity with the test format.

More worryingly, however, repetitive test-taking could also be driven by the more brazen practices of the test-preparation industry. In addition to offering test-driven classroom instruction, English language centres in China are also reported to sell candidates authentic test items (called *jijing*, Yan, 2015). These are compiled either through the power of social media (where candidates themselves publish test papers they have taken on online chatroom

websites), or by sending associates to take the test and memorise the questions (Yan, 2015). Although some of the largest test-preparation centres were fined in recent years for the breach of copyright of high-stakes tests and hundreds of test scores withdrawn from candidates, very little research exists on just how much repetitive test-taking affects the validity of IELTS and other high-stakes test scores.

Overview of the present study

In Trenkic & Hu (under review), we established that intensive IELTS-preparation programmes can boost IELTS scores by about 0.5 band without an associated improvement on alternative proficiency tests, measured immediately after the intervention. In this study, we recruited Chinese students at a UK university to explore how IELTS test-preparation practices affect students' language proficiency on arrival, and how well IELTS scores, so affected, predict academic outcomes. The study addressed three research questions:

1) How different is the English proficiency in university students who met the language entry requirements by attending IELTS-preparation programmes compared to those who met the same requirements without attending such programmes?
2) Does English proficiency differ in students who achieved the same IELTS result depending on the number of attempts it took them to achieve it?
3) How well do IELTS scores, affected by the test-preparation industry, predict academic outcomes of international students?

By focusing on these questions, our aim was to start and contribute to the debate on the effects of the test-preparation industry on both the extrapolation validity and the predictive validity of high-stakes language test scores.

Method

Participants

One hundred and fifty-three (138 female) Chinese students attending a UK university participated in this study. Their median age was 23 years (range 21-28), and they all spoke Mandarin Chinese as their dominant language. They were graduates of recognized Chinese universities, and were, at the time of testing, studying for a one-year masters degree at a UK university. All participants sat at least one IELTS test prior to arriving in the UK. Eighty were enrolled on programmes requiring an overall IELTS score of at least 7 (e.g. TESOL, Applied linguistics, English literature), and 73 were studying linguistically less demanding

programmes requiring an IELTS score of 6.5 (e.g. various business, management or marketing programmes) or 6.0 (e.g. engineering, finance, or music subjects).

Instruments and measures

Participants self-reported their IELTS test-preparation histories and the number of IELTS attempts. Their final IELTS scores (with which they were accepted for their current programme), as well as the scores on the initial IELTS attempt for those who took the test more than once, were recorded from official certificates.

We administered two additional tests of English proficiency: the Duolingo English Test and a C-test. The Duolingo English Test is a computer-based adaptive test developed and owned by Duolingo. At the time of our study, the test consisted of four types of tasks: listening to short utterances and writing them down; speaking / reading aloud short sentences; a vocabulary task (deciding whether a string of letters is a word or nonword); and a version of a cloze test. Given the test's adaptive nature where a computer algorithm decides the next item based on how well the test-taker has answered the previous one, the length of the test and the number of test items may vary between test-takers. On average, it took participants in our study 25-30 minutes to complete the test. The test is scored automatically on the scale from 1 to 100.

A C-test is a type of a cloze test that estimates the test-taker's general language proficiency based on their ability to restore incomplete words in a text. C-tests are designed by applying the RULE OF TWO: 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 (e.g. *essential*), the larger half is deleted (e.g. *esse*____). To provide the initial context, the first sentence of each text is left unchanged, as are names, numbers and one-letter words throughout the text. Our C-test consisted of 5 short texts, with 100 blanks across them, ordered in an increasing level of difficulty. The sum of correctly restored words (scale 0-100) was used in the analyses. Cronbach's alpha for the internal consistency of the scale was .817. The instrument is available on the IRIS repository, www.iris-database.org.

Both the Duolingo English Test and the C-test have a narrower operationalisation of the construct of English proficiency than IELTS. Whereas IELTS operationalises English proficiency as an ability to communicate through listening, speaking, reading and writing in academic contexts, neither Duolingo nor C-test test communicative competence directly. Rather, they measure linguistic knowledge (vocabulary; implicit grammar patterns) and

lower-level skills (e.g. speech segmentation in listening; suprasegmental skills in speaking) which underpin and predict communicative ability (cf. Daller, Mueller & Wang-Taylor, this issue; Brenzel & Settles, 2017). Although the appropriateness of the Duolingo English Test for university admissions purposes has been questioned (Wagner & Kunnan, 2015), its reported reliability indices (internal reliability coefficient 0.96; split-half-reliability 0.96; test-retest reliability 0.84; Settles 2016) and its criterion-related validity (Duolingo-IELTS r=0.70; Duolingo-TOEFL r=0.71, Brenzel & Settles, 2017) appear good. Similarly high are the reported internal and criterion-related reliability indices for C-tests (Daller, Mueller & Wang-Taylor; Dörnyei & Katona, 1992).

The purpose of including the additional measures of English proficiency was not to arbitrate which test is a better measure of the construct or a better predictor of academic success. Rather, assuming a correlation between different measures of English proficiency, we expected that higher scores on IELTS, irrespective of the route by which they were achieved, would be reflected in higher scores on the independent proficiency measures. Conversely, a systematic difference on independent proficiency measures between those who achieved a particular IELTS scores through coaching and/or repeated test-taking and those who achieved them without such support would indicate some fundamental difference in their mastery of English.

As the correlation between the Duolingo scores and the C-test scores in our study was significant and moderate (r=.458, p=.000), a composite proficiency score, created by summing the z-scores from both tests, was used in regression analyses as an independent index of English proficiency. Both tests were also positively correlated with the IELTS scores: there was a significant and strong correlation between the C-test scores and the IELTS scores (r=.560, p=.000), and a significant and moderate correlation between Duolingo and IELTS (r=.395, p=.000). Although these correlations were not as strong as those reported in the previous literature, they validate the assumption that higher scores on one measure of English proficiency map onto higher scores on another measure.

As one of our questions concerned academic outcomes, we also measured participants' working memory and non-verbal intelligence as potential confounding variables. Non-verbal intelligence was assessed using the Matrix Reasoning subset from the *Wechler Abbreviated Scale of Intelligence II* (WASI-II, Wechsler, 2011). In this task, participants view a series of geometrical forms arranged according to an implicit logical principle, and select the form that completes the matrix from a set of options. The scale has 30 items and the sum of correct answers (0-30) was used in the analyses. Participants' working memory was

measured using an auditory forward digit-span task in both English and Chinese. The span was calculated by averaging the highest three sequences that a participant repeated correctly. For the regression analyses, a composite digit span score was used as an index of working memory, calculated by adding z-scores from both versions.

Participants' academic success was operationalized as the weighted average mark at the end of the taught component of their masters programme.

Design and procedures

Participants were recruited through posted adverts in autumn 2016, soon after starting their programmes. They were tested individually in a single session lasting approximately 75-80 minutes. The tests were administered in the following order: Duolingo English Test, matrix reasoning, digit span in Chinese and English, and C-test. During the same session their IELTS histories and other demographic data were collected. Nine months later, at the end of the taught component of their masters, academic marks were collected from the University registry services with the participants' consent.

Participants received a small payment for their participation. The study was approved by the Department of Education Ethics Committee, University of York.

Analyses

Means and standard deviations were calculated for all measures, and where the data was not normally distributed, median and mode were calculated, too. Bivariate correlations were used to establish the relationship between different proficiency measures. The independent *t*-test (for normally distributed data) and the Mann-Whitney test (for non-normally distributed data) were used to compare the IELTs scores and independent proficiency scores (Duolingo and C-test) of students who attended IETLS-preparation programmes with those who did not.

Hierarchical regressions were used to explore the role of IELTS test-preparation industry (attendance at coaching programmes, the number of test attempts), above and beyond IELTS scores, on the English proficiency as measured by alternative tests. The same method was used to investigate the ability of IELTS scores affected by the IELTS-preparation industry to predict students' academic success.

Results

IELTS coaching

Of the 153 participants in our study, 87 (57%) reported having attended IELTS-preparation programmes. Table 1 summarises participants' IELTS, Duolingo and C-test scores for the whole group of 153, and divided by whether they attended an IELTS-preparation programme. It also presents the statistical comparison of the English proficiency scores between those who attended and those who did not attend IELTS-preparation programmes. The results show that although the two groups arrived with roughly similar IELTS scores, the IELTS-coached group scored significantly lower on the two independent measures of English proficiency: their Duolingo scores were on average 5.5 points below the group who met the language entry requirements without attending IELTS-preparation training, and the C-test scores were about 3.2 points lower.

Table 1. *Comparison of English proficiency scores between participants who attended IELTS-preparation programmes and those who did not undergo such training.*

	N	IELTS M(SD)	Duolingo M(SD)	C-test M(SD)
All participants	153	6.72 (.48)	56.31 (15.51)	44.79 (10.76)
Attended IELTS prep course	87	6.66 (.47)	53.90 (15.05)	42.33 (10.16)
Didn't attend IELTS prep course	66	6.81 (.48)	59.49 (.15.64)	48.03 (10.73)
Test statistics		U=2391.50	t(151)=2.24	t(151)=3.35
		p=.058	p=.027	p=.001
		r=.15	r=.18	r=.26

In the group that underwent IELTS coaching (n=87), we explored the relationship between the length of the programme (range 1 to 24 weeks, Mdn=4; M=5.84; SD=3.85) and the IELTS score gains (range -.50 to 2.00, Mdn=.50, M=.55, SD=.49), and found that they were not significantly correlated, r=.193, p=.073. This suggests that there may be a limit to how much dedicated IELTS-preparation courses can boost the scores, which the previous research estimates at about half a band (Trenkic & Hu, under review).

We also wished to understand how much of the variance in language proficiency, measured on alternative tests on arrival, is explained by students' attendance / non-attendance on IELTS-preparation programmes, above and beyond that explained by the difference in their IELTS scores. To do that, we fitted a linear hierarchical regression model with the composite proficiency score on arrival (sum of Duolingo and C-test z-scores) as an outcome variable, and with IELTS scores and attendance at IELTS-preparation programmes as predictor variables. The regression model is summarised in Table 2. On its own (Model 1),

the final IELTS scores explained 30.5% of variance in the composite English proficiency score. Attendance at an IELTS-preparation programme (Model 2) significantly improved the model by explaining a further 2.4% of the variance. This corroborates the finding that among students presenting with the same IELTS scores on entry to university, those who achieved that score by attending IELTS-preparation programme could do less well on alternative tests of English proficiency compared to those who achieved it without coaching. In other words, IELTS was a less stable indicator of proficiency for students who attended dedicated test-preparation programmes.

Table 2. *Regression model using final IELTS overall score, attendance on IELTS-preparation programmes and number of test attempts to predict English proficiency scores on alternative tests upon participants' entry to university (n=153).*

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

Note: Final model $F(3,149)=26.00$, $p=.000$

Number of attempts

From the whole sample of 153, only 18 participants were accepted with their first attempt IELTS score, while 171 repeated the test at least once. The median number of attempts was 3 (M=3.17, SD=1.49). The median gap between the first and the last attempt was 212 days (about 7 months), and the median gap between any two attempts was 91 days (3 months). Table 3 breaks down the English proficiency scores (Duolingo, C-test, IELTS) and IELTS gains by the number of IELTS attempts and reports correlations between each of these variables and the number of attempts.

Table 3. *English proficiency scores (Duolingo, C-test, IELTS) and IELTS gains, and their relationship with the number of IELTS attempts.*

IELTS attempts	N	%	IELTS gain M (SD)	IELTS score on arrival	Duolingo	C-test
1	18	11.8		7.03 (.44)	63.78 (19.96)	50.94 (10.41)
2	37	24.2	.41 (.50)	6.80 (.45)	57.87 (16.83)	45.60 (9.97)
3	42	27.5	.39 (.39)	6.70 (.50)	57.17 (13.32)	45.83 (11.44)
4	31	20.3	.66 (.40)	6.63 (.43)	53.74(13.76)	41.90 (9.75)
5	12	7.8	.67 (.65)	6.46 (.62)	54.17 (15.67)	39.00 (11.45)
6	9	5.9	.78 (.26)	6.61 (.41)	42.11 (8.30)	39.33 (8.57)
7	3	2.0	.50 (.00)	6.83 (.29)	59.33 (8.63)	52.67 (4.93)
8	1	0.7	.50	6.50	52	45
			r=.379	*r*=.-238	*r*=.-235	*r*=.-218
			p=.000	*p*=.003	*p*=.003	*p*=.007

All English proficiency measures, including IELTS scores, were significantly negatively correlated with the number of IELTS attempts: students who have repeated IELTS test more times tend to arrive with significantly lower proficiency in English than students who have taken the test fewer times. This does not mean that repeated test-taking lowers the scores – on the contrary, there was a positive correlation between the number of attempts and the IELTS gain from the first to the last test. Rather, students with lower initial proficiency take the test more times to reach the minimum entry requirements.

The critical question that we sought to answer was: does repeated test-taking affect the extrapolation validity of IELTS scores? In other words, does English ability differ in students who achieve the same IELTS result depending on the number of attempts it took them to achieve it? If repeated test-taking inflates IELTS scores, we would expect students who arrive having achieved a particular IELTS score on fewer attempts to attain higher scores on alternative tests of English than students who achieved the same score on more attempts. Figure 1 confirms that within each IELTS half-band in the range between 6.0 and 7.5, there was a negative correlation between the number of IELTS attempts a student has taken to achieve that particular score and the composite score on the two alternative English tests.

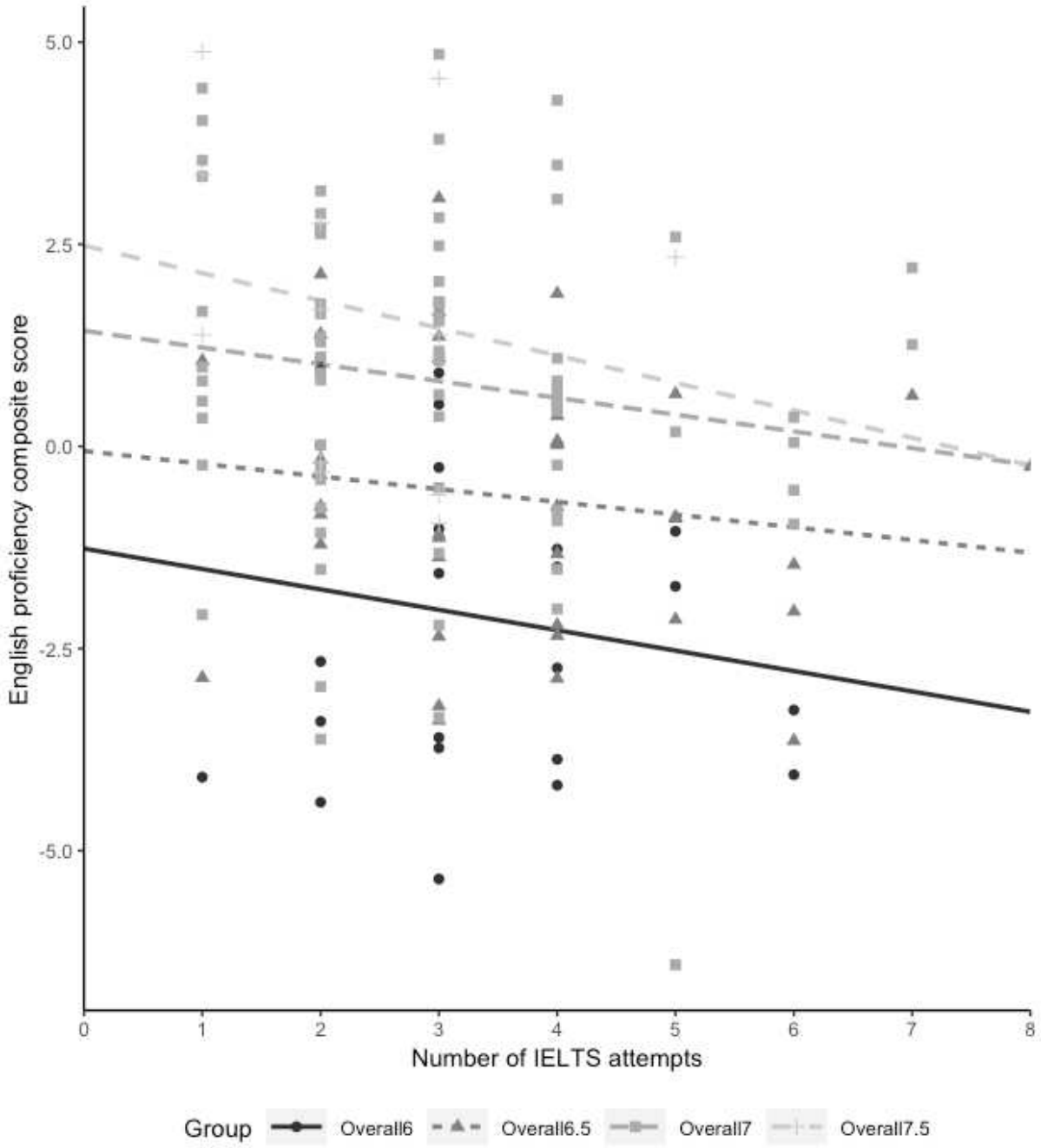

Figure 1. The relationship between the number of IELTS attempts and the composite score that participants attained on two alternative English tests, within each IELTS half-band (6.0 to 7.5).

To statistically confirm whether the number of attempts additionally boost IELTS scores above and beyond the boost provided by the IELTS-preparation programmes (Model 2 in Table 2), we added to the model as the next predictor the number of IELTS attempts. Because of small number of students who attempted IELTS more than 6 times (n=3), we grouped them together into a category "6 or more attempts" (n=13). Adding the number of

IELTS attempts improved the model fit by explaining a further 1.5% variance of the composite proficiency score (Model 3), but this change was slightly short of the significance level (p=.067).

The final regression model thus suggest that dedicated, curriculum-narrowing test preparation has more power to inflate IELTS scores than do repeated tests on their own. The effect of repeated testing was smaller, and the present sample of 153 participants was not sufficiently large to confidently detect (or rule out) a statistical significance of small effects, for which a sample of about 600 participants is needed (Field, 2005).

Academic success

The results above corroborate the view that the IELTS test-preparation industry can boost test-takers' scores without reliably improving their ability to do well on another English test, so that students who arrive at university having met the entry requirements by attending dedicated test-preparation programmes, and to a lesser extent those who repeated the test several times, arrive with less stable proficiency in English than those who met the entry requirements without it. If, as it appears, the test-preparation industry undermines the extrapolation validity of IELTS scores, it is important to understand whether it also interferes with the ability of IELTS scores, as a measure of English proficiency, to predict academic outcomes in the ranges where such a relationship would be expected.

Table 4 summarises the weighted average grade, overall IELTS scores, matrix reasoning scores, and working memory (composite) scores for all participants, and divided by the linguistic demand of the programme (students enrolled on programmes with IELTS requirement of band 6.5 or below vs students enrolled on programmes with IELTS requirement of at least 7 overall).

Table 4. *Comparison of the weighted average grade, IELTS scores, non-verbal intelligence and working memory between participants enrolled on linguistically more and linguistically less demanding programmes.*

	N	Non-verbal intelligence M(SD)	Composite WM (digit span) M(SD)	IELTS M(SD)	Weighted average mark M(SD)
All participants	153	20.95 (3.11)	.00 (1.70)	6.73 (.48)	59.92 (7.49)
Participants on linguistically more demanding programmes	80	20.71 (2.90)	.03 (1.70	6.96 (.27)	58.31 (6.41)
Participants on linguistically less demanding programmes	73	21.22 (3.33)	-.03 (1.72)	6.47 (.52)	61.68 (8.21)
Test statistics		$t(151)=1.01$ $p=-.316$ $r=.08$	$t(151)=.23$ $p=.815$ $r=.02$	$U=4563.50$ $p=.000$ $r=.58$	$t(151)=-2.85$ $p=.005$ $r=.23$

Participants studying linguistically more demanding subjects did not differ from participants studying linguistically less demanding subject in their non-verbal intelligence or working memory, as measured by WASI II Matrix reasoning task and the forward digit span, respectively. They did differ in the average IELTS score with which they arrived, with participants enrolling on programmes demanding higher IELTS scores typically arriving with the required higher scores (M=6.96, SD=0.48) than participants enrolling on programmes demanding lower scores (M=6.47, SD=0.27). However, participants on linguistically less demanding programmes went on to achieve a higher weighted average grade (62) compared to participants in linguistically more demanding programmes, who achieved on average a grade of 58. This difference was significant, even though the effect was small (Table 4). Whatever the cause – and there could be several potential contributors which go beyond the scope of this paper – it suggests that it is important to take the linguistic demand of a programme into account when considering the predictive power of IELTS for academic outcomes. Most importantly, the average IELTS scores for both groups indicated English proficiency levels below the test-developers' recommended thresholds (IELTS 7.5 for more linguistically-demanding and 7.0 for linguistically less-demanding programmes), and thus in the range where English proficiency should be predictive of academic success.

To test the predictive power of IELTS on academic outcomes, we ran a hierarchical regression model, with weighted average grade as an outcome variable (Table 5). In the first block, we entered non-verbal intelligence and working memory as control variables known to influence academic success (Model 1). IELTS scores were entered in the next block (Model 2), and the linguistic discipline (linguistically more or less demanding), as a potential moderating variable, in the final block (Model 3).

Table 5. Regression models of the predictors of academic success.

Model	B	Coefficients SE	ß	t	*p*	R^2	ΔR^2	ΔR^2 sig.
1						.018	.018	.259
Constant	57.13	4.13		13.85	.000			
Non-verbal intelligence	0.13	0.20	.06	0.68	.496			
Working Memory	0.54	0.36	.12	1.51	.133			
2								
Constant	45.90	9.51		4.83	.000	.029	.011	.192
Non-verbal intelligence	0.14	0.19	.06	0.71	.480			
Working Memory	0.51	0.36	.12	1.44	.151			
IELTS score	1.66	1.26	.11	1.31	.192			
3								
Constant	29.41	9.80		3.00	.003	.136	.107	.000
Non-verbal intelligence	0.07	0.19	.03	0.39	.698			
Working Memory	0.50	0.34	.11	1.48	.141			
IELTS score	4.76	1.40	.31	3.40	.001			
Linguistic demand	-5.73	1.34	-.38	-4.27	.000			

Note: Final model $F(4,148)=5.81$, $p=.000$

The final model shows that IELTS scores with which participants arrived, and the linguistic demand of the programme on which they were enrolled, were both significant unique predictors of the average academic mark achieved. Non-verbal intelligence and working memory, while positively linked to academic success, were not statistically significant predictors in our sample. Specifically, the model predicts that an increase in IELTS score of one band (in the range 5.5 – 8.0) results in an average improvement of the weighted average mark of 4.76 points. The result, thus, shows that even though many of the IELTS scores in our sample were affected by the test-preparation industry, the predictive validity of the test for academic outcomes was not lost.

Furthermore, the fact that the linguistic demand of the programme was a significant contributor to participants' academic marks has an important methodological implication when considering the role of language proficiency – and IELTS as a measure of it – on academic outcomes. When all participants in our study were grouped together without accounting for linguistic demand of the programme (Model 2), IELTS was not predictive of academic outcomes – the unaccounted linguistic demand of the programme obscuring the effect of the language proficiency on performance. It is only when the linguistic demand is taken into account that the effect of proficiency becomes apparent.

To understand better how IELTS scores influence academic outcomes in linguistically more vs linguistically less demanding programmes, we ran two separate regression analyses, with weighted average grade as an outcome and IELTS score as a predictor variable (Tables 6 and 7). The results show that in linguistically more demanding programmes, IELTS scores accounted for 14% of the variance in academic performance ($F(1,78)=12.93$, $p=.000$), with

each band increase in IELTS overall scores leading to an increase in the weighted average mark of about 8.85 points (or 4.43 points for each half band). In linguistically less demanding programmes, IELTS scores accounted for 6% of the variance in academic performance ($F(1,71)=4.13$, $p=.046$), with each increase of one band in IELTS scores leading to an increase in the weighted average mark of about 3.69 points (or 1.85 for each half band). In other words, language is a stronger predictor of academic outcomes in linguistically more than linguistically less demanding academic disciplines.

Table 6. *Regression model using final IELTS overall score to predict academic outcomes (weighted average grade) in linguistically more demanding programmes (n=80)*

Model	B	Coefficients SE	ß	t	*p*	R^2	F	*p*
1						.142	(1.78)= 12.93	.001
Constant	-3.28	17.14		-.19	.849			
IELTS score	8.85	2.46	.38	3.60	.001			

Table 7. *Regression model using final IELTS overall score to predict academic outcomes (weighted average grade) in linguistically less demanding programmes (n=73)*

Model	B	Coefficients SE	ß	t	*p*	R^2	F	*p*
1						.055	(1,71) =4.13	.046
Constant	37.85	11.76		3.22	.002			
IELTS score	3.69	1.81	.24	2.03	.046			

Discussion

IELTS-preparation industry undermines the extrapolation validity of IELTS scores

Our previous work found that intensive curriculum-narrowing coaching programmes have a potential to boost IELTS scores without generalising to other English tests. Specifically, in [Trenkic & Hu, under review], we observed that a group of Chinese students undergoing a 4-week programme in Shanghai experienced an average boost of about half a band in IELTS scores without a corresponding increase on other proficiency measures: Online Oxford Placement Test, vocabulary knowledge, and sentence comprehension. Here, we focused on the population of Chinese masters students in the UK. We tested their English proficiency on arrival on two independent measures: Duolingo English Test, and a C-test. We found that students who had attended IELTS-coaching programmes to meet the language entry requirements scored significantly lower on both the Duolingo English Test and the C-test compared to students who had met the entry requirements without attending such programmes. In the regression model, the attendance at an IELTS-coaching programme was a

significant negative predictor of the independently measured proficiency, even after accounting for individual differences in IELTS scores with which the students arrived.

The above findings corroborate and extend the results from Trenkic & Hu (under review) in several important ways. First, by using two new measures – the Duolingo English Test and a C-test – this study confirms that IELTS-coaching programmes can boost IELTS scores without a corresponding improvement on another measure of proficiency. Taken together, the results strongly suggest that IELTS-coaching programmes can lead to IELTS scores that overstate test-takers' mastery of English.

The findings also underscore the persistency of the effect. In Trenkic & Hu (under review), language proficiency was measured immediately after the preparation programme, in the context of the language centre where the programme was delivered. Here, we show that the same discrepancies between IELTS scores and alternative measures of proficiency are evident in students enrolled on masters programmes in the UK, months after the coaching programmes were undertaken. This rules out the possibility that coaching which has an immediate effect on boosting IELTS scores might generalise with a delay to other proficiency measures. University applicants who meet the entry requirements after attending IELTS-coaching programmes start their degree programmes with less stable proficiency in English than students who attained the same IELTS scores without coaching.

The findings of the present study are also more generalizable. In Trenkic & Hu (under review) we followed participants through a 4-week intensive IELTS-coaching programme in Shanghai. Participants in the present study attended a range of programmes, differing in length, provider and location. Despite that, the effect (boosting of IELTS scores without corresponding improvement on other proficiency measures) was the same, suggesting that the finding is not provider-specific. This suggests that the test-preparation industry has identified what practices result in the quickest gain in scores, and that similar, curriculum-narrowing programmes are probably the norm across different training centres.

Looking at the effects of repeated test-taking on boosting IELTS scores, we found that within each IELTS half-band, there was a negative correlation between the number of attempts it took a student to achieve that particular result and their English proficiency measured on arrival on a new test. This suggests that repeated test-taking can also inflate IELTS scores, but not to the same extent as IELTS-coaching programmes do. In line with previous research on repeated test-taking (Green, 2005; Zhang, 2008), the effect was small. Studies with larger samples may confirm this effect as genuine, but it is likely to remain small, and primarily of theoretical rather than practical significance.

In sum, our study shows that practices encouraged by the IELTS-preparation industry – test-coaching and to a lesser degree repeated test-taking – can inflate the scores. As a result, many international students arrive with IELTS scores indicating proficiency levels higher than they can demonstrate on an alternative test.

IELTS is a good predictor of academic success, despite the score inflation

In the present study, IELTS scores at the point of acceptance to a university predicted academic outcomes in a sample of Chinese masters students in the UK. In particular, the rise of one IELTS band in linguistically more-demanding disciplines resulted in an increase of about 9 points in academic success (out of 100), or a nearly whole degree classification higher. This result is in line with the data reported for a similar population in Trenkic and Warmington (2019). In linguistically less-demanding disciplines, the effect was smaller but still significant, with an increase of one IELTS band corresponding to the difference of about 4 points in the average grade. By including the measures of non-verbal reasoning and working memory, we ruled out the possibility that the observed positive relationship between English proficiency and later outcomes is due to variation in the students' general cognitive ability.

The results of our study attest to the robustness of IELTS as a measure of readiness to study in English, despite any inflation in scores caused by the test-preparation industry. This might be because the effect of the industry is relatively small (our data suggests that the score inflation is limited to about half an IELTS band), and also because the large majority of IELTS scores in our sample were affected by the test-preparation industry: some by coaching, some by repeated test-taking, and some by both. In fact, only 11 out of 153 participants achieved the required scores without attending a coaching programme or repeating the test at least once. When the majority of students arrive with scores suggestive of higher level of proficiency than their actual ability, the relative ranking amongst them may remain reasonably stable (having engaged in similar practices to attain the scores).

The predictive validity of IELTS scores for academic success underscores two important messages. First, that well-developed language and literacy skills are critical for success in tertiary education, in any academic subject. Limited mastery of these skills present a barrier to learning and achievement. Second, language entry requirements for all programmes, but especially for linguistically-demanding disciplines, are set at a level considerably lower than the threshold after which language ceases to be a barrier to

performance. For most international EFL students, as a consequence, the proficiency of English with which they arrive constraints what they can achieve academically. Or to put it differently, many international students are intellectually capable of doing much better than their mastery of English allows them to.

Methodological implications

On the methodological level, the study confirms the importance of accounting for disciplinary differences when investigating language as a predictor of academic success (Feast, 2002). First, the results suggest that some disciplines award marks in higher ranges than others, and that in directly comparing academic results across disciplines we may not, in fact, be comparing like with like. Second, the results confirm that some disciplines are more linguistically demanding than others, in that the language plays a more critical role in learning and demonstrating what one has learnt. Therefore, when disciplinary differences, and in particular the linguistic demand of the programme, are not taken into account, language proficiency, and IELTS as a measure of it, may appear irrelevant for academic success. As our study demonstrates, it is only when the effect of the linguistic demand is statistically partialled out that the predictive power of IELTS scores may become evident.

We also note that as the relationship between language proficiency and academic success is expected to be asymptotic (i.e., the linear relationship can be approximated only over a portion of the curve), the findings from our analyses (linear regressions) may not hold beyond the ranges of IELTS scores collected here.

Conclusion and practical implications

Our study confirms that English proficiency, and IELTS as its measure, are a strong predictor of academic success of international students in UK higher education. Being a strong predictor indicates that the level of English with which many international students arrive is below the threshold that would enable them to perform academically to the best of their ability. Like much of previous research, our study shows that students who arrive with scores recommended by test developers (and thus higher than the typical minimum entry requirements) do better than equally capable students who only meet their programmes' minimum entry requirements, but that they, in turn, do better than the peers who miss the entry requirements and gain entry through attending English remedial courses. This confirms that by setting the minimum language requirements below the test-takers' recommendations,

receiving universities put international students at a disadvantage that is difficult to overcome.

Although our study validates IELTS as a strong predictor of academic success, the results also indicate that the intensive test-preparation industry may be undermining the extrapolation validity of scores. We found that students who arrive having met the language entry requirements by attending dedicated IELTS-preparation programmes, and to a lesser extent those who repeated the test several times, do less well on alternative test of English proficiency on arrival than students who met the entry requirements without such help. In that sense, the test-preparation industry seems to add to the situation where international students arrive with the level of language skills that is still a barrier to what they can achieve.

International students accept their offers in good faith, believing that if the university has accepted their qualifications, their English skills must be good enough to allow them to fulfil their academic potential. For those who find out that their English is not strong enough to allow them to learn and perform at the true level of their ability, this risks jeopardising their educational experience, their mental health and wellbeing, and their future employment prospects. As part of their duty of care, universities should therefore set their entry requirements prudently and with awareness that some of the scores with which international students present will overstate applicants' actual linguistic ability. It is also important to explore ways in which students disadvantaged in their studies by the level of the English ability with which they are accepted could be helped to achieve their academic potential. This may include better language support provision but also special assessment arrangements, such as extra time in exams.

References

Abraham, C., Richardson, M., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, *138*, 353–387.

August, D., & Shanahan, T. (eds.) (2006). Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth. Mahwarh, NJ: Lawrence Erlbaum Associates & Centre for Applied Linguistics.

Brenzel, J., & Settles, B. (2017). The Duolingo English Test – design, validity, and value. *DET Whitepaper (Short)* (englishtest.duolingo.com/resources).

Crawford, I., & Wang, Z. (2015). The impact of individual factors on the academic attainment of Chinese and UK students in higher education. *Studies in Higher Education*, *40*, 902–920.

Daller, M., Müller, A., & Wang-Taylor, Y. (this issue). The C-test as predictor of academic success of international students. *International Journal of Bilingual Education and Bilingualism.*

Daller, M., & Phelan, D. (2013). Predicting international student study success. *Applied Linguistics Review*, *4*, 173–193.

Daller, M. H., & Xue, H. (2009). Vocabulary knowledge and academic success: A study of Chinese students in UK higher education. In B. Richards, H.M. Daller, D.M. Malvern, P. Meara, J. Milton, J. Treffers-Daller (eds.), *Vocabulary studies in first and second language acquisition: The interface between theory and application*, pp. 179–193. London: Palgrave Macmillan.

Dörnyei, Z. & Katona, L. (1992). Validation of C-test amongst Hungarian EFL learners. *Language Testing, 9*, 187-206.

Eddey, P., & Baumann, C. (2011). Language proficiency and academic achievement in postgraduate business degrees. *International Education Journal: Comparative Perspectives, 10,* 34-46.

Elder, C., Bright, C., & Bennett, S. (2007). The role of language proficiency in academic success: Perspectives from a New Zealand university. *Melbourne Papers in Language Testing*, *12*, 24–58.

Feast, V. (2002). The impact of IELTS scores on performance at university. *International Education Journal 3(*4), 70-85.

Field, A. (2005). *Discovering statistics using SPSS* (2nd edition). London: Sage.

Green, A. (2005). EAP study recommendations and score gains on the IELTS Academic Writing test. *Assessing writing, 10,* 44-60.

Green, A. (2007). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education*, 14(1), 75-97

Hakuta, K., Butler, Y. G., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* Policy report by the University of California Linguistic Minority Research Institute. Retrieved from http://escholarship.org/uc/item/13w7m06g

Hamid, M.O. (2015). Policies of global English tests: test-takers' perspectives on the IELTS retake policy. *Discourse: Studies in the Cultural Politics of Education, 37*, 472-487.

Iannelli, C., & Huang, J. (2014). Trends in participation and attainment of Chinese students in UK higher education. *Studies in Higher Education*, *39*, 805–822.

IELTS (2014). Guide for educational institutions, governments, professional bodies and commercial organisation.

Kieffer, M. J. (2008). Catching up or falling behind? Initial English proficiency, concentrated poverty, and the reading growth of language minority learners in the United States. *Journal of Educational Psychology*, *100*, 851–868.

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

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*, 58–79.

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*, 111-121.

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice. *Educational Psychologist, 17*, 67-91.

Morrison, J., Merrick, B., Higgs, S., & Le Métais, J. (2005). Researching the performance of international students in the UK. *Studies in Higher Education*, *30*, 327–337.

Murray, N. (2010). Considerations in the post-enrolment assessment of English language proficiency: Reflections from the Australian context. *Language Assessment Quarterly*, *7*, 343–358.

Oliver, R., Vanderford, S., & Grote, E. (2012). Evidence of English language proficiency and academic achievement of non-English-speaking background students. *Higher Education Research & Development*, *31*, 541–555.

Prevoo, M. J., Malda, M., Mesman, J., & van IJzendoorn, M. H. (2016). Within-and cross-language relations between oral language proficiency and school outcomes in bilingual children with an immigrant background: A meta-analytical study. *Review of Educational Research*, *86*, 237–276.

Read, J., & Hayes, B. (2003). The impact of IELTS on preparation for academic study in New Zealand. In R. Tulloh (Ed.), *IELTS research reports 2003* (pp. 153–205). Canberra: IELTS Australia.

Roche, T., & Harrington, M. (2013). Recognition vocabulary knowledge as a predictor of academic performance in an English as a foreign language setting. *Language Testing in Asia: A Springer Open Journal*, *3*(12), 133–147.

Settles, B. (2016). The reliability of Duolingo English Test scores. *Duolingo Research Reports DRR-16-02.*

Trenkic, D., & Hu, R. (under review). Coaching for language assessment in China: How effective is it?

Trenkic, D., & Warmington, M. (2019). Language and literacy skills of home and international university students: How different are they, and does it matter? *Bilingualism: Language and Cognition*, *22*, 349-365. doi:10.1017/S136672891700075X 1-17.

Wagner, E. & Kunnan, A. J. (2015). The Duolingo English test (test review). *Language Assessment Quareterly, 12,* 320-331.

Wechsler, D. (2011). Wechsler abbreviated scale of intelligence, 2nd edition (WASI-II). Oxford: Pearson.

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*, 196-218.

Yan, A. (2015). Test of credibility: How Chinese exam "cheats" threaten students' dreams of studying abroad. Retrieved from https://bit.ly/2OMyqU8

Zhang, Y. (2008). Repeater analyses for TOEFL iBT. Research Memorandum. Education Testing Services. Retrieved from https://www.ets.org/Media/Research/pdf/RM-08-05.pdf.