

**Effects of Test Preparation on Test Performance -
the case of the IELTS and TOEFL iBT Listening Tests¹**

Thi Nhan Hoa Nguyen

University of Melbourne

Abstract

The effect of test preparation on test performance has been debated for a long time in the testing area in general and in language testing in particular. Researchers are divided on this issue due to various factors such as language proficiency, background knowledge, course length and the test which candidates have to prepare for.

This study examines the effect of two test preparation programs (the IELTS preparation course and the iBT TOEFL preparation course) on 95 Vietnamese test takers destined for overseas study in English-speaking countries. The research focuses on the listening sub-test of these two test batteries. The students from the two test preparation groups took both the IELTS and the iBT TOEFL listening tests.

Findings reveal that the effect of test preparation can be seen more clearly on the performance of the IELTS listening test than on that of the TOEFL iBT listening test. The IELTS preparation group performed significantly better than the TOEFL preparation group on the IELTS listening test. In contrast, the TOEFL preparation group performed better than the IELTS

¹ The findings in this paper were presented at the conference "Teaching English to Speakers of Other Languages (TESOL) in the Internationalization of Higher Education in Vietnam", Hanoi (Vietnam), May 2007

preparation group on the TOEFL iBT listening test, but this difference was not significant. Reasons behind these different effects of test preparation on test performance are discussed.

1. Introduction: the necessity of a test preparation or coaching program

The effect of a test preparation or coaching program on test performance has been an issue of concern for all parties involved: test developers, researchers, teachers and students/test takers. From theoretical and pedagogical perspectives, researchers and test designers often highlight the value of a test which can measure the underlying ability of test-takers independently of how well they have prepared or crammed for the test. In other words, if coaching leads to a higher test score regardless of whether the candidate's ability has improved, the construct validity of the test is questionable. From a practical perspective, however, teachers and students/test takers often feel the need to be well-prepared before being tested and believe in the benefit of test preparation, especially before an important test.

In Vietnam, coaching programs are often highly valued due to the "exam culture" characteristics of society. It is thus likely that if a student sits an exam, they will have already attended a test preparation course. For example, students who plan to study in English speaking countries often attend an IELTS or TOEFL Preparation Course. Such programs have attracted thousands of Vietnamese candidates every year. Vietnamese students often expect that if they attend TOEFL or IELTS test preparation programs, then their scores on all skills will improve, especially on listening skills as listening has been considered to be the most challenging for them². Their expectations, however, are not always met.

² IELTS Annual Review 2000 to 2006 reported that the listening sub-bands of both Academic and General Training candidates from Vietnam were often lower than other sub-bands of reading, writing and speaking.

The reasons for the gap between their expectations and what actually occurs have not been researched in-depth to date in Vietnam. It is thus argued that from theoretical, practical and local perspectives, it is useful to investigate the effect of test preparation on test scores in general, and on listening test scores in particular in a Vietnamese setting.

2. The relationship between test preparation programs and test score improvement

In the last two decades, a number of studies have investigated the relationship between test preparation programs and test performance scores (Anderman & Power, 1980; Bachman, Davidson, Ryan, & Choi, 1995; Bangert, Kulik, & Kulik, 1983; Brown, 1998; Celestine & Ming, 1999; Geranpayeh, 1994; Hayes & Watt, 1998; Hayes & Read, 2004; Powers, 1985, 1986; and Thiel, 1995). The findings from this research, however, are to some extent contradictory: while a relationship between test preparation and test performance is acknowledged in some studies, doubts about this relationship have also been voiced. This complex picture can be seen in research conducted in both non-language and language testing contexts.

2.1 Research on the effect of coaching programs on non-language tests

In research into non-language tests, conflicting findings can be seen regarding the effect of coaching programs. Anderman and Power (1980) used the true-experiment method to study the effect of a special preparation program on the verbal part of the Scholastic Aptitude Test (SAT). The subjects were students at eight secondary schools. The researchers found that the special preparation program had very little impact on students' total scores in the SAT verbal section. Powers (1985) studied the effect of a special preparation program on scores of academic aptitude tests by using a representative sample (3%) from 5,107 candidates. His finding was similar to that of Alderman and Powers (1980): the coaching program had little influence on test-takers' scores.

In contrast, Bangert et al (1983) used a "meta-analysis" or "the analysis of analyses" method to investigate the effectiveness of coaching programs on achievement test scores. By statistically analysing the results of a large collection of individual studies, they concluded that coaching boosted achievement scores and that there was a positive correlation between the length of coaching period and the achievement score achieved. Similarly, Powers (1986) used a quantitative summary method to synthesize the results of ten previous studies on the effect of test practice on a number of test item characteristics. He found a strong relation between test preparation and (a) the length and complexity of the test instructions, and (b) the format of the test item.

2.2 Research on the effect of preparation programs on language tests

A conflicting picture can also be seen in the research on the effect of preparation programs on language tests. Using stratified sampling to compare the performance of 398 Malaysian students of Science and Arts at three levels of proficiency, Celestine and Ming (1999) found that IELTS preparation did not make a significant difference to the scores of either average or high proficiency students from different disciplines. They went further by explaining that a preparation course did not have any effect on the test scores because 'IELTS is a test of proficiency thus the knowledge of test-taking strategies cannot substitute for fluency' (p. 46). Similarly, when comparing the two test batteries: FCE and TOEFL on several aspects such as test content, test scores, and test preparation, Bachman et al (1995) reported that test preparation did not produce a significant gain in test scores although in their pilot study they had found some effect of test preparation on the FCE test performance.

Another study referring to the effect of IELTS preparation on test performance in Asian students was carried out by Hayes and Read (2004) in New Zealand. The pre-test and post-test of 23 students taking either of two courses: a test-focused and skill-based course (12 students) or an EAP focused and topic-based course (11 students) revealed that though

approximately half of the student population improved their scores, the difference was not significant. An earlier study by Hayes and Watt (1998) also focused on the effect of the IELTS test preparation on the test performance of Asian students. They found that a two-month program with a testing rather than a teaching focus did not improve students' performance. A similar finding was reported by Elder & O'Loughlin in their study of IELTS score gains of 112 students after a three-months of intensive English study in either Australian or New Zealand. Although the average score gain of these students was half a band overall with a slightly greater average improvement for listening, there was no advantage for the subset of students who had taken a course focusing specifically on test preparation.

In contrast, a number of studies have suggested that testing preparation programs do affect test performance. Brown (1998) compared the performance on an IELTS test of students in an IELTS preparation course and a more broadly focussed EAP course at the Hawthorn English Language Centre in Melbourne, Australia. He found a positive effect of the IELTS preparation program on the students' performance. However, the sample sizes of the two groups in Brown's study were rather small (9 vs. 5) and only writing skills were researched. Thus the findings of his study need to be interpreted with caution. Brown called for 'a replication of the study with a larger population sample' (p. 36). Hayes and Read (2004) found a significant difference in the mean score in the IELTS listening sub-test of 12 Asian students after a course focussed on IELTS preparation in an institution in New Zealand though their overall score improvement was not significant. In a larger scale study, Geranpayeh (1994) examined the comparability of TOEFL and IELTS scores across two groups (group A: 113 subjects and group B: 103 subjects). Group B had gone through the TOEFL preparation course and were more familiar with this test than group A. A test preparation effect was found as group B performed significantly better than group A.

The point that emerges from these studies on the effects of coaching programs is that researchers have been split in their findings regarding the effect of language test preparation on test performance. The differences in researchers' findings might be due to many factors such as the kinds of tests and/or test formats involved, the sample sizes, the background disciplines, the background cultures, and the language proficiency levels of the studies' participants. Consequently, more empirical research on this issue is needed to better understand the effect of test preparation on test performance.

3. The study

3.1 Research questions

This study, which is part of a larger multifaceted comparison of the IELTS and the internet-based TOEFL listening test, aimed to address the question of the relationship between test preparation and test performance on each of these measures and hence to determine whether they were equally amenable to coaching. It addressed two main research questions:

RQ 1: *What is the effect of test preparation on test performance on the IELTS listening test?*

RQ 2: *What is the effect of test preparation on test performance on the TOEFL iBT listening test?*

3.2 Research contexts and participants

Participants for the study were 95 Vietnamese students who were doing either the IELTS or TOEFL preparation courses in Hanoi for the purpose of maximizing their chances of studying overseas in English-speaking countries. The IELTS preparation group consisted of 48 candidates at the Australian Development Scholarships Project; the TOEFL preparation

group comprised 47 candidates: (i) 34 with Vietnamese Government Scholarships at the Language Training course of the Hanoi Foreign Studies University (HUFS) and (ii) 13 candidates at the Vietnamese American Training College (VACT).

The IELTS preparation group studied English for Academic Purposes and IELTS preparation. The IELTS preparation was 4-6 hours a week in which about 1.5 to 2 hours were devoted to listening skills such as listening skills training and doing IELTS listening practice tests. The program had been running for 10 weeks. The English language proficiency of this group of students ranged from IELTS band 5 to IELTS 6.5 with the listening sub-band from 4.5 to 6.5. The TOEFL iBT was introduced in the TOEFL preparation program for about 2 hours to all participants on the first day. After that, instructions were focused on teaching listening skills in general and on the TOEFL iBT listening test in particular. For the participants in the TOEFL group at Hanoi Foreign Studies University, the training for the TOEFL iBT listening test ran for approximately 4 hours per day and lasted for 2 weeks. For the participants of the TOEFL group at the Vietnamese American Training College, time constraints allowed for only a 2-hour per day training course on the TOEFL iBT listening test of 4 weeks' duration. The English language proficiency of the TOEFL preparation group ranged from 400 to 590 on the Institutional TOEFL PP with the listening sub-band from 10 to 31.

3.3 Instruments

Since for reasons of test security it was not possible to obtain operational versions of either test, the test materials used in the study were the IELTS Specimen listening test versions 2005 and the TOEFL iBT practice test online 2005. While this may place certain constraints on the validity of the study, it should be said that the IELTS Specimen practice test published by the British Council, IDP IELTS Australia, and University of Cambridge ESOL Examinations is closer to the real IELTS test than any

other commercial IELTS practice tests available. Similarly, the TOEFL iBT practice test available online 2005 closely simulates the real TOEFL iBT test.

3.4 Procedures

3.4.1 Data collection

All participants in the study took both the IELTS and TOEFL practice tests. In order to avoid any possibility of a test practice effect, 48 participants took the *IELTS Specimen 2005* listening test first and the other 47 took the iBT TOEFL listening test first. The data collection design can be summarized as follows (Table 1).

	<i>IELTS preparation group</i>		<i>TOEFL preparation group</i>	
	<i>Group 1</i>	<i>Group 2</i>	<i>Group 3</i>	<i>Group 4</i>
	<i>24 students</i>	<i>24 students</i>	<i>24 students</i>	<i>23 students</i>
Step 1	IELTS	TOEFL iBT	IELTS	TOEFL iBT
Step 2	TOEFL iBT	IELTS	TOEFL iBT	IELTS

Table 1: The data collection procedure in the study

3.4.2 Data analysis

The effect of test preparation on test performance was investigated from two perspectives: (1) a comparison of test performance (and of mean scores in particular) between two groups (the IELTS preparation group and the TOEFL preparation group), and (2) a differential item functioning (DIF) analysis which identifies variation in items difficulty between groups, or more precisely those items which are relatively more difficult for the IELTS preparation group than for the TOEFL preparation

group and vice versa. T-tests (using Version 13 of the SPSS program, 1994) were applied for the mean score comparison and the Rasch-based test of parameter invariance available in the Quest program (Adams & Toon, 2002), was used for the DIF analysis. The number of test takers in this study was only 95, which is admittedly rather small for the detection of DIF (see Camilli and Shepard 1994, McNamara and Roever 2006), and thus the findings reported below need to be interpreted with caution.

4. Results

4.1 RQ 1: What is the effect of test preparation on test performance of the IELTS listening test?

4.1.1 Analysis of raw scores

A summary of basic information on the IELTS listening test performance across the two test preparation groups is provided in Table 2 below.

<i>IELTS listening scores</i>	<i>Mean</i>	<i>Minimum</i>	<i>Maximum</i>	<i>Std Deviation</i>
IELTS preparation group (n = 48)	20.94	13.00	31.00	4.35
TOEFL preparation group (n = 47)	18.34	9.00	31.00	5.39

Table 2: Descriptive statistics of the IELTS listening test for each test preparation group

Evidence suggestive of a test preparation effect can be seen in the raw scores of both groups of test takers. Though the maximum score of the IELTS preparation group was similar to that of the TOEFL preparation group (31), the minimum score of the IELTS preparation group was higher than that of the TOEFL preparation group (13 vs. 9) as was the

overall mean score (20.94 vs. 18.34). The standard deviation of the IELTS preparation group was smaller than that of the TOEFL preparation group (4.35 vs. 5.39) indicating a greater level of homogeneity among the former group.

To examine if this difference between the two groups was statistically significant, an independent t-test was used (after confirming the assumption of a normal score distribution underlying the use of this statistic). This analysis showed a significant advantage for the IELTS test preparation group with regard to their mean scores on the IELTS listening test ($t = 2.586$, $p = .011$). The effect size of this mean score difference was medium at $d = 0.48$ (Cohen, 1988). We can therefore conclude that there was a clear effect of test preparation on the IELTS listening test.

4.1.2 Analysis of the number of attempted items

Further evidence of the effect of test preparation on IELTS test performance can be seen in the number of items attempted in the listening test by each test preparation group as follows (Table 3).

Number of attempted items	The IELTS preparation group (N = 48)		The TOEFL preparation group (N = 47)	
	Raw number of attempted items (N = 40)	% of attempted items	Raw number of attempted items (N = 40)	% of attempted items
Minimum	32	80	16	40
Maximum	40	100	40	100
Standard Deviation	1.88	4.86	6.56	16.39
Mean	38.50	96.25	30.55	76.38

Table 3: The number of attempted items in the IELTS listening test by test preparation groups

The IELTS preparation group attempted more items than the TOEFL preparation group in terms of minimum number (32 vs. 16) and average number (38.50 vs. 30.55). In addition, only 3 (or 6.38%) of the TOEFL preparation group as compared to 20 (or 41.67%) of the IELTS preparation group attempted all items (40/40). These figures together with the fact that the IELTS preparation group attempted on average approximately 20% of items more than the TOEFL preparation group can be taken as further confirmation of the effect of test preparation. An independent t-test shows that the percentage of IELTS items attempted by the IELTS preparation group was significantly higher than that of the TOEFL preparation group ($t = 7.889$, $p < .001$).

The reason for this significant difference might be that the IELTS listening test has a high proportion (70%) of items requiring written production (short answers) of one to three words. The IELTS test preparation group had been prepared for this test and thus they did better than their TOEFL

preparation counterparts not only with respect to the mean score but also in terms of the number of attempted items. The TOEFL preparation group, in contrast, was likely to be more familiar with the multiple-choice format and thus they tried most of the multiple-choice items but failed to complete a number of items requiring productive answers. In short, the significant difference in the percentage of items attempted by the two test preparation groups provides further evidence of *the effect of test preparation on test performance as far as the IELTS listening test is concerned*.

4.1.3 A DIF analysis of items

A detailed DIF analysis of all items in the IELTS listening test across the two groups was undertaken. The Chi-square analysis showed that there were 9 items showing significant differences between the two test preparation groups (p value <.05): items 2, 12, 14, 16, 20, 21, 26, 27 and 36. Interestingly, the first 5 items belong to section 1 and 2: designed to measure everyday spoken English whereas the remaining 4 items fall in the section measuring academic English listening ability. Everyday spoken English was thought to be easier than academic English by test takers. However, the DIF analysis showed that everyday spoken English and academic English were not equally easy or difficult across the two test preparation groups. The following graph (Figure 1), yielded by the Quest program illustrates these points.

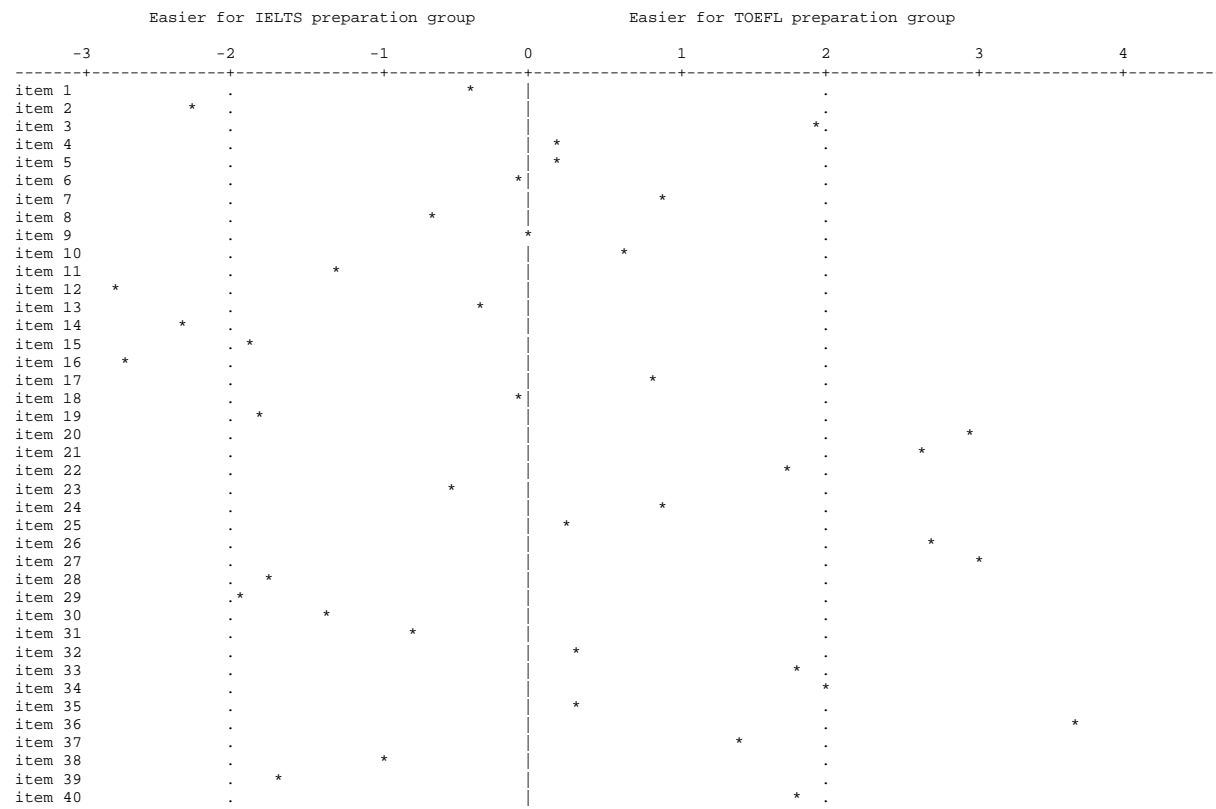


Figure 1: Plot of standardised differences in the IELTS listening test between the two test preparation groups

From the 0 point in the graph, all items on the left side were easier for the IELTS preparation group and all items on the right side were easier for the TOEFL preparation group. The items which were outside the parallel dotted lines showed evidence of bias: they were either significantly easier for the IELTS preparation group (items 2, 12, 14, 16) or significantly easier for the TOEFL preparation group (items 20, 21, 26, 27, 34, 36). These nine items constitute 22.5% of all items in the IELTS listening test.

What is most striking about the graph in Figure 1 are the points outside the parallel dot lines. These points demonstrate that:

- (i) with the exception of item 20, all items assessing everyday spoken English were significantly easier for the IELTS preparation group.
- (ii) all items assessing academic English were significantly easier for the TOEFL preparation group.

We could speculate that the TOEFL preparation group was more confident with academic English items perhaps because the TOEFL/TOEFL iBT listening test tasks were related to academic/university topics and the practice materials in their preparation course mirrored this academic orientation. Likewise, the fact that TOEFL prepared candidates performed at a lower level than the IELTS preparation group on the everyday spoken items may relate to the TOEFL/TOEFL iBT listening test not containing this genre. This reinforces and perhaps explains the previous finding of a significant *effect of test preparation on the IELTS listening test*.

4.2 RQ 2: What is the effect of test preparation on test performance on the TOEFL iBT listening test?

4.2.1 Analysis of raw scores

A summary of basic information in the TOEFL iBT listening test performance across the two test preparation groups is summed up in Table 4 below.

<i>TOEFL iBT listening scores</i>	<i>Mean</i>	<i>Minimum</i>	<i>Maximum</i>	<i>Std. Deviation</i>
IELTS preparation group (N = 48)	17.77	7.00	30.00	5.15
TOEFL preparation group (N = 47)	19.04	6.00	30.00	6.26

Table 4: Descriptive statistics of the listening scores across two test preparation groups

Unlike in the IELTS listening test, the effect of test preparation cannot be seen very clearly in the raw scores of the TOEFL iBT listening test. Firstly, the minimum scores across the two test preparation groups were similar (7 vs. 6). In addition, the two test preparation groups obtained a similar maximum score of 30. Although, the mean score in the TOEFL iBT listening test of the TOEFL preparation group was marginally higher than that of the IELTS preparation group (19.04 vs. 17.77) an independent t-test analysis showed that the difference in the mean scores of two groups was not significant: 17.77 vs. 19.04, $t = -1.081$, $p = 2.83$. This difference was also rather small as shown by Cohen's d ($d = .25$). These data indicate that there was *no significant effect of test preparation in the performance of the TOEFL iBT listening test*.

4.2.2 Analysis of the number of attempted items

In terms of the number of test items attempted, the two groups were also very similar in that they all tried every item in the test (34/34). This might be due to two factors: (i) the TOEFL iBT is designed in a way that test takers cannot move to the next item before attempting the previous one, (ii) this test only has multiple-choice or multiple-choice like format questions with a clock on the screen to let test takers know how much time is left for them to answer the remaining items. Therefore, test takers undertaking the TOEFL iBT in this research could not miss an item at random and they also tried to reach the last item within the given time. The fact that this behaviour was common to both test preparation groups is further evidence that *there was no significant difference between the two test preparation groups in the performance of the TOEFL iBT listening test.*

4.2.3 The DIF analysis of items

The DIF analysis procedure used for the TOEFL iBT listening test was to the same as that for the IELTS listening test (as described in Data analysis 3.4.2). The X^2 (Chi-square) analysis showed only four items for which there were significant differences in performance between the two test preparation groups (p values <.05): items 9, 10, 12 and 30. Figure 2 illustrates which items showed bias in favour of one or other test preparation group.

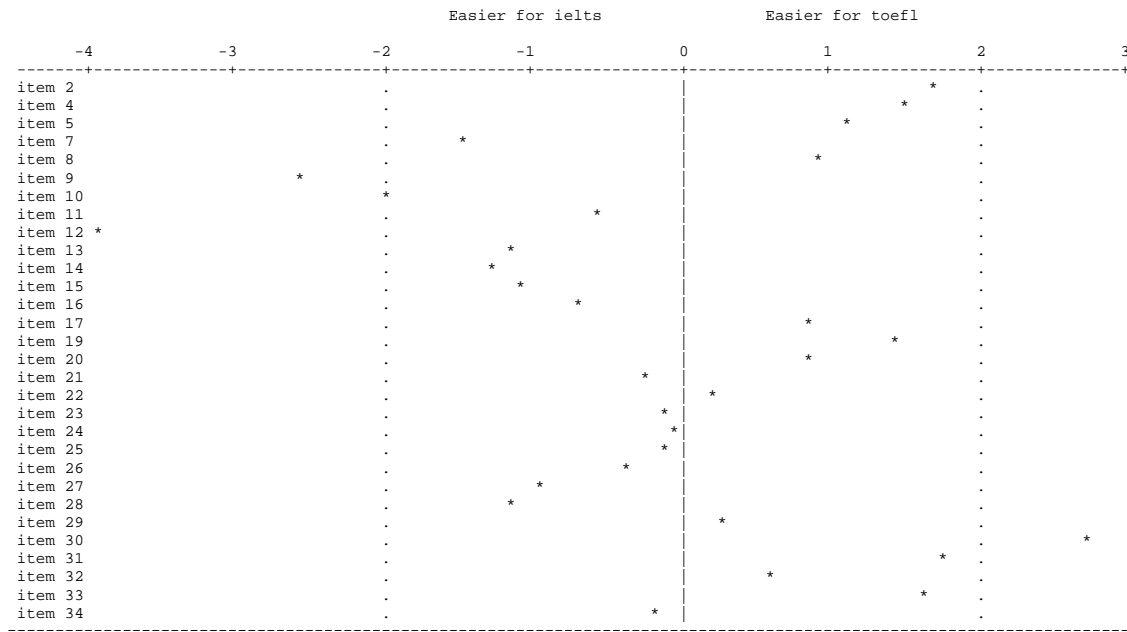


Figure 2: Plot of standardised differences in the TOEFL iBT listening test between two groups of test preparation

Note: item 1 and 6 have perfect score and thus do not appear on this map

From the 0 point in the graph, all items on the left side were easier for the IELTS preparation group and all items on the right side were easier for the TOEFL preparation group. The items which were outside the parallel dot lines showed evidence of bias: they were either significantly easier for the IELTS preparation group (items 9, 10, 12) or significantly easier for the TOEFL preparation group (item 30). These four biased items constituted 12.50% of all items in the TOEFL iBT listening test and belong to three different lectures of the TOEFL iBT listening test. Content analysis showed that item 9 and 30 ask for comprehension of details; item 10 requires a comprehension of implicature, and item 12 measures comprehension of global information.

Item 9, 10 and 12 were easier for the IELTS preparation group whereas item 30 was easier for the TOEFL preparation group. As items 9 and 30 asked for detailed information and each group did better in one item only, it cannot be said that the two groups were different in their ability to comprehend detailed information. Item 12 asked for global information regarding the topic of the lecture. The IELTS group found this item less challenging than did the TOEFL preparation group. However, there was only one item so there is not enough evidence to conclude that the IELTS group did better than the TOEFL preparation group in comprehending global meaning. The same caution should be applied in explaining the bias evident for item 10 which asks what the lecturer's implied as he said: *'I don't have to write that on the board, do I?'* Although the IELTS preparation group performed better on this item than the TOEFL preparation group this cannot be interpreted as evidence of superior ability with implicature. In any case, the group which did better on this TOEFL iBT item was the IELTS preparation group rather than the TOEFL preparation group

which, if it does have any meaning, can be seen as counter evidence for a test preparation effect.

6. Discussion and Conclusion

The various analyses of test scores (mean scores, number of attempted items and DIF analysis) of the two listening tests across the two test preparation groups provide suggestive evidence that the effect of test preparation (i) was significant on test performance of the IELTS listening test, and (ii) not significant on test performance of the TOEFL iBT listening test. These findings partly support and partly contradict those reported in the literature on the relationship between test preparation and performance in general and test preparation and performance on the IELTS and TOEFL in particular.

Firstly, the finding that there is a significant effect of test preparation on test performance of the IELTS listening test was contradictory to that of Celestine and Ming (1999) and Hayes and Watt (1998), who found no IELTS test preparation effect on test takers' performance scores. However, it was consistent with the findings of Elder and O'Loughlin (2003), and Hayes and Read (2004). Hayes and Read (2004) found a significant difference in the mean score in the IELTS listening sub-test of 12 Asian students after a course focus on IELTS preparation in an institution in New Zealand. Similarly, Elder and O'Loughlin (2003) found that the gain of test takers on the listening sub-band after three months of an intensive English program including some IELTS preparation was dramatic.

Secondly, the finding that there is no significant effect of test preparation on test performance of the TOEFL iBT listening test was consistent with what Bachman et al (1995) found in their main study as they compared two EFL batteries: the FCE (First

Certificate in English) and the IELTS (International English Language Testing Systems). It was, however, contradictory to that of Geranpayeh (1994) who found that among the two groups, the one, which had gone through the TOEFL preparation course, performed significantly better in the TOEFL test than the one, which had no test preparation. However, Geranpayeh (2004) used the old TOEFL as test materials for his research, rather than the new TOEFL which was not available until 2004. The results reported here might indicate that the new TOEFL (or the TOEFL iBT) test is likely to be better than the old TOEFL test in terms of construct validity. The TOEFL iBT is a more construct valid test because it is less amenable to test preparation than was the old TOEFL test: test takers' performance was not affected by a test preparation effect such as listening/test taking strategies taught in the preparation course. Such strategies cannot compensate for any lack of language skill.

There might be various reasons for the difference in the effect of test preparation across these two listening tests. The first reason might be the IELTS listening test is more complex than the TOEFL listening test in terms of format: test takers have to read questions, listen to the stimulus, and write down the answer at the same time. These combined activities are challenging and without sufficient practice and preparation, test takers might get lost.

In addition, there are several different question formats in the IELTS listening test and even in one section of this test the formats vary. For example, in section 1 of the IELTS listening test, the question formats are: locating positions on a map, multiple-choice, and filling in missing information in a table. Using different formats can avoid the test method effect on test performance but it may have a negative effect on test takers:

they have to be flexible and ready to switch on and off to provide different types of answers: selective or productive. This flexibility is likely to be developed by practising and getting more familiar with the test. Therefore it is not surprising that the TOEFL preparation group who were less familiar with the IELTS test attempted a significantly smaller percentage of items than their IELTS preparation counterparts. Another reason for the significant effects of test preparation on IELTS test performance was mentioned in the DIF analysis above: the IELTS listening test is richer than the TOEFL iBT listening test in genres: it has both academic and everyday spoken English whereas the TOEFL iBT only has academic-related genres; thus the IELTS preparation group which was familiar with and had practiced both genres did better than the TOEFL preparation group which had only prepared for the academic genre.

In contrast, the effect of test preparation cannot be seen clearly in the TOEFL iBT listening test probably due to the consistent format of the test: all questions are multiple-choice or multiple-choice like, all topics are academic-related, test-takers have time to listen, to note down information, and then to read and answer questions. Therefore test takers can concentrate on listening and then answering the questions without worrying about getting lost, locating the wrong item, or putting one, two or three words in a gap. Thus, we might tentatively draw the conclusion that the TOEFL iBT listening test is a more valid measure of test-takers' listening ability since it is not as amendable to coaching as the IELTS listening test.

Finally, it should be noted that in this research, the IELTS preparation group had a longer preparation course compared to their TOEFL preparation counterparts. More importantly, there was no independent listening proficiency test to measure if the two groups were equal in listening ability at the beginning of

the study. These are other factors contributing to the more visible effect test preparation had on test performance in the IELTS than in the TOEFL iBT listening test. These issues need to be acknowledged as limitations of this study. Further research aimed at overcoming these limitations is needed to confirm whether the IELTS listening test is indeed more sensitive to test preparation than the TOEFL iBT listening test.

References

- Adams, R. J., & Toon, K. S. (2002). Quest T90: the interactive test analysis system: A.C.E.R.
- Alderman, D. L., & Powers, D. E. (1980). The Effects of Special preparation on SAT-Verbal Scores. *American Educational Research Journal* 17(2), 239-251.
- Bachman, L. F., Davidson, F., Ryan, K., & Choi, I. C. (1995). *An investigation of comparability of two tests of English as a foreign language*. Cambridge: Cambridge University Press.
- Bangert, R. L., Kulik, J. A., & Kulik, C. C. (1983). Effects of Coaching Program on Achievement Test Performance. *Review of Educational Research*, 53(4), 571-585.
- Brown, J. D. H. (1998). An Investigation into Approaches to IELTS Preparation, with particular focus on the Academic Writing Component of the Test. In S. Wood (Ed.), *IELTS Research Reports* (Vol. 1, pp. 20-37).
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. CA: Sage.
- Celestine, C., & Ming, C. S. (1999). The Effect of Background Disciplines on IELTS Scores. In R. Tullloh (Ed.), *IELTS Research Reports* (Vol. 2, pp. 36-51). Canberra: IELTS Australia Pty Limited.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New Jersey: Lawrence Erlbaum Associates, Inc.
- Elder, C., & K. O'Loughlin. (2003). Investigating the relationship between intensive English language study and band score gain on IELTS. In R. Tolloh (Ed.), *IELTS Research Report* (Vol. 4, pp. 207-254). Canberra: IELTS Australia Pty Limited.
- Geranpayeh, A. (1994). Are score comparisons across language proficiency test batteries justified?: an IELTS - TOEFL

-
- comparability study. *Edinburgh working papers in applied linguistics* 5, 50 - 65.
- Hayes, B., & Read, J. (2004). IELTS preparation in New Zealand: preparing students for academic module. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing. Research contexts and methods* (pp. 97-112). Mahwah: Lawrence Erlbaum Associates.
- Hayes, B., & Watt, L. (1998). An IELTS preparation course for Asian students; when practice doesn't make perfect. *EA Journal*, 16, 15-21.
- McNamara, T., & Roever, C. (2006). Validity and the social dimension of language testing. *Language Learning*, 56(Supplementary 2), 9-42.
- Powers, D. E. (1985). Effects of Coaching on GRE Aptitude Test Scores. *Journal of Educational Measurement*, 22(2), 121-136.
- Powers, D. E. (1986). Relation of Test Items Characteristics to Test Preparation/ Test Practice Effects: a Quantitative Summary. *Psychological Bulletin*, 100(1), 67-77.
- Thiel, T. (1995). *An Analysis of the Evolution of the IELTS and an Investigation of Its Validity*. University of Tasmania, Hobart.