JESLA | Journal of the European Second Language Association

Teaching to the test: The effects of coaching on English-proficiency scores for university entry

RESEARCH

DANIJELA TRENKIC
RUOLIN HU
*Author affiliations can be found in the back matter of this article

WHITE ROSE UNIVERSITY PRESS
Universities of Leeds, Sheffield & York

ABSTRACT

Despite arriving with the required language qualifications, many international students struggle with the linguistic demands of a university degree. Using the International English Language Testing System (IELTS) as an example, this study explored how short but intensive preparation programmes may affect high-stakes English-proficiency test scores with which students apply for university places. The participants were 89 Chinese speakers of English as a foreign language in Shanghai. They were tested twice, four weeks apart, on IELTS and three other measures of English ability: The Oxford Online Placement Test, a vocabulary test, and the speed and accuracy of sentence comprehension. Between the two testing points, 45 participants underwent test-specific training consisting of previous IELTS papers, offered by a large test-preparation establishment with a network of over 1,000 training centres. The remaining 44 participants did not engage in any test preparation at the time. Teaching to the test led to a half a band rise in IELTS scores above the gain from test repetition alone, suggesting that the training was effective. Importantly, the IELTS gain did not generalise to the other measures of English ability; the groups performed similarly on all other language tests at both times. This suggests that test-specific, curriculum-narrowing courses could be inflating the scores with which international students apply for university places, with important consequences for test-developers, universities and students.

CORRESPONDING AUTHOR:
Danijela Trenkic
University of York, GB
danijela.trenkic@york.ac.uk

KEYWORDS:
English-language proficiency assessment; language testing; test preparation; coaching; international students; higher education; IELTS

TO CITE THIS ARTICLE:
Trenkic, D., & Hu, R. (2021). Teaching to the test: The effects of coaching on English-proficiency scores for university entry. *Journal of the European Second Language Association*, 5(1), 1–15. DOI: *https://doi.org/10.22599/jesla.74*

DOI: 10.22599/jesla.74

1. INTRODUCTION

As millions of international students prepare to demonstrate their readiness to pursue university education in English as a condition for enrolment, a powerful language test-preparation industry has emerged to meet the rising demand. In a bid to help their clients achieve the required scores on high-stakes exams in a time-efficient way, many language schools and training centres offer dedicated test-preparation programmes.

But do such programmes work? Do they reliably improve scores, and are the scores improved in this way trustworthy? In the context where many international students struggle with the linguistic demands of their programmes (Murray, 2010) and where as a group they experience lower academic success than home students (Morrison et al., 2005), the question of whether, and to what extent, the language test-preparation industry may be subverting the validity of scores is an important one to address. Here we explore this question in the context of China, a country which currently sends the largest number of students abroad and has a well-established test-preparation industry.

1.1. LANGUAGE PROFICIENCY AS A CRITERION FOR UNIVERSITY ENTRY

In countries where admission to higher education is academically selective, universities set criteria to choose students with the capacity to learn quickly and perform well in their studies. Scholastic aptitude is typically the chief criterion, usually indexed by candidates' previous academic success, and in some cases by scholastic aptitude tests. However, because the capacity to acquire new knowledge also depends on proficiency in the language of instruction (Elder et al., 2007; Daller & Phelan, 2013; Trenkic & Warmington, 2019), most universities require international students to demonstrate their linguistic readiness to study on one of the approved language-proficiency tests.

At universities teaching in English, one of the most widely accepted tests is the academic version of the International English Language Testing System (IELTS Academic, henceforth IELTS), which is owned and administered by the Cambridge Assessment English, the British Council and the International Development Program Australia Consortium (henceforth, the Consortium). It is a test constructed, trialled and validated for university admissions purposes, and it operationalises English proficiency as the ability to communicate through listening, speaking, reading and writing in academic contexts. It comprises four sections testing each of these skills, with the results reported on a 9-band scale (including half-point scores) for each skill and as the overall test score (the average of four skills scores).

The test does not have a pass/fail boundary. Instead, in its guidance to educational institutions, the Consortium lists scores between 7.5 and 9.0 as acceptable for studying linguistically demanding academic courses, and scores of 7.0 and above as sufficient for linguistically less demanding courses (IELTS, 2019). Each university, however, sets its own minimum requirements, with the decisions usually guided by pragmatism and compromise (Deygers & Malone, 2019). The requirements thus differ by institution and by programme, with scores between 6.0 and 7.0 (corresponding to the range from the mid B2 to low C1 level of the Common European Framework of Reference) typically accepted for unconditional entry (Feast, 2002; Green, 2007).

1.2. THE EFFECTS OF TEST PREPARATION ON THE VALIDITY OF SCORES

When the question of validity in high-stakes language assessment is raised, the concern is typically with score validity: Whether trustworthy inferences about a candidate's language proficiency can be drawn from the behaviours elicited by the test and from the resulting test scores. In relation to test preparation, the question is whether particular activities could reliably lead to higher scores, and whether in the process, they could be undermining the validity of scores.

Theoretically, test preparation can improve test scores via three main mechanisms, each with different consequences for score validity. First and foremost, on tests that appropriately measure the intended construct, scores should improve along with the underlying skill. Thus, test-preparation activities that focus on the development of language proficiency should improve scores on language-proficiency tests with no consequence, positive or negative,

DOI: 10.22599/jesla.74

for their validity. The limitation of this approach as a mechanism for improving test scores, however, is that it is not quick. Language proficiency is an ability that develops over a long time—typically over many years of education and practice—and therefore significant progress is difficult to achieve through short programmes of intervention.

The second mechanism that leads to higher scores involves countering some of the common threats to score validity. For example, familiarising students with the format of the test can both reduce evaluative anxiety and ensure that the manner in which the skill is tested does not take them by surprise. Although such activities do not change the underlying ability being tested, they do make it easier for candidates to demonstrate it on the test, and in so doing, stand to improve the validity of the score (Messick, 1982).

The third mechanism, in direct contrast to the last, leads to higher scores by actively exploiting threats to validity. For example, training students in test-taking strategies and answer-selection tricks seeks to achieve high scores by taking advantage of the construct-irrelevant properties of a test (Messick, 1982). Furthermore, narrowing the curriculum to only focus on practising previous tests and parallel items (for example, teaching writing in only one genre that is assessed on the test) may result in higher scores through exploiting both the construct-irrelevant properties of a test and the test's construct under-representation (Haladyna & Downing, 2004).

Unlike activities that target language-proficiency development, both those that minimise and those that exploit common threats to validity have shown promise in quick score gains. For example, acquiring even a basic familiarity with the format through sitting a test once seems to statistically improve the chances of getting a higher score during the following attempt. Zhang (2008) analysed 2,000 candidates who took the internet-based Test of English as a Foreign Language (TOEFL iBT, henceforth TOEFL), a test which, like IELTS, measures academic English proficiency in listening, reading, speaking and writing. The study showed small but reliable gains in scores when the test was repeated within a single month.

Curriculum-narrowing test preparation is also shown to lead to score gains over a period of one to two months. For example, focusing on the relationship between test preparation and test performance of 14,593 students taking the TOEFL in China, Liu (2014) found that after controlling for participants' English proficiency on an independent test prior to the study, the strongest predictor of TOEFL scores was preparation activities that involved intense and short-term practice using similar test formats (TOEFL practice tests and previously used authentic test items). Similar findings were reported by Xie (2013) in a study that followed a group of 1,003 undergraduate students as they prepared for the national College English Test Band 4 in China. This study recorded participants' test scores, both before and after a two-month preparation period and collected detailed information about their preparation activities. As in Liu (2014), the effects of test preparation on test performance were significant but small, explaining 2.4% of the variance in the final scores after controlling for the proficiency with which the participants started the preparation. Crucially, the only test-preparation practices that made a significant positive contribution to the final scores were those that were of a curriculum-narrowing nature: The intensive practice and memorisation of specific test items and of earlier versions of the test. Xie (2013) argued that as activities that enhance the target construct in a very narrow domain, they effectively weaken the extrapolation link from the test scores to untested behaviours.

1.3. IELTS PREPARATION AND SCORE VALIDITY

When it comes to the effects of IELTS preparation on its test scores, the previous literature was predominantly focused on understanding how long it takes to observe reliable improvements in scores following different training regimes. Much of this research was conducted in Anglophone countries (United Kingdom, Australia, New Zealand), often involving the English for Academic Purposes (EAP) programmes offered by the receiving universities (e.g., Brown, 1998; Archibald, 2001; Elder & O'Loughlin, 2003; Rao, et al., 2003; Read & Hayes, 2003; Green 2005, 2007). Of those, the study that touched most directly on the question of test preparation and score validity was that reported in Green (2005, 2007), which explored IELTS writing score gains amongst 476 international students across 15 institutions in the UK.

DOI: 10.22599/jesla.74

The study focused on learners enrolled in courses preparing them for academic study in the United Kingdom over periods of between three and ten weeks. Most of the participants (n = 331) were attending EAP pre-sessional courses provided by universities in the United Kingdom as an enrolment condition for international students with language scores falling short of the requirements for unconditional entry. Although these courses involved some training resembling an IELTS writing test (e.g., timed writing practice), they were not actively preparing students for the test. The rest of the participants were either enrolled in programmes whose curriculum combined a general EAP training with IELTS test preparation (n = 60) or were following special IELTS-preparation programmes (n = 85).

In line with research exploring the link between test preparation and score gains in other high-stakes tests, Green's (2005, 2007) study found a small but significant gain in IELTS writing scores (roughly 2/10 of an IELTS band) across programmes. Echoing the findings from Xie (2013) and Liu (2014), the only course parameter correlating significantly with score gains was class activities similar to the test (Table 5 in Green, 2007, p. 90). Despite this, no discernible differences in score gains were observed between the three different programme types.

What this means for the validity of scores is not entirely clear. Green (2007) argued, on the strength of the finding that different programmes led to similar score gains, that teaching to the test was not necessarily more effective in boosting IELTS scores than teaching the targeted skill, thus dismissing the power of dedicated test-preparation programmes to exploit test characteristics and undermine the validity of IELTS scores.

There are, however, at least two other plausible explanations. Theoretically, and as discussed earlier, test scores could improve through different mechanisms, with different consequences for score validity. Thus, despite the similar level of gain, the extrapolation link from tested to untested behaviours may still have been weaker for scores that improved through dedicated test-preparation programmes (assuming they improved only a subset of the writing domain, as tested by IELTS) compared to those achieved through broader EAP courses (assuming they improved academic writing more broadly construed). Alternatively, and probably more likely here, the three nominally different programme types may, in fact, not have been too dissimilar from each other. They could have led to the improvement in scores via the same mechanisms: Through striving to develop the underlying skill, fortified by activities that, by design or otherwise, familiarised students with the test format. Although the actual practices were not documented, given that all of the courses were described as preparing international students for academic study, it seems reasonable to assume that they were all making an earnest effort to help the students improve their English, and that none was designed to actively game the test. Furthermore, the finding that class activities similar to the test were the only part of the training that was significantly associated with score gains strongly suggests the possibility that different programmes, more intent on gaming the test, may still hold the potential to subvert score validity.

In sum, findings from previous research suggest that concentrating one's efforts on the repetitive practice of specific test formats can lead to quick and significant, if not very large, score gains in high-stakes language assessment. But whether, and if so to what extent, this actually undermines score validity remains unresolved. Specifically, Xie's (2013) argument that curriculum-narrowing training weakens the extrapolation link from the test performance to what one is capable of demonstrating in a different format requires empirical verification in studies that combine both tests for which the specific training was provided and alternative measures of language proficiency. Furthermore, to be able to confidently attribute score gains to test preparation, we need to be able to tease apart the effects of test preparation from the effects that arise through test repetition and from the baseline gain through a growth in ability regardless of test preparation. This is why studies with a control group are needed (Messick & Jungeblut,1981). We incorporated both features (a control group and alternative measures of proficiency) in the design of the present study.

1.4. LANGUAGE TEST PREPARATION IN CHINA

Large-scale written examinations and test-driven classroom instruction appear firmly rooted in Chinese society (Spolsky, 1995). Language test-preparation centres are widely used and argued to be sites of "the most egregious negative washback" (Matoush & Fu, 2012, p. 113), where

negative washback is understood as test-preparation activities that do not actively encourage language learning but focus instead on raising scores through exploiting test characteristics. Typical activities include test-taking strategies, repetitive practice, and memorisation of parallel test items.

In addition to practice tests and retired test papers that are publicly available, some centres are also reported to collate leaked items from active tests (Yan, 2015). Even though some of the major test-preparation providers have incurred large fines for the breach of copyright and hundreds of high-stakes test results have been annulled, their services remain highly praised and sought after by both students and their parents. As Matoush and Fu (2012, p. 114) observe,

> the competitive nature of attaining success in a heavily populated nation with limited university places has contributed to a disproportionate focus on test-preparation and parents willing to pay high prices for classes taught by those who seem able to predict test items.

Students themselves acknowledge that their goal in attending such programmes is not to improve their language skills but to hone the techniques needed to pass the test (Ma & Cheng, 2016).

1.5. OVERVIEW OF THE PRESENT STUDY

Given the ubiquity of test-preparation centres in China and assuming that the intensity and focus of test-driven instruction may, at least in part, be culturally specific, we set the present study in a large training centre in Shanghai. The study addressed two research questions:

(1) How much can IELTS scores improve through a brief but intensive preparation course in China?

(2) Do such gains generalise to other measures of language proficiency?

To evaluate the effectiveness of coaching and address the first question, we recruited a group of students undertaking a four-week preparation course and compared their gains on the IELTS test with that of a control group of uncoached participants. We approached the second question by comparing whether the gains in IELTS scores can be generalised to participants' performance on three other English-ability tests.

2. METHOD

2.1. PARTICIPANTS

Eighty-nine Chinese speakers of English as a foreign language participated in the study. Forty-five were recruited through a large training centre in Shanghai where they were enrolled on an IELTS-preparation course (the coached group). The other 44 served as a control group and were recruited through snowball sampling, asking the coached group participants to invite friends of a similar age and educational background. Participants in the control group were not attending any test-preparation training at the time.

All participants were from mainland China, received their education in Mandarin Chinese and spoke it as their dominant language. They started learning English in school at the age of 11 and had never lived abroad. None reported having had previous experience with the IELTS test or IELTS test-preparation courses. At the time of testing, the majority of participants were in full-time education (10 final-year high school students and 67 university undergraduates), six were recent university graduates preparing for further studies abroad, and six were university graduates in employment. The median age of the participants was 21 years (range 18–37). The groups were similar in the demographic factors of age, gender balance and level of education (see ***Table 1***).

Table 1 Demographic information about the sample.

PARTICIPANT GROUP	GENDER	AGE MEDIAN (RANGE)	FINAL YEAR HIGH SCHOOL STUDENTS	UNIVERSITY UNDERGRADUATE STUDENTS	RECENT UNIVERSITY GRADUATES	UNIVERSITY GRADUATES (EMPLOYED)
Coached group	23 female; 21 male	20 (18–37)	6	32	6	1
Control group	26 female; 19 male	21 (18–34)	4	35	0	5

DOI: 10.22599/jesla.74

2.2. DESIGN

Participants were tested twice, four weeks apart, on IELTS and on three other tests, measuring English proficiency and related enabling skills. These were Oxford Online Placement Tests (OOPT), a vocabulary test and a speed and accuracy of sentence comprehension test (the latter two from Baddeley, Emslie & Nimmo Smith, 1992). Different versions of test papers were counterbalanced across participants and between times.

Between the two testing points, the coached group underwent an intensive four-week IELTS-preparation programme. It was offered by one of the leading schools for language teaching and test-preparation training in China with a national network of over 1,000 providers. The course curriculum consisted of previous IELTS papers from the Cambridge IELTS book series (Books 6, 7, 8, 9, and 10 were used at the time of data collection), supplemented by a collection of speaking and writing topics featured at recent IELTS exams (practice known as *Ji Jing* 机经[1]), put together by candidates and tutors. Eighteen hours of preparation and practice was devoted to each part of the test (Reading, Listening, Writing and Speaking). The instruction involved the identification of common topics across papers, a close analysis of the format, grammar and lexis of previous tests and model answers and a range of test-taking strategies, from predicting where in the text the answer to a question is likely to be, to memorising paragraphs and practising to repurpose them for similar topics in speaking and writing tasks. Participants followed the course at different points during 2015 and 2016, but the overall teaching and learning objectives, the course length and the main materials remained consistent.

The participants in the control group were not undergoing any test preparation at the time. The design allowed us to tease apart the effects of test preparation from the effects of test repetition or any growth in ability (e.g., through the consolidation of previous knowledge) that may have occurred regardless of the coaching programme.

2.3. MEASURES

2.3.1. IELTS academic test

The format of the IELTS test consists of four components: Listening, Reading, Writing and Speaking. The components are administered in this order, with the first three being tested in a controlled group setting, and the last one in a face-to-face individual interview. The Listening and Reading tests are allocated 40 and 60 minutes, respectively. The Writing test takes 60 minutes and the Speaking test is between 11 and 14 minutes long. The results are expressed on a 1 to 9 scale, in half-point increments.

We used authentic past IELTS examination papers from the Cambridge IELTS book series (Books 1 and 2; Jakeman & McDowell, 1996; University of Cambridge Local Examination Syndicate, 2000). The selected papers were not used by the training centre in teaching. However, they were from the same book series as the papers used in teaching and were similar to them in length, format and content. The format of the active IELTS test at the time of our study (2015–2016) did not differ substantially from that of the retired papers we used. When administered in a controlled environment, retired tests are expected to provide an accurate indication of the candidate's likely performance on the active IELTS test (Cambridge University Press, 2020). Test-retest reliability of the papers in our study was good, $r = .79$.

Writing tasks were scored by two experienced IELTS instructors, following the IELTS procedures and criteria.[2] Interclass correlations (ICC, ***Table 2***) indicate a high degree of inter-rater reliability.

Table 2 Interclass correlation estimates of inter-rater reliability on IELTS writing tasks.

TIME AND WRITING TEST COMPONENT	INTERCLASS CORRELATION	95% CONFIDENCE INTERVALS
Time 1, IELTS Writing part 1	.978	.966–.985
Time 1, IELTS Writing part 2	.972	.958–982
Time 2, IELTS Writing part 1	.986	.979–.991
Time 2, IELTS Writing part 2	.976	.964–.984

1 *Ji Jing* 机经 roughly translates as 'experience of taking the actual test'.

2 For full IELTS assessment criteria and score calculations, see IELTS (2020).

DOI: 10.22599/jesla.74

Performance on the speaking component was assessed by one experienced IELTS speaking instructor using the published IELTS criteria, and the listening and reading tests were scored using the answer key provided with the tests.

2.3.2. Online Oxford Placement Test (OOPT)

The OOPT is a widely used test in second-language learning research to evaluate participants' proficiency in English, as well as by language schools, universities or employers to determine the level of training that their recruits need. It tests knowledge of grammar, as well as the ability to understand literal and implied meanings of English words, phrases and sentences. It assesses this knowledge through listening and reading only. Compared to IELTS, which taps directly into the four language skills, the OOPT can be seen as an index of the linguistic knowledge that underpins those skills.

The OOPT is a computer-adaptive test; it adapts to the ability level of each test-taker. It does so by selecting items based on how the test-taker answered the previous question. If the last question was answered correctly, the next one is selected to be more difficult. If the wrong answer is given, the test goes on to select an easier question next. Because of its adaptive nature, the number of items and the length of the test are not fixed. There is no set time-limit, but the test usually takes between 30 and 40 minutes to complete (Purpura, 2009).

Scores are generated automatically on a scale of 120. For an in-depth account of the score calculation algorithm (see Pollitt, 2009, and for a summary see Hu, 2018). On its website (*www.oxfordenglishtesting.com*), the OOPT is reported as validated on 19,000 test-takers in 60 countries. Test-retest reliability in our study was $r = .61$.

2.3.3. Speed and Capacity of Language-Processing (SCOLP) test

The SCOLP (Baddeley et al., 1992) has two components that between them test vocabulary and the speed and accuracy of sentence comprehension. Although not a test of English proficiency *per se*, SCOLP provides measures of those abilities that are critical for language use. Vocabulary knowledge, in particular, is a precondition for all other language skills and is strongly correlated with other measures of language proficiency (Roche & Harrington, 2013; Trenkic & Warmington, 2019). The speed at which written language is processed and how much of it is understood are also strong indicators of developing language proficiency.

The vocabulary component of the test, also known as The Spot the Word Test, is a version of the yes-no vocabulary task (Meara & Buxton, 1987), providing an index of receptive vocabulary knowledge in English. Participants performed a silent lexical decision task on 60 pairs of items, containing one real and one pseudo word (e.g., *kitchen – harrick*). The target words ranged in frequency of occurrence from common to extremely rare, thus providing a measure of the richness of the participants' vocabulary (number of correctly identified words out of 60). Previous research has established that measuring vocabulary size through yes-no judgements correlates highly with other standard measures of vocabulary knowledge and is a good measure of L2 proficiency (Roche & Harrington, 2013). The test-retest reliability was $r = .88$.

The sentence-processing component of SCOLP contains 100 short sentences, half of which are true (e.g., *Dogs have four legs; Birds have wings*) and half are false (e.g., *Dogs have wings; Birds have four legs*). The participants' task is to verify the statements as quickly as they can. The test was administered using a pen and paper format. The total reading times and accuracy scores (scale 0 to 100) were used in the analyses. The test-retest reliability was .94 for the speed of reading and .92 for accuracy. The full SCOLP test (vocabulary task + sentence-processing component) took participants between 10–15 minutes to complete.

Although clearly different in format, scope and purpose, both IELTS and these additional measures test aspects of the same trait (i.e., English-language ability). Score gains on one, therefore, may be expected to be reflected, at least to some degree, in score gains on the others. If corresponding improvements are absent, this should raise validity concerns.

2.4. PROCEDURE

Participants sat the IELTS Listening, Reading and Writing components, in this order, in a controlled group setting similar to the authentic IELTS exam. Later that day or the following

DOI: 10.22599/jesla.74

morning, they were individually tested on the IELTS Speaking component by an experienced IELTS speaking instructor. The following day, the OOPT was administered in a controlled group setting. Finally, the participants sat the two components of the SCOLP Test. Testing at both Time 1 (T1) and Time 2 (T2) followed the same steps.

2.5. ANALYSIS

Preliminary analysis showed that the scores on most measures were not normally distributed. As the assumption of normality of distribution was not met, we used non-parametric tests in our main analyses: The Wilcoxon signed rank test to compare changes in the within-group performance between T1 and T2 and the Mann-Whitney's *U* test to compare groups on the gain in scores achieved between T1 and T2.

3. RESULTS

3.1. IELTS SCORES AT T1 AND T2

The median overall score of the participants at the start of our study (T1) was band 6.0. According to the IELTS band descriptors, this level denotes a competent user who has an effective command of the language despite some inaccuracies, inappropriate usage and misunderstandings and who can use and understand fairly complex language, particularly in their own field (IELTS, 2019).

Both groups saw some gain in IELTS scores from T1 to T2.[3] The coached group's median scores rose by half a band: The Listening, Reading and the Overall scores moved from 6.0 to 6.5, and the Writing and Speaking scores moved from 5.5 to 6.0 (***Table 3***). The mean of the overall score improved 6/10 of a band, from 5.66 (*SD* = 0.87) to 6.26 (*SD* = 0.60). The shift was evident for the majority in the coached group. For example, the Overall IELTS score improved for 36 out of 45 participants and for the remaining nine participants it remained unchanged. No participant saw a drop from T1 to T2 (see ***Figure 1***). The effect sizes were large (medium in the case of Writing), and the Wilcoxon signed rank test confirmed all the gains as statistically significant (***Table 3***).

Table 3 Comparison of Time 1 and Time 2 scores in the coached group (n = 45).

MEASURE	TIME 1			TIME 2			*z*	*p*	EFFECT SIZE *r*
	M (*SD*)	MEDIAN	RANGE	*M* (*SD*)	MEDIAN	RANGE			
IELTS overall	5.66 (.87)	6.0	2.0–7.0	6.26 (.60)	6.5	4.5–7.0	5.46	.000	.58
IELTS listening	5.81 (1.15)	6.0	2.0–8.0	6.59 (.75)	6.5	5.0–8.0	4.98	.000	.53
IELTS reading	6.00 (1.11)	6.0	2.0–7.5	6.73 (.96)	6.5	4.0–8.5	4.90	.000	.52
IELTS writing	5.42 (.98)	5.5	1.5–7.0	5.87 (.62)	6.0	4.5–7.0	4.25	.000	.45
IELTS speaking	5.29 (.71)	5.5	3.0–6.5	5.91 (.57)	6.0	4.0–7.0	5.13	.000	.54
Sent. reading speed	448.00 (149.68)	410	270–813	417.62 (128.07)	401	250–764	–3.51	.000	.37
Sent. comp.	73.71 (9.28)	75	50.0–87.5	73.84 (10.07)	76	47.0–90.0	0.98	.327	.10
Vocabulary	33.30 (4.60)	32.5	20–41	32.97 (4.57)	33	20–42	1.56	.119	.16
OOPT	51.31 (12.56)	54	21–73	53.02 (15.22)	53	17–87	0.93	.350	.10

In contrast, the control group's gains were small (***Table 4***). The median values remained unchanged: The Listening, Reading, Speaking the Overall scores were 6.0 and the Writing score was 5.5 both times. A small shift was only observable in the means, which for all scores increased by about 1/10 of a band. The change reached significance for the Writing and the Overall scores, but the effect sizes were small and driven by a handful of participants (nine in each case); for the majority, the scores remained unchanged, and for some they were reduced at T2 (***Figure 2***).

3 One participant in the coached group received an overall IELTS score of 2.0 at T1. This was lower compared to all other participants (their initial scores ranging from 4.0 to 7.5). This participant also showed a higher gain at T2 (2.5 bands) compared to the rest (range –0.5 to 2.0). To check the effect that this has had on the results, we repeated all the analyses with this participant excluded; the results did not change. For the sake of completeness, the full set of data is reported here.

Mann-Whitney's test statistics confirmed that IELTS gains experienced by the coached group were statistically larger than those seen in the control group (***Table 5***). The effect size for group differences was medium in the case of Writing and large for the other three skills.

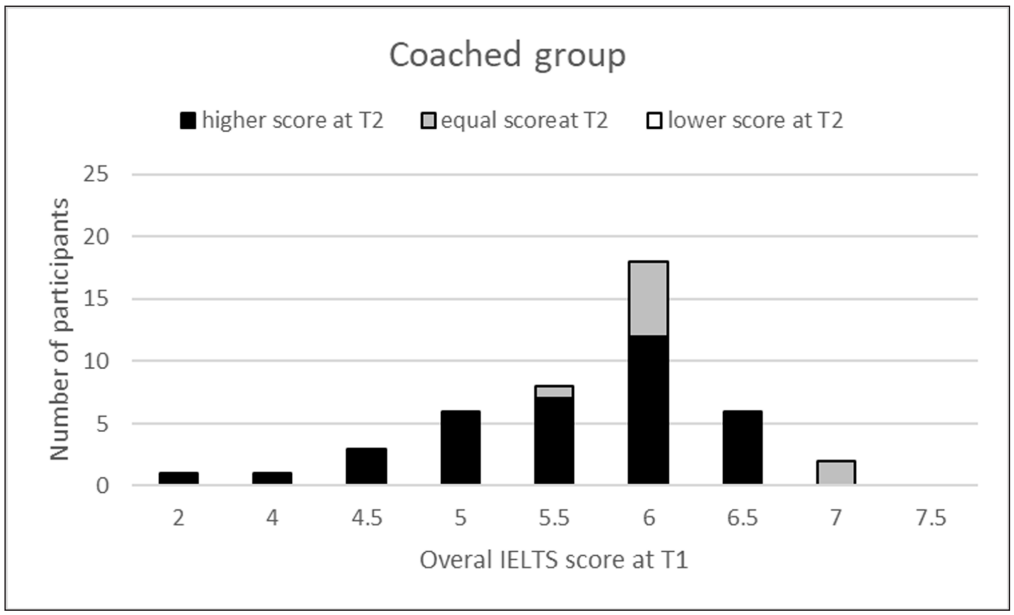

Figure 1 The proportion of participants in the coached group who improved the overall IELTS score at T2, broken down by the initial IELTS score.

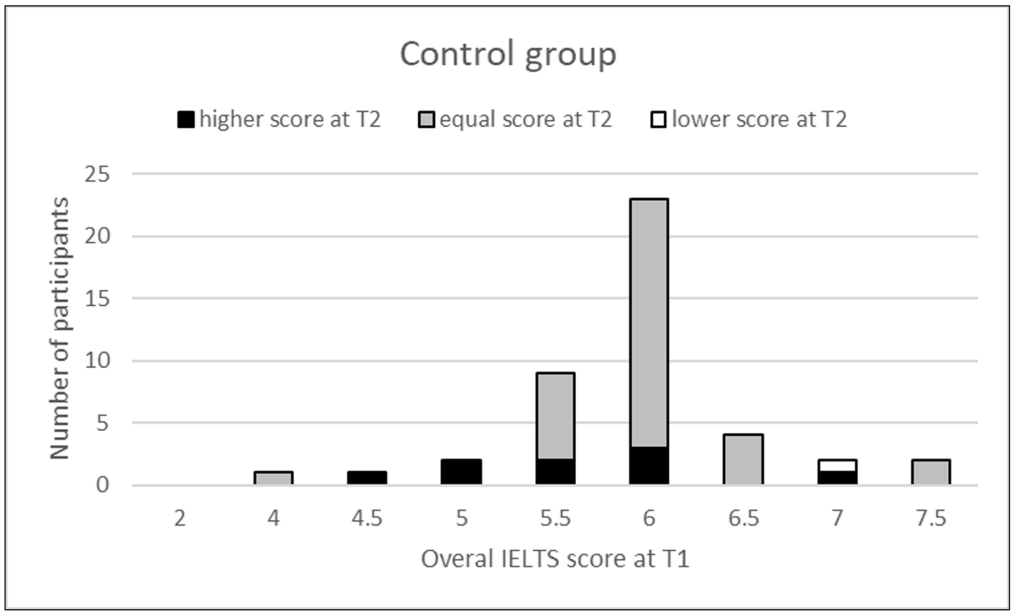

Figure 2 The proportion of participants in the control group who improved the overall IELTS score at T2, broken down by the initial IELTS score.

Table 4 Comparison of Time 1 and Time 2 scores in the control group (n = 44).

MEASURE	TIME 1			TIME 2			*z*	*p*	EFFECT SIZE *r*
	M (*SD*)	MEDIAN	RANGE	*M* (*SD*)	MEDIAN	RANGE			
IELTS overall	5.93 (.64)	6.0	4.0–7.5	6.02 (.61)	6.0	4.0–7.5	2.53	.011	.27
IELTS listening	6.08 (.90)	6.0	4.0–8.0	6.15 (.87)	6.0	4.0–8.5	1.60	.109	.17
IELTS reading	6.10 (.76)	6.0	5.0–8.0	6.19 (.71)	6.0	5.0–8.0	1.80	.074	.19
IELTS writing	5.49 (.67)	5.5	4.0–7.0	5.58 (.66)	5.5	4.0–7.0	2.14	.033	.23
IELTS speaking	5.70 (.68)	6.0	3.5–7.0	5.76 (.69)	6.0	3.5–7.5	1.73	.083	.18
Sent. reading speed	412.39 (115.68)	366	242–808	386.86 (95.25)	358	227–617	–4.24	.000	.45
Sent. comp.	76.86 (8.64)	77	52.0–94.5	77.56 (9.11)	77	55.0–97.0	1.72	.085	.18
Vocabulary	33.51 (7.57)	33	13–56	34.52 (6.87)	33	17–56	2.07	.039	.22
OOPT	51.66 (14.86)	49.5	17–91	53.61 (14.38)	56	19–86	1.28	.202	.14

DOI: 10.22599/jesla.74

The results of the control group demonstrate that simply repeating an IELTS test may in some cases lead to an improvement of overall score or some of the sub-scores. This could be a reflection of test-takers' improved familiarity with the test, the growth in the underlying ability or a measurement error. In the context of the present study, this is important because it suggests that a small part of the gain observed in the coached group may have stemmed from the same sources and may not be related to the test-preparation activities. The group difference in score gains, however, suggests that the largest part of the coached group's gain is most likely coming from the test-preparation activities.

Table 5 Group comparison (coached vs control) of Time 1-Time 2 score gains.

MEASURE	COACHED (*n* = 45)		CONTROL (*n* = 44)		MANN-WHITNEY *U*	*z*	*p*	EFFECT SIZE *r*
	MEDIAN	RANGE	MEDIAN	RANGE				
IELTS overall	.05	0–2.5	0	−0.5–0.5	1633.00	5.79	.000	.61
IELTS listening	1	−0.5–3.0	0	−0.5–0.5	1579.00	5.10	.000	.54
IELTS speaking	0.5	0–2.5	0	−0.5–0.5	1599.00	5.42	.000	.58
IELTS reading	0.5	−0.5–2.5	0	−0.5–1.0	1536.00	4.69	.000	.50
IELTS writing	0.5	−0.5–3.0	0	−0.5–1.0	1343.50	3.27	.001	.35
Sent. reading speed	−26	−188–68	−13	−191–29	1975.00	0.70	.485	.07
Sent. comp.	0.50	−22–12.50	0.75	−8–6.5	929.50	−0.50	.619	.05
Vocabulary	0.5	−6–7	1.0	−5–14	953.50	−0.30	.763	.03
OOPT	1	−43–39	2	−26–25	968.50	−0.18	.860	.02

3.2. OTHER SCORES AT T1 AND T2

To assess whether the IELTS gains generalise to alternative measures of English ability, we considered the participants' performance on the other three tests. Both groups made similarly large gains on their sentence reading speed: The sentences were read faster at T2 than at T1 (***Tables 3*** and ***4***). However, neither group saw an improvement in how many sentences they understood correctly. The increase in reading speed, therefore, is most likely a reflection of familiarity with the task. But the absence of improvement in comprehension suggests that the underlying knowledge on which the comprehension is based has not changed substantively between T1 and T2.

The control group also showed small but significant changes to the vocabulary score, which the coached group did not. However, this effect was small, and the group difference in gain was not significant. Finally, no significant improvement in English proficiency from T1 to T2 was detected for either group on the widely used measure of English proficiency, the OOPT.

In sum, in contrast to the IELTS test, both groups performed very similarly on all other tests at both T1 and T2. Put differently, the coached group's improvement on IELTS scores was not accompanied by a corresponding improvement in English-language ability as measured by the other tests in our study.

4. DISCUSSION

4.1. SUMMARY OF FINDINGS AND THEIR SIGNIFICANCE

In this study, we followed a group of 45 students as they prepared for an IELTS exam in a large test-preparation centre in China. After an intensive four-week preparation programme, their scores improved by 6/10 of an IELTS band on average. Although the participants with the lowest scores on entry stood to gain the most, even the participants with initial IELTS results of 6.0 and 6.5 reliably increased their scores after the training. In a control group of 44 students who were not undergoing any test preparation at the time, repeating the test four weeks apart also led to a statistically significant gain in the overall score, but at 1/10 of an IELTS band it was substantially smaller than that observed in the coached group. On the tests of vocabulary, sentence comprehension and the OOPT, the two groups performed similarly to each other at both testing points, with no measurable improvement from the first test. The large gains in IELTS scores observed in the coached group did not generalise to other measures of English ability.

DOI: 10.22599/jesla.74

The findings extend the current state of research in several important ways. First, they demonstrate that test-preparation centres in China may be more efficient in raising IELTS scores than preparation programmes in the United Kingdom, Australia and New Zealand, where much of the previous research was conducted. Although some of the earlier studies report a similar level of gain to the one observed here, this was typically achieved over two to three times longer periods of training (cf. Brown, 1998; Elder & O'Loughlin, 2003). In other studies, the reported gains were substantially smaller (Green, 2005, 2007; Read & Hayes, 2003). Furthermore, while previous studies observed training-induced gains primarily in students starting with IELTS scores of 5.5 and below, here we show that even students with initial scores of 6.0 and 6.5 can reliably improve their results. This has particularly important consequences for university admissions, as IELTS scores just half a band higher—6.5 and 7.0—are a typical requirement for unconditional entry at many universities. It suggests that for students who need a final push to get them over this important threshold, some short preparation courses may, indeed, be effective.

Second, by evaluating the gains of the coached group against a control group, we were able to tease apart the effects of test preparation from those of test repetition. Similar to the results reported for TOEFL (Zhang, 2008), the results of our study show that for test-takers not previously familiar with the IELTS test format, merely repeating the test four weeks later statistically improves the chance of getting a higher score. This highlights the importance of attaining familiarity with the format of a high-stakes test in order to perform to one's true level of ability (Messick, 1996). Critically, the difference in score gains between the coached and the control group demonstrates that teaching to the test can lead to a half a band rise in IELTS scores, over and above the improvement that could be attributed to familiarity with the test format or experiential growth in ability.

Finally, by including the additional measures of English ability and finding that the IELTS gain did not generalise here, our study empirically confirms that even in high-stakes language tests designed to minimise threats to validity through authentic and direct assessment, curriculum-narrowing practices and instruction in test-taking strategies have the power to weaken the extrapolation link from coached scores to what students are able to demonstrate in alternative contexts (Xie, 2013). Thus, while some courses may indeed be effective in raising scores, they appear to undermine the validity of such scores.

4.2. LIMITATIONS, ALTERNATIVE EXPLANATIONS, AND DIRECTIONS FOR FURTHER RESEARCH

As our study was conducted in a single test-preparation centre, it is important to consider whether and how far the findings generalise. In another study with 153 Chinese university students in the United Kingdom (Hu & Trenkic, 2019), we observed the same effect: Among students arriving with an identical IELTS score, those who achieved it by attending IELTS-preparation programmes did less well on alternative tests of English proficiency compared to students who achieved the same score without coaching. This occurred even though the participants attended a range of IELTS-preparation programmes, differing in length, provider and location. The finding suggests that similar curriculum-narrowing test-preparation programmes, with similar outcomes for IELTS scores and their validity, may be the norm across different training centres.

At present we do not know whether the same practices are employed by the language test-preparation industry outside of China and if so, whether they have the same effects. In addition to being the result of the properties of a training regime, the effects we observed here and in Hu and Trenkic (2019) could be a consequence of the broader cultural milieu in which these programmes were embedded. Memorisation is hugely important in Chinese education, not least because of the need for verbal rote learning of hundreds of logographic characters. But it is also greatly valued and admired in other contexts, as evidenced by the popularity of TV contests based on rote memory performance (Mattys et al., 2018). It is therefore possible that test-preparation courses that rely in great part on the narrowing of the curriculum and memorising answers to probable test questions are particularly effective for quick score gains in cultures that put a premium on verbal rote memory.

DOI: 10.22599/jesla.74

Unlike IELTS which tests academic reading, writing, listening and speaking, the alternative measures of English ability in this study all tested linguistic knowledge that underpins these skills. Could it be that coaching improved participants' academic reading, writing, listening and speaking skills without changing their general language proficiency, thus explaining the gain in IELTS scores without a change on other measures? This might be feasible but is theoretically unlikely. Academic language skills are tightly linked to and directly dependent on well-developed general language proficiency (Hoover & Gough, 1990). In fact, general language proficiency becomes more, not less, critical as academic language skills develop. For example, in both monolingual and bilingual school-age populations, measures of general language proficiency (such as vocabulary knowledge) explain greater variance in academic language skills (such as reading comprehension) in higher grades than in lower grades (Geva & Farnia, 2012). This is why a substantial gain on measures of academic reading, writing, speaking or listening without an associated improvement on measures of general language proficiency raises questions about the robustness and the interpretation of the gain. Future research could probe this further by including both measures of general language proficiency and alternative tests of academic language skills (e.g., TOEFL).

5. IMPLICATIONS AND CONCLUSIONS

Using IELTS as an example, our findings indicate that short but intensive curriculum-narrowing courses can reliably improve scores in high-stakes language-proficiency tests, but that such test-specific gains do not readily generalise to other measures of English ability. This raises important implications for test-developers, test-users and test-takers alike.

We wish to stress that our results do not question the soundness of IELTS as a test of English proficiency, nor its appropriateness for university admissions purposes. Rather, and to paraphrase Goodhart's law (Strathern, 1997), they underscore that every measure that becomes a target, also becomes a target for gaming. English-language tests for university entry, even when designed to measure language proficiency directly and using as authentic tasks as possible, appear to be no exception.

One way to lessen the attractiveness of gaming a test—and to make the alternative of more gradual gains through language-developing more appealing—is to space out the opportunities for taking the test. Until 2006, IELTS had a 90-day resit rule: A rule that acknowledged that language development takes time and that quick gains in scores are unlikely to reflect a corresponding improvement in English proficiency. The removal of this rule seems to have played directly into the hands of the extreme end of the test-preparation industry and encouraged other gaming behaviours (e.g., see Hamid, 2016, for a case study of a serial repeater who took IELTS 14 times within eight months, including three attempts within a single month). Given the results of the present study and of Hu and Trenkic (2019), re-instating the resit rule to protect the validity of the score interpretation and use would be a step in the right direction.

For receiving universities who use proficiency tests for admissions purposes, our results suggest that many students may have weaker English than their qualifications indicate, and therefore more extensive measures need to be in place to support their learning. Furthermore, universities themselves may be fuelling the dubious practices of the test-preparation industry by making offers that are conditional on unrealistic improvements in English proficiency within an application cycle. For test-takers, our study confirms that test-specific preparation programmes may help them to cross the threshold needed for a university place. However, it also raises concerns that without a corresponding improvement in English proficiency, this may be putting them at a distinct disadvantage in their studies.

Previous research has demonstrated that well-developed language and literacy skills are essential for success in every academic subject. Specifically in the context of higher education, students who arrive with English-proficiency scores recommended by test-developers do on average better than students who only meet the (typically lower) minimum requirements set for their programmes (Trenkic & Warmington, 2019); the latter, however, outperform peers who bypass these requirements altogether by gaining a direct entry into university through different pathway options (Eddey & Baumann, 2011; Oliver et al., 2012). If, as our study suggests, the test-preparation industry helps candidates improve test scores with

which they apply for university places but without an equivalent gain in their underlying proficiency, the gap between the level of English with which they arrive and that needed for fulfilling their academic potential may be even greater than previously assumed. Thus, although by no means the only culprit, the language test-preparation industry appears to be contributing to the situation where many international students struggle with the linguistic demands of their programmes and are constrained by their level of English in terms of what they can achieve.

AUTHOR AFFILIATIONS

Danijela Trenkic *orcid.org/0000-0001-6340-6030*
University of York, GB

Ruolin Hu *orcid.org/0000-0003-2153-4944*
University College London Institute of Education, GB

REFERENCES

Archibald, A. (2001). Targeting L2 writing proficiencies: Instruction and areas of change in students' writing over time. *International Journal of English Studies, 1*(2), 153–174.

Baddeley, A. D., Emslie, H., & **Nimmo-Smith, I.** (1992). *The speed and capacity of language processing (SCOLP) test*. Thames Valley Test Company.

Brown, J. D. (1998). Does IELTS preparation work? An application of the context-adaptive model of language program evaluation. *International English Language Testing System (IELTS) Research Reports, 1*, 20–37.

Cambridge University Press. (2020). Cambridge English Catalogue. *Cambridge English exams; IELTS 14*. *https://www.cambridge.org/gb/cambridgeenglish/catalog/cambridge-english-exams-ielts/ielts/ielts-14*

Daller, M., & **Phelan, D.** (2013). Predicting international student study success. *Applied Linguistics Review*, 4, 173–193. DOI: *https://doi.org/10.1515/applirev-2013-0008*

Deygers, B., & **Malone, M. E.** (2019). Language assessment literacy in university admission policies, or the dialogue that isn't. *Language Testing, 36*, 347–368. DOI: *https://doi.org/10.1177/0265532219826390*

Eddey, P., & **Baumann, C.** (2011). Language proficiency and academic achievement in postgraduate business degrees. *International Education Journal: Comparative Perspectives, 10*(1), 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.

Elder, C., & **O'Loughlin, K.** (2003). Investigating the relationship between intensive English language study and band score gain on IELTS. *International English Language Testing System (IELTS) Research Reports, 4*, 207–254.

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

Geva, E., & **Farnia, F.** (2012). Developmental changes in the nature of language proficiency and reading fluency paint a more complex view of reading comprehension in ELL and EL1. *Reading and Writing, 25*, 1819–1845. DOI: *https://doi.org/10.1007/s11145-011-9333-8*

Green, A. (2005). EAP study recommendations and score gains on the IELTS Academic Writing test. *Assessing writing, 10*, 44–60. DOI: *https://doi.org/10.1016/j.asw.2005.02.002*

Green, A. (2007). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education, 14*, 75–97. DOI: *https://doi.org/10.1080/09695940701272880*

Haladyna, T. M., & **Downing, S. M.** (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement, 23*(1), 17–27. DOI: *https://doi.org/10.1111/j.1745-3992.2004.tb00149.x*

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*, 472–487. DOI: *https://doi.org/10.1080/01596306.2015.1061978*

Hoover, W. A., & **Gough, P. B.** (1990). The simple view of reading. *Reading and Writing, 2*, 127–160. DOI: *https://doi.org/10.1007/BF00401799*

Hu, R. (2018). *The effect of IELTS test preparation and repeated test taking on Chinese candidates' IELTS results, general proficiency and their subsequent academic attainment*. PhD thesis, University of York. DOI: *https://doi.org/10.1080/13670050.2019.1691498*

Hu, R., & **Trenkic, D.** (2019). 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*, 1–16. DOI: *https://doi.org/10.1080/13670050.2019.1691498*

IELTS. (2019). *Guide for Educational Institutions, Governments, Professional Bodies and Commercial Organisation*. *https://www.ielts.org/-/media/publications/guide-for-institutions/ielts-guide-for-institutions-uk*

IELTS. (2020). *IELTS Scoring in Detail*. *www.ielts.org/ielts-for-organisations/ielts-scoring-in-detail*

Jakeman, V., & **McDowell, C.** (1996). *Cambridge Practice Tests for IELTS 1*. Cambridge University Press.

Liu, O. L. (2014). Investigating the relationship between test preparation and TOEFL iBT® performance. *ETS Research Report Series, 2*(1), 1–13. DOI: *https://doi.org/10.1002/ets2.12016*

Ma, J., & **Cheng, L.** (2016). Chinese students' perceptions of the value of test preparation courses for the TOEFL iBT: Merit, worth, and significance. *TESL Canada Journal, 33*(1), 58–79. DOI: *https://doi.org/10.18806/tesl.v33i1.1227*

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. DOI: *https://doi.org/10.1080/1358684X.2012.649176*

Mattys, S. L., Baddeley, A., & **Trenkic, D.** (2018). Is the superior verbal memory span of Mandarin speakers due to faster rehearsal? *Memory and Cognition, 46*(3), 361–369. DOI: *https://doi.org/10.3758/s13421-017-0770-8*

Meara, P., & **Buxton, B.** (1987). An alternative to multiple choice vocabulary tests. *Language Testing,* 4, 142–154. DOI: *https://doi.org/10.1177/026553228700400202*

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice. *Educational Psychologist, 17*, 67–91. DOI: *https://doi.org/10.1080/00461528209529246*

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241–56. DOI: *https://doi.org/10.1177/026553229601300302*

Messick, S., & **Jungeblut, A.** (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89*, 191–216. DOI: *https://doi.org/10.1037/0033-2909.89.2.191*

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. DOI: *https://doi.org/10.1080/03075070500095762*

Murray, N. (2010). Considerations in the post-enrolment assessment of English language proficiency: Reflections from the Australian context. *Language Assessment Quarterly, 7*, 343–358. DOI: *https://doi.org/10.1080/15434303.2010.484516*

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. DOI: *https://doi.org/10.1080/07294360.2011.653958*

Pollitt, A. (2009). *The Oxford Online Placement Test: The Meaning of OOPT Scores*. *www.oxfordenglishtesting.com*

Purpura, J. (2009). *The Oxford Online Placement Test: What does it Measure and How?* *www.oxfordenglishtesting.com/*

Rao, C., McPherson, K., Chand, R., & **Khan, V.** (2003). Assessing the impact of IELTS preparation programs on candidates' performance on the General Training reading and writing test modules. *International English Language Testing System (IELTS) Research Reports, 5*, 236–262.

Read, J., & **Hayes, B.** (2003). The impact of IELTS on preparation for academic study in New Zealand. *International English Language Testing System (IELTS) Research Reports,* 4, 153–205.

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. DOI: *https://doi.org/10.1186/2229-0443-3-12*

Spolsky, B. (1995). *Measured words*. Oxford University Press.

Strathern, M. (1997). Improving ratings: Audit in the British university system. *European Review, 5*, 305–321.

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: *https://doi.org/10.1017/S136672891700075X*

University of Cambridge Local Examination Syndicate. (2000). *Cambridge IELTS 2 Student's Book with Answers: Examination Papers from the University of Cambridge Local Examinations Syndicate*. Cambridge University Press.

Xie, Q. (2013). Does test preparation work? Implications for score validity. *Language Assessment Quarterly, 10*, 196–218. DOI: *https://doi.org/10.1080/15434303.2012.721423*

Yan, A. (2015). Test of credibility: How Chinese exam "cheats" threaten students' dreams of studying abroad. *South China Morning Post*. Retrieved May 6, 2020, from *https://www.scmp.com/news/china/money-wealth/article/1874818/test-credibility-how-chinese-exam-cheats-threaten-students*

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

DOI: 10.22599/jesla.74

TO CITE THIS ARTICLE:
Trenkic, D., & Hu, R. (2021). Teaching to the test: The effects of coaching on English-proficiency scores for university entry. *Journal of the European Second Language Association*, 5(1), 1–15. DOI: *https://doi.org/10.22599/jesla.74*

Submitted: 08 September 2020
Accepted: 27 November 2020
Published: 16 February 2021

Journal of the European Second Language Association, is a peer-reviewed open access journal published by White Rose University Press.

WHITE ROSE UNIVERSITY PRESS
Universities of Leeds, Sheffield & York