Assessment in Education: Principles, Policy & Practice

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/caie20

Washback to learning outcomes: a comparative study of IELTS preparation and university pre-sessional language courses

Anthony Green a

a University of Bedfordshire, UK

Published online: 25 Apr 2007.

To cite this article: Anthony Green (2007) Washback to learning outcomes: a comparative study of IELTS preparation and university pre-sessional language courses, Assessment in Education: Principles, Policy & Practice, 14:1, 75-97, DOI: 10.1080/09695940701272880

To link to this article: http://dx.doi.org/10.1080/09695940701272880

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the “Content”) contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions
Washback to learning outcomes: a comparative study of IELTS preparation and university pre-sessional language courses

Anthony Green*
University of Bedfordshire, UK

This study investigated whether dedicated test preparation classes gave learners an advantage in improving their writing test scores. Score gains following instruction on a measure of academic writing skills—the International English Language Testing System (IELTS) academic writing test—were compared across language courses of three types; all designed for international students preparing for entry to UK universities. Course types included those with a test preparation focus, those designed to introduce students to academic writing in the university setting and those combining the two. In addition to IELTS academic writing test scores, data relating to differences in participants and practices across courses were collected through supplementary questionnaire and test instruments. To take account of the large number of moderating variables and non-linearity in the data, a neural network approach was used in the analysis. Findings indicated no clear advantage for focused test preparation.

Introduction

Students are increasingly willing to travel internationally for their education and the English language has, in this context, served both as a major attraction of UK education and as a potential barrier to successful study. Students who struggle with the language are unlikely to benefit fully from university study. For this reason, before accepting international students onto their courses, receiving universities will usually, in addition to evidence of intellectual capacity, require evidence of sufficient English language ability.

With the growth in the international student population, there has been a corresponding increase in the use of language tests to screen applicants for language ability.

*Centre for Research in English Language Learning and Assessment, University of Bedfordshire, Putteridge Bury, Hitchin Rd, Luton, Bedfordshire, LU2 8LE, UK. Email: tony.green@beds.ac.uk
In the UK, the most widely recognized test of English for academic purposes is the International English Language Testing System (IELTS), jointly owned by the British Council, IDP (International Development Programme) Australia and Cambridge ESOL (English for Speakers of Other Languages) Examinations (www.ielts.org). The test consists of four modules (reading, writing, listening and speaking) and is available in two versions: academic (the version used for university admissions purposes and in this study) and general training (used for vocational training and immigration purposes). Requirements differ by institution, but universities will typically require a score of 6 or 7 on the nine band IELTS scale for unconditional entry. Over the decade from 1995 (when the test was last revised) to 2005, the worldwide IELTS candidature expanded by 1,000% from around 50,000 to over 500,000 (IELTS, 2005). The high-stakes nature of the test, coupled with its rapid growth, has attracted a burgeoning international industry in test preparation courses and text books.

This raises questions about how well test preparation courses fit with preparation for academic study. The educational consequences of using tests to regulate access to opportunities have worried educators ever since examinations came into widespread use for this purpose (see, for example, Herbert, 1889). It has often been asserted that tests tend to ‘narrow the curriculum’ and that ‘what is tested is what gets taught’. The effect of testing on teaching and learning is termed washback in the applied linguistics literature, although terms such as impact and testing consequences are often used with this sense in general education. In the case of English language screening tests like IELTS, it has been said that training to take a test may not develop the full range of academic language skills required for university study, particularly in the area of academic writing (Deakin, 1996; Turner, 2004).

A contrast is sometimes made between the content of IELTS and courses in English for Academic Purposes (EAP) which seek to address the specific language and study skills requirements of academic study (Blue, 2000). However, Hawkey (2006) observes that IELTS writing is a direct test that requires candidates to produce academic text types (an essay and a report). The test is, according to Hawkey, widely perceived by both teachers and students as meeting the needs of its constituency of intending undergraduates and postgraduates. IELTS test preparation might therefore be expected to be more relevant to academic writing needs than preparation for alternative tests that do not include a direct test of writing, or that assess writing skills through story-telling or informal letter-writing tasks.

The study reported here investigates test-directed courses, courses in EAP and courses that include both strands to determine whether dedicated test preparation is successful in delivering greater short term gains in writing test scores over the length of a course. To take account of the numerous variables that might moderate the relationship between course type and course outcomes (such as differences in learner background, course length and exposure to English outside class), the study employs an artificial intelligence approach to the data analysis using a neural network. This approach has the advantage over traditional statistical techniques that it is better able to take account of the complex, non-linear associations between the variables involved.
International student access to higher education in the UK

To gain access to university courses in the UK, international students will usually need to present an acceptable score on a recognized screening test such as IELTS or its American equivalent, the TOEFL (Test of English as a Foreign Language, Educational Testing Service, Princeton, USA) test. If their scores fall below the level required for entry, they may nonetheless be accepted on the condition that they successfully complete a language course provided by the university to bring their skills up to the required level. In effect, international students who fall below the immediate entry requirements are faced with a choice between taking a test preparation course in the hope of improving their test scores and so gaining unconditional entry, enrolling on an intensive language course provided by their chosen university or finding a course that combines these elements. These three types of course are the focus of the study reported here.

To become successful writers, international students need to acquire culturally embedded knowledge. This might include learning about the role of writing in studying an academic subject, permissible uses of source material (including the avoidance of plagiarism) and the place of critical thinking in written work (Horowitz, 1986; Zamel & Spack, 1998; Casanave, 2002; Turner, 2004). EAP courses designed for international students by universities will usually take account of these needs. However, for reasons of equity, access and practicality, such skills, which may be required to varying extents in local contexts, may not properly be addressed in an international language test such as IELTS (Fulcher, 1999; Clapham, 2000).

If it is true that IELTS does not test the full range of skills required by international students to support academic writing and that what is tested is what gets taught, it is possible that learners choosing to rely on IELTS preparation courses may, by concentrating closely on the demands of the test, achieve better test results without necessarily acquiring some of these key academic writing skills.

Investigating washback

In investigating the effects of a test on educational practices a number of perspectives are available to the researcher. One approach is to consider test design features and their likely effects on instruction. Chapman and Snyder for example (2000) suggest a typology of test features that may impact on educational practices. In relation to IELTS, Moore and Morton (1999) compare university writing assignments with IELTS writing Task 2 (see Appendix for an example) and conclude that test preparation will involve a relatively narrow range of writing activities in the classroom. Thorp and Kennedy (2003), Mickan and Slater (2003) and Coffin (2004) arrive at similar conclusions on the basis of IELTS test responses. However, none of these studies incorporated evidence of what actually happens in IELTS preparation classes.

Empirical washback studies have usually involved investigating the effect of testing on aspects of educational systems. Hughes (1993) draws a useful distinction between effects on participants, processes and outcomes. The influence of a test on participants...
A. Green

(the teachers, learners and materials writers preparing for a test and the perceptions and attitudes they bring to the task), leads them to modify their processes (teaching and learning behaviours) and these in turn impact on products (learning outcomes including knowledge of target skills and test scores). Although the effects of standardized testing on educational outcomes and the interpretability of test scores have been central questions in the general education literature (see, for example, Amrein & Berliner, 2002; Rosenshine, 2003), products have attracted relatively little attention from language testing researchers. No previous washback study has linked together test design, participant, process and product variables to trace the influence of a language test through teaching and learning processes to test score outcomes.

A popular approach to investigating washback has been through questionnaire and observation based case studies of participants and processes (Alderson & Wall, 1993; Shohamy et al. 1996; Burrows, 1998; Ferman, 2004; Qi, 2004; Watanabe, 2004; Cheng, 2005; Wall, 2005). A number of such studies have addressed the effects of the IELTS test on the attitudes and classroom practices of teachers. Deakin’s (1996) questionnaire-based study, together with Brown (1998), Hayes and Read (2004) and Hawkey (2006), which include classroom observation data in addition to surveys, suggest that the greater the concentration on IELTS test demands in test preparation, the greater the tendency to reduce or exclude attention to academic writing skills that fall outside the immediate requirements of the test (notably the integration of sources, diversity of text type and academic register). However, because they focus only on IELTS preparation classes, most of the studies have been unable to establish how teachers might change their behaviour if they were not preparing students to take a test.

Brown (1998) is able to cast some light on this by including a comparison with a university course in EAP that was not directed towards IELTS. He found that this course, which aimed to prepare students for academic study rather than to take the test, incorporated elements that were not typically featured in IELTS preparation. These included the collection and integration of source materials, referencing, routine redrafting and a requirement that all students should produce a piece of extended, project-based writing. However, in common with the other observational studies, he failed to include a systematic description of how the design of the IELTS test might be expected to influence teaching and learning. This limits the claims he is able to make about whether it is the test that has caused the differences he observed between courses or variation in teacher beliefs about learning of the kind found in other washback studies (see Burrows, 1998, for a discussion of teacher responses to assessment).

Equally, none of the studies succeed in relating differences in course content to differences in course products in the form of test score gains. Both Hayes and Read (2004) and Brown (1998) include test scores as a variable, but are prevented by restricted sample sizes from reaching meaningful conclusions on this issue.

Weir and Green (2002) drew on a review of theories of academic writing and a model of writing task concerns from Weigle (2002) to build a systematic description of the IELTS test. From this they made predictions about the likely impact of the design of the test on classroom practices and carried out an observational study to
investigate these predictions. On the basis of the study, they were able to categorize the courses offered to international students preparing to study in the UK into three types according to participant expectations and course content:

1. Test preparation programmes that focus specifically on the IELTS test. Participants take these courses with the aim of succeeding on the test. Course content is closely modelled on the test design (see Appendix for an example of the IELTS academic writing test). IELTS preparation courses typically involve frequent timed writing practice with tasks modelled on the test. Writing topics are typically of general interest such as ‘sport’ or ‘the environment’ and responses are based on personal opinions or given data.

2. Pre-sessional courses in English for Academic Purposes (EAP). These are courses provided by universities to introduce language skills required specifically in university settings and to orient international students to the university environment before they begin their studies. Students take the course to gain entry to university and to learn the skills required for academic study. Course content is broader than on IELTS preparation courses. These courses may include some timed writing practice, but are not restricted to data description and brief discursive essays (the format of the two IELTS compositions). Attention is given to such academic writing skills as integrating sources and summarizing texts. Topics include both general interest topics and topics drawn from students’ chosen disciplines. Responses are based both on personal opinion and on textual sources or on data collected by students.

3. Courses that combine features of both. These combination courses are pre-sessional EAP courses which include an explicit IELTS preparation strand. Students on these courses are preparing both to enter a specific university and for the IELTS test.

To learn more about the implications of test preparation in this context, this study builds on Weir and Green’s (2002) categorization to compare the success of courses of these three types in improving students’ IELTS writing scores. The main focus of the paper is on whether learners on courses involving IELTS preparation (Course Types 1 and 3) are able to make greater gains in their test scores than those on courses that do not concentrate on the test (Course Type 2).

**Variables affecting language learning**

Investigating differences in product between courses raises an important issue for the researcher. Account must be taken of the effects of variables other than course type that are likely to influence rate of learning and that may be present to different degrees in each of the courses included. Such variables might be expected to include *inter alia* participant variables such as the level of the students’ ages, native language backgrounds, writing abilities and general language proficiency at course entry together with process variables including the length and intensity of the courses and opportunities to use English outside class.
Numerous variables have been suggested and investigated as having the potential to promote or retard the rate and degree of second language acquisition and so might mediate the washback effect (see Skehan, 1989; Spolsky, 1989; Ellis, 1994). In addition to amount and type of instruction, variables researched include learner background (age, social class, first language [L1]), and psychological factors including intelligence, personality, motivation, language aptitude and language learning strategies. In a review, Spolsky (1989) lists some 79 key variables that may exert an influence and suggests that these interact in complex ways.

There is plentiful evidence that individual learner differences are an important factor in mediating EAP test success. Weir (1983) incorporated a questionnaire addressing a number of social and psychological indicators, finding features such as age and educational experience to be significant predictors of EAP test performance. Hawkey (1982) included personality and attitude factors. Although he found language proficiency to be a strong predictor of EAP test success, cognitive style and attitude also proved to be significant. In a regression study of overall IELTS score gains on ten to twelve week preparation courses, Elder and O’Loughlin (2003) found that the kinds of accommodation that learners had lived in, course level, educational qualifications, and reading proficiency together provided the best predictions of gain in overall IELTS scores. A moderate relationship between nationality and score gain also emerged. These findings underline the importance of addressing such variables in the current study.

While Hawkey (1982) and Elder and O’Loughlin (2003) used linear regression for their analyses, the increasing recognition of the need for explanatory models and the complexity of the interrelationship between large numbers of learner variables have led to innovation in statistical analysis. In recent years, researchers exploring language gain (although not in the context of washback research) have used techniques such as structural equation modelling (Wen & Johnson, 1997; Purpura, 1998), cluster analysis (Skehan 1986) and neural networks (Boldt & Ross, 1998; Hughes-Wilhelm, 1997, 1999). These techniques have been employed because they are able to accommodate a wide range of interacting variables and to model their relationships in predicting test outcomes.

Hughes-Wilhelm’s (1997, 1999) project investigating rates of progress through an intensive English programme at a US university is of particular relevance to the current study as it involves an EAP course in the destination country, relating learner background and learning processes to course outcomes. An array of 70 student background variables was collected through hour-long interviews with participants. Fifty-seven language learning variables were selected as potential predictors of success (success being operationalized through class grades, rate of progress and course completion). Although entry proficiency (TOEFL test score) was the best single predictor, accuracy of prediction was considerably enhanced by the inclusion of other learner characteristics. Among the additional features contributing most to prediction were:

- communicative use of English both in school (reading and writing) and outside formal education (speaking and listening)
IELTS language courses

- school success and prior experience of academic study
- type and source of exposure to English, and
- individual characteristics (self-confidence; attitudes and motivations when learning English) (Hughes-Wilhelm, 1999, p. 449).

In relation to IELTS, if preparation programmes are successful at exploiting the characteristics of the test and if pre-existing differences between groups of learners and course organisation are taken into account, we would expect to see greater improvement in IELTS scores on dedicated test preparation courses (Type 1 and Type 3) than on those pre-sessional EAP courses not directed towards the test (Type 2).

The study

Aims

This study investigates washback to outcomes by comparing learner performance on the three course types. IELTS writing tests were administered at course entry and exit and a gain score for each learner calculated as the simple difference between these entry and exit scores. As both participant and process variables other than course type might account for any differences in mean score gains found in the study, data relating to course length, course intensity (hours of study per week) and individual characteristics, beliefs and attitudes considered likely to mediate washback were accessed through questionnaires and course documentation. A predictive model incorporating these variables was developed to account for score gains and to establish whether courses involving IELTS preparation (course Types 1 and 3) were more successful than Type 2 courses in improving IELTS writing scores in this context.

Participants and settings

Participants represented an opportunity sample of international students preparing for academic study at fifteen institutions in the UK. These institutions were selected following an earlier survey of UK course providers; they were willing to participate and were conveniently located. In all 663 students volunteered to participate in the research with 476 of these completing both entry and exit forms of the IELTS academic writing test: an overall response rate of 71.8%. Over 50 nationalities were represented, the largest cohort being from the Peoples’ Republic of China (34%) with large numbers also coming from Taiwan (19%).

Following Weir and Green (2002), the courses were categorized into three types (Table 1). the largest numbers of student participants (331) were studying on one of six Type 2 pre-sessional EAP courses (with no IELTS component) provided by three universities for intending international students. It was more difficult to access large numbers of learners on IELTS preparation courses as these were often provided by smaller, private institutions and always involved smaller numbers of learners than the pre-sessional EAP courses. Sixty individuals, each studying at one of four universities, were attending Type 3 combination courses: pre-sessional EAP courses with IELTS
preparation strands. The remaining 85 participants were attending Type 1 *IELTS preparation* courses at one of 2 state sector colleges of further education or one of five private sector language colleges. Courses ranged from 4 to 14 weeks in length and involved between 15 and 28 hours per week of classroom instruction. All but three of the learners involved were working towards entry to UK higher education. As all were either intending to take IELTS or to enter university at the end of their courses, these students could be said to represent the kinds of learner for whom the IELTS academic writing test is intended.

**Procedures**

To measure writing score gains, all participants took an IELTS writing test. Four IELTS writing test tasks (two versions of Task 1 and two of Task 2) were made available by the IELTS partner organizations and assembled into four linked forms: one to be administered at course entry and one at course exit at each participating centre. A crossed design was used to ensure that no learner encountered the same task on both occasions. Each test was scored by at least two trained IELTS writing examiners using the official nine-band rating scale. Differences in the scores awarded by these examiners were reconciled through a multi-faceted Rasch procedure which takes account of differences in the severity of individual raters (Linacre, 1988).

Of key relevance to the research questions are the improvements made by learners in their IELTS writing scores as measured through the IELTS writing tests administered at course entry and exit. However, multiple participant and process variables were identified through the literature review as having the potential to moderate the effects of instruction on score gains. To access these variables a variety of supplementary instruments were used.

Clearly, pre-existing differences between learners at course entry (participant variables such as age, first language, motivation and previous experience of learning English) and differences in experience (process variables such as course length, use of English outside class) needed to be taken into account. These were targeted through supplementary research instruments. Drawing on theories of second language learning (Skehan, 1989; Spolsky, 1989) and studies of learning gains on EAP courses (Hawkey, 1982; Weir, 1983; Hughes-Wilhelm, 1999; Elder & O’Loughlin, 2003) a set of moderating variables was identified that might affect learning gains. These included the participant variables (Hughes, 1993) that learners brought to their courses and process variables that captured their experience during their courses.

<table>
<thead>
<tr>
<th>Course Type</th>
<th>Institutions</th>
<th>Number of Students</th>
<th>Mean Course Length</th>
<th>Mean Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1: IELTS preparation</td>
<td>7</td>
<td>85</td>
<td>9 weeks</td>
<td>21 hrs/ wk</td>
</tr>
<tr>
<td>Type 2: Pre-sessional EAP</td>
<td>3</td>
<td>331</td>
<td>8 weeks</td>
<td>24 hrs/ wk</td>
</tr>
<tr>
<td>Type 3: Combination</td>
<td>5</td>
<td>60</td>
<td>9 weeks</td>
<td>23 hrs/ wk</td>
</tr>
</tbody>
</table>
Instruments relating to these variables were developed, piloted and collated into questionnaires to be administered at course entry and exit. Each learner responded to 291 questionnaire items, 160 grammar and vocabulary test items and composed four IELTS writing test responses. A full discussion of these lengthy questionnaire instruments is beyond the scope of this paper, but they are provided as appendices in Green (2007, forthcoming) and an outline of questionnaire content is given in Table 2 below.

In summary the following instruments were used to access participant and process variables:

1. Questionnaire A. A questionnaire for students addressing participant variables (to be administered at course entry)
2. Questionnaire B. A questionnaire for students addressing process variables (to be administered at course exit).
3. Two forms of a test of vocabulary based on the Vocabulary Levels Test (Nation, 1990) and a test of grammar for academic purposes developed from Weir (1983). These were also administered in a crossed design at course entry and exit.
4. Two brief protocol forms addressing (page 1) knowledge about the IELTS academic writing component and (page 2) the use of test-taking strategies. These were administered with each writing test—the first page before the test, the second after the test.

The course entry questionnaire (Questionnaire A) targeted participant variables including learner background (taking account of first language, region of origin, educational experience and previous use of English) as well as motivation for study. An example question would be ‘I am taking this course because I want to get a good grade on IELTS’. Responses would be given using a five-point Likert scale ranging from ‘strongly agree’ to ‘strongly disagree’.

Other sections of the questionnaire targeted attitudes towards the host culture (e.g., I usually enjoy meeting English people), confidence in own writing ability (e.g., I don’t think I write in English as well as other students), expectations of course content (e.g., I expect to take practice tests in class), educational history including any previous IELTS score, and future intention to take an IELTS test. At the same time, all participants were administered a measure of their general English language proficiency (a grammar and vocabulary test); were tested on their knowledge of the format and content of IELTS; and, immediately following their first IELTS writing test, were given questions about their use of test-taking strategies (e.g., I read the question carefully and underlined or highlighted key words).

At the end of their courses, together with their second IELTS writing test, all students took a second test of English grammar and vocabulary. As at entry, the test was administered together with a test of knowledge of the IELTS writing test design and questions about test-taking strategies. A questionnaire administered at course exit targeted process variables including details of class activities (e.g., I learned how to organize an essay to help the reader to understand); learning strategy use (e.g., I
studied vocabulary in my free time); self-assessed learning gains (e.g., After taking this course, I would feel more confident about writing assignments at university), use of English in and out of class and approach to learning (e.g., I often think about and criticize things that I hear in lessons or read in books).
Data analysis

Relating participant and process variables to writing score gains involved comparing test scores at course entry with test scores at course exit. Initially t-tests and analysis of variance were used to compare entry and exit scores and to make tentative comparisons between course types. Correlations were employed to explore the participant and process variables and to establish which of these might have an effect on score gains. However, a more sophisticated approach to analysis was required in order to establish whether there were substantive differences between course types once the many differences between courses and students (accessed through the questionnaires and tests) had been taken into account. An artificial intelligence application was chosen for this analysis: a neural network developed using NeuroSolutions 4.16 (Principe et al., 2000).

The choice of a neural network method of analysis was dictated by the large number of interrelated variables (or features in neural network nomenclature) addressed by the study. Neural networks were originally designed to emulate human learning by reproducing the structure of the human brain. They are widely used in text and speech recognition, financial forecasting and a variety of engineering applications. They differ from more familiar statistical methods such as multiple regression or multivariate analysis of variance (MANOVA) in several respects. Linear regression and MANOVA models are programmatic, involving a model determined by the researcher and tested against the data. In contrast, neural networks are adaptive; data is presented to the network case by case and the system ‘learns’ by adjusting its parameters through an automatic process of feedback, governed by rules set by the researcher. Somewhat different outcomes with varying levels of predictive accuracy will be obtained on each training occasion. For this reason each network was trained four times with the initial network weights being reset on each occasion. Average values over the four training runs are presented in the findings.

The advantages of neural networks for studies of this kind lie in this adaptivity. Unlike traditional statistical models, they require no assumptions about the linearity of relationships between variables or about patterns of distribution in the data. They are robust with respect to missing or incomplete data and can operate with large numbers of variables relative to the number of cases: a liberal rule of thumb being ten cases for each variable (Garson, 1998). Neural networks are not yet widely used in educational research, but have been employed as an alternative to multiple regression or analysis of covariance in studies that, in common with the research reported here, predict language course outcomes from multiple sources of data (Hughes-Wilhelm, 1997, 1999; Boldt & Ross, 1998).

Although neural networks are flexible in handling large numbers of input features, generalization may be poor if too many are used. To reduce the number of features for inclusion in the model, scores on each feature were compared with writing score gains and only those that displayed significant correlations were retained.

Following this initial screening, three sets of neural networks were developed from variables that were significantly (p<0.05) correlated with writing scores at
course exit. The first was based on participant features; a second on process and a third incorporated the strongest predictors from the other two sets. A series of networks were constructed. Two-layer networks (with input and output elements) effectively perform linear regression, while three-layer networks include a third or ‘hidden’ layer to allow for non-linear relationships between the input and desired sets. It was thus also possible to compare the success of linear and non-linear approaches to prediction.

In determining the most appropriate design for the networks, the number of input and output processing elements (PEs) (shown as circles in Figure 1) is determined by the nature of the research questions. The number of PEs in the input layer (labelled $X_1$ to $X_3$ in Figure 1) is equal to the number of features in the input set (the number of independent variables under investigation). The number of PEs in the output layer is equal to the number of features in the desired set (the dependent variables). The number of PEs in the hidden layer (marked as $U_1$ and $U_2$ in Figure 1) is not determined by any theory or set of predetermined guidelines, but is established through experimentation with the aim of finding the simplest architecture capable of solving the problem.

In the case of this study, the number of PEs at input ranged from just one (initial IELTS writing score) to 34 (for the widest range of participant, process and product features). A single PE formed the output layer, corresponding to the single feature being predicted (scores on the exit IELTS writing test). The number of PEs in the central, hidden layer was determined for each case through experimentation.

Each network was trained with randomized sequences of individual cases or input vectors. Sixty percent (286 cases) of the input vectors formed a training set. The networks are initially trained with this set of data, so that the machine ‘learns’ the

Figure 1  Example of a neural network architecture with three inputs, two PEs in the hidden layer and one output
problem. As each case is passed through the network, the predicted value of the output (in this case the exit writing score) is compared with the actual value. The error—the discrepancy between the predicted and actual value—is used to adjust the weighting (the lines labelled W1 to W8 in Figure 1) applied to the inputs in predicting the output.

Twenty percent (95 cases) of the input set was used as a cross-validation set. The cross-validation set is used to help the researcher to decide when the network has ‘learnt’ from the data. A further 20% were retained as a testing set. The testing set is used to test the accuracy of the network in applying the prediction model to data that has not been used in the training process.

If the process of training continues unchecked it is possible for a network to ‘overtrain’, obtaining very accurate results for the specific set of data under consideration, but not for previously unseen cases. To prevent such an outcome, the training was halted at the point of optimum generalization; the point at which there was least error in estimation of the cross-validation set. This point is reached when the error term—the mean squared error (MSE)—calculated between the predicted values generated by the neural network for the cross-validation data set and the actual values in the desired set reaches its lowest value. If overtraining occurs, the MSE will continue to decline in the training set, but will increase for the cross-validation set as generalization of prediction declines.

**Results**

*Writing score gains*

The purpose of the analysis was to establish whether learners on each course type had made measurable gains in test scores following instruction in writing skills and whether there were any significant differences between course types in this respect once additional moderating participant and process variables such as course length and learner background had been taken into account.

Paired t-tests were used to investigate whether learners had made score gains following their courses. The results indicated that a significant gain in writing scores had indeed occurred on all three course types (Course type 1, $t = -3.071$; Course Type 2, $t = -5.973$; Course Type 3, $t = -3.657$; $p < .01$ in all cases). Taken as a whole, the learners improved their IELTS academic writing scores by an average of 0.207 of a band on the nine-band IELTS scale.

The mean score gains on the three course types are displayed in Table 3. For comparison, the mean score on the IELTS academic writing test reported on the IELTS website for the worldwide candidature in 2005 was 5.69. The results displayed in Table 3 show that learners on all course types improved their IELTS writing scores, but that mean score gains were limited relative to the nine band IELTS scale. Learners on Type 3 combination courses made the most gain in their scores, but also started from the lowest levels at course entry.

There is little support here for the belief expressed by course providers and teachers in Weir and Green (2002) that courses directed towards the IELTS test would be
more effective than broader-based pre-sessional EAP courses in boosting IELTS writing scores. In brief, the picture to emerge of writing score gains was as follows:

- Writing score gains occurred across course types.
- Analysis of variance indicated there were no significant ($F = 1.331, p > .01$) differences in writing score gains between course types.
- Only one in ten learners made gains of more than one band, while one third scored lower on the second test. Only exceptionally did learners improve their writing test scores by a full band.

Score gains varied with initial writing scores. Results reflected a degree of regression to the mean: the higher the initial writing score, the lower the gain. The limitations in gains are suggestive of a ceiling effect in the test, but as the mean scores at exit were below the worldwide mean performance on IELTS writing, this seems unlikely. However, the finding is consistent with the frequently observed ‘plateau effect’ in language learning. At lower levels of ability relatively short periods of instruction can result in measurable improvements in proficiency, but at higher levels considerably longer periods are usually required.

**Prediction models**

While the analyses above are suggestive, they do not take account of the array of moderating variables that might have affected the average gain being made on each course type and, thus, neural network analyses were used to take account of these variables.

Data from the questionnaires and grammar and vocabulary tests provided over 100 separate features for each participant. Of these, a majority did not display a significant ($p < 0.05$) correlation with exit scores or with score gains and so would not be expected to contribute to predicting exit writing scores. To limit the number of features used as input to the neural network and so to improve the possibility of accurate prediction, these features were discarded.

The 34 features retained for the neural network analysis are displayed in Tables 4, 5 and 6. Correlations ($r$) with writing score gains are displayed for each feature,
IELTS language courses

89

together with a partial correlation, controlling for the effect of the initial (course entry) writing score. Where prediction is based on individual questionnaire items (rather than scales), as these are ranked data, non-parametric correlations (Spearman’s rho–\(r_s\)) are also provided.

It is interesting to note that the feature displaying the highest correlation with writing score gains was not course length (which was correlated at \(r=.269\)), but initial writing score (\(r=-.530\)). This indicates that students with higher initial writing scores made less gain than their lower scoring counterparts. Other features that displayed relatively high correlations with writing score gain included the grammar and vocabulary measures, use of test-taking strategies, self-assessed improvement in writing ability and self-confidence in English writing ability.

Entering these features to the neural network (Tables 7, 8 and 9), initial writing scores again contributed the most to predictions of writing score gain. This feature alone accounted for approximately 25% of the variance in the exit writing scores in the training set for both linear and non-linear models (the latter providing marginally
better accuracy of prediction). Course length, in contrast, contributed little to prediction, accounting for less than 8% of the variance in the exit writing scores.

Tables 7–9 show the performance of the three prediction models. Table 7 is based on thirty-four participant and process features, Table 8 on process variables alone and Table 9 on participant features alone. These results indicate that the participant features selected contribute more to prediction than do the process features.

### Table 5. Correlations between process features and score gains

<table>
<thead>
<tr>
<th>Course parameters</th>
<th>r</th>
<th>rₜ</th>
<th>partial r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Course length (weeks)</td>
<td>.269</td>
<td>.114</td>
<td></td>
</tr>
<tr>
<td>Questionnaire B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1: Course outcome</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO05 (I learned ways of improving my English Language test scores)</td>
<td>.162</td>
<td>.185</td>
<td>.128</td>
</tr>
<tr>
<td>CO06 (I learned words and phrases for describing graphs and diagrams)</td>
<td>.104</td>
<td>.108</td>
<td>.021</td>
</tr>
<tr>
<td>CO08 (I learned how to organize an essay to help the reader to understand)</td>
<td>.075</td>
<td>.118</td>
<td>.112</td>
</tr>
<tr>
<td>CO09 (I learned how to communicate my ideas effectively in writing)</td>
<td>.105</td>
<td>.107</td>
<td>.032</td>
</tr>
<tr>
<td>CO18 (My teacher corrected my grammar mistakes in my written work)</td>
<td>.093</td>
<td>.083</td>
<td>.022</td>
</tr>
<tr>
<td>CO19 (The activities we did in class were similar to the ones on the IELTS test)</td>
<td>.174</td>
<td>.201</td>
<td>-.050*</td>
</tr>
<tr>
<td>Section 2: Learning Strategies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learning strategy use</td>
<td>.113</td>
<td>.085</td>
<td></td>
</tr>
<tr>
<td>Section 3: Self-assessed gain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-assessed aptitude</td>
<td>.128</td>
<td>.175</td>
<td></td>
</tr>
<tr>
<td>Self-assessed writing score gains</td>
<td>.179</td>
<td>.200</td>
<td></td>
</tr>
<tr>
<td>Satisfaction with study context</td>
<td>.115</td>
<td>.084</td>
<td></td>
</tr>
<tr>
<td>Section 5: Approaches to learning</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Meaning-based approach scale</td>
<td>.083</td>
<td>.058</td>
<td></td>
</tr>
<tr>
<td>Test Knowledge Form</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test knowledge gain</td>
<td>.057</td>
<td>.010</td>
<td></td>
</tr>
<tr>
<td>Change in test task demands</td>
<td>.155</td>
<td>.150</td>
<td></td>
</tr>
<tr>
<td>Test strategy use</td>
<td>.168</td>
<td>.271</td>
<td></td>
</tr>
</tbody>
</table>

*Negative correlations indicate that higher scores on the variable are associated with lower score gains

<table>
<thead>
<tr>
<th>Exit Tests</th>
<th>r</th>
<th>partial r</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Exit Writing score)</td>
<td>.407</td>
<td></td>
</tr>
<tr>
<td>Exit Grammar score</td>
<td>-.167*</td>
<td>.132</td>
</tr>
<tr>
<td>Exit Vocabulary score</td>
<td>-.096*</td>
<td>.185</td>
</tr>
</tbody>
</table>

*Negative correlations indicate that higher scores on the variable are associated with lower score gains
Participant features treated separately (Table 9) account for around 45% of the variance in exit writing scores in the training set ($r = 0.675$) while process features (Table 8) account for 31% ($r = 0.556$). The inclusion of process features does add to the predictive power of the model, but also makes generalization to the cross-validation (c.v.) and testing sets less stable (which could be attributed to the inclusion of too many input features in relation to the number of cases). The full set of 34 features (Table 7) accounts for 52% of the variance in exit writing scores in the training set ($r = 0.72$), but this falls to just 41% in the testing set (0.636).

By pruning the networks—removing features from the model—it is possible to establish which features contribute the most to prediction (Table 10). This process suggested that learners who had low initial writing and grammar scores, studied on longer courses, were educated beyond secondary level and believed that they were good at learning to write in English, would achieve the highest writing score gains. The model also suggested that there were advantages in having a positive orienta-

<table>
<thead>
<tr>
<th>Set</th>
<th>r</th>
<th>mse</th>
<th>Error %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average training</td>
<td>0.720</td>
<td>0.050</td>
<td>6.127</td>
</tr>
<tr>
<td>Average c.v.</td>
<td>0.597</td>
<td>0.080</td>
<td>7.781</td>
</tr>
<tr>
<td>Testing</td>
<td>0.636</td>
<td>0.062</td>
<td>6.631</td>
</tr>
<tr>
<td>Linear training</td>
<td>0.710</td>
<td>0.052</td>
<td>8.009</td>
</tr>
<tr>
<td>Linear c.v.</td>
<td>0.561</td>
<td>0.085</td>
<td>6.109</td>
</tr>
<tr>
<td>Linear testing</td>
<td>0.642</td>
<td>0.061</td>
<td>6.878</td>
</tr>
</tbody>
</table>

Table 8. Performance measures for prediction based on 17 process features (desired feature = exit writing score)

<table>
<thead>
<tr>
<th>Set</th>
<th>r</th>
<th>mse</th>
<th>Error %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average training</td>
<td>0.556</td>
<td>0.072</td>
<td>7.262</td>
</tr>
<tr>
<td>Average c.v.</td>
<td>0.413</td>
<td>0.102</td>
<td>8.654</td>
</tr>
<tr>
<td>Testing</td>
<td>0.399</td>
<td>0.089</td>
<td>8.062</td>
</tr>
<tr>
<td>Linear training</td>
<td>0.529</td>
<td>0.075</td>
<td>7.343</td>
</tr>
<tr>
<td>Linear c.v.</td>
<td>0.396</td>
<td>0.104</td>
<td>8.715</td>
</tr>
<tr>
<td>Linear testing</td>
<td>0.436</td>
<td>0.084</td>
<td>7.943</td>
</tr>
</tbody>
</table>
tion towards the host culture and, of interest in relation to washback, in developing strategies to improve writing test scores. However, these features added relatively little to the level of prediction achieved with just three test features: initial writing score, initial Grammar and Vocabulary test scores. These three features alone accounted for 38% of the variance in the exit writing scores ($r = 0.646$) in the training set (Table 11) and prediction generalized well to the testing set ($r = 0.599$).

Adding Course Type as an input feature to the above models did not contribute to any improvement in the prediction of score gains. From this it can be concluded that learners pursuing a Type 1 or Type 3 test preparation course did not obtain any measurable advantage over those studying on Type 2 pre-sessional EAP courses in terms of test score gains. However, learners intending to take the test, whether on Type 2 pre-sessional EAP or Type 1 or Type 3 IELTS courses, did make significantly ($F = 56.9$, $p < .01$) greater gains than those with no plan to take the test again (although the effect was perhaps too small to be considered meaningful). In this

Table 9. Performance measures for prediction based on 17 participant features (desired feature = exit writing score)

<table>
<thead>
<tr>
<th>17 participant features</th>
<th>1 hidden layer with 4 PEs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r</td>
</tr>
<tr>
<td>Average training</td>
<td>0.675</td>
</tr>
<tr>
<td>Average c.v.</td>
<td>0.621</td>
</tr>
<tr>
<td>Testing</td>
<td>0.619</td>
</tr>
<tr>
<td>Linear training</td>
<td>0.658</td>
</tr>
<tr>
<td>Linear c.v.</td>
<td>0.614</td>
</tr>
<tr>
<td>Linear test</td>
<td>0.618</td>
</tr>
</tbody>
</table>

Table 10. Performance measures for prediction based on 9 selected features (desired feature = exit writing score)

<table>
<thead>
<tr>
<th>9 presage features</th>
<th>1 hidden layer with 3 PEs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r</td>
</tr>
<tr>
<td>Average training</td>
<td>0.646</td>
</tr>
<tr>
<td>Average c.v.</td>
<td>0.683</td>
</tr>
<tr>
<td>Testing</td>
<td>0.599</td>
</tr>
<tr>
<td>Linear training</td>
<td>0.682</td>
</tr>
<tr>
<td>Linear c.v.</td>
<td>0.594</td>
</tr>
<tr>
<td>Linear test</td>
<td>0.623</td>
</tr>
</tbody>
</table>
context, it seems to be washback to the learner (possibly in the form of motivation to succeed) rather than washback to the programme, which has the greater relevance to test score gains. However, the very small differences found cannot be taken to indicate that test preparation strategies make a substantial contribution to score outcomes.

Conclusions

In the context of IELTS preparation, this study has cast doubt on the power of dedicated test preparation courses to deliver the anticipated yields. Learners on courses that included IELTS preparation (Type 1 and Type 3) did not improve their scores to any greater extent than those on Type 2 pre-sessional EAP courses (with or without an IELTS preparation component).

Evidence from Weir and Green (2002) indicated that learners did not always share teacher understanding of what the test required of them or of what happened in classes. The prediction model would appear to suggest that in this context the response of the individual learner to the demands of the test might influence outcomes to a greater extent than their choice of course and the content of their classes. This is a question that might repay further investigation focused on individual learners, their learning goals and understanding of test demands.

It is encouraging for the validity of the IELTS tests both that scores were found to improve following instruction in academic writing and that there was no significant difference in terms of score gains between those studying on pre-sessional EAP programmes and those engaging in dedicated IELTS preparation courses. There is no evidence here that course providers were able, through dedicated test preparation practices, to exploit test design characteristics to boost scores. However, the complementary questions of whether learners on the pre-sessional EAP courses that were not directed towards the test make greater improvements in those academic writing skills not covered by the test, and the contribution of instruction in those skills to academic success also require investigation.

Table 11. Three test features: Initial Writing, Vocabulary and Grammar tests

<table>
<thead>
<tr>
<th>set</th>
<th>r</th>
<th>mse</th>
<th>Error %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average training</td>
<td>0.624</td>
<td>0.063</td>
<td>6.475</td>
</tr>
<tr>
<td>Average c.v.</td>
<td>0.592</td>
<td>0.068</td>
<td>7.205</td>
</tr>
<tr>
<td>Testing</td>
<td>0.619</td>
<td>0.076</td>
<td>7.551</td>
</tr>
<tr>
<td>Linear training</td>
<td>0.625</td>
<td>0.064</td>
<td>6.481</td>
</tr>
<tr>
<td>Linear c.v.</td>
<td>0.592</td>
<td>0.068</td>
<td>7.201</td>
</tr>
<tr>
<td>Linear test</td>
<td>0.615</td>
<td>0.077</td>
<td>7.581</td>
</tr>
</tbody>
</table>
Further research is also needed into IELTS preparation programmes in countries other than the UK and into the impact of the Reading, Listening and Speaking subtests to explore whether test preparation is more effective in these contexts. In the context addressed by this study at least, there is little apparent benefit to learners in concentrating on the test at the expense of broader academic writing skills—a finding that both teachers and learners might do well to note.

More generally, this study suggests that relatively narrow test preparation may not necessarily be more effective than wider-ranging alternatives in delivering improvements in test scores. If it is more generally found to be the case that ‘teaching to the test’ is no more effective in boosting test scores than teaching the targeted skills (at least for a test that covers these skills to the same extent as IELTS writing), this will have profound implications for the relationship between teaching and testing.

Acknowledgements

This study was partly funded by the British Council through the IELTS joint funded research programme. Cambridge ESOL Examinations contributed the two versions of the IELTS writing test.

Notes on contributor

Tony Green is principal lecturer in assessment at the Centre for Research in English Language Learning and Assessment, University of Bedfordshire. Until recently he worked at Cambridge ESOL examinations as part of a team involved in the validation of the IELTS test. His research interests include the testing of academic writing skills and interfaces between testing and curriculum.

References


Herbert, A. (Ed.) (1889) *The sacrifice of education to examination: letters from all sorts and conditions of men* (London, Williams & Norgate).


A. Green


Appendix. Example IELTS writing test (from *The IELTS Handbook, 2005*)

**WRITING TASK 1**

You should spend about 20 minutes on this task.

_The chart below shows the different modes of transportation used to travel to and from work in one European city, in 1950, 1970 and 1990. Write a report for a university lecturer describing the information below._

Write at least 150 words.

![Chart showing modes of transportation](image)

**WRITING TASK 2**

You should spend about 40 minutes on this task.

Present a written argument or case to an educated reader with no specialist knowledge of the following topic.

As computers are being used more and more in education, there will soon be no role for the teacher in the classroom.

To what extent do you agree or disagree?

You should use your own ideas, knowledge and experience and support your arguments with examples and relevant evidence.

Write at least 250 words.