Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China

Qi Luxia

a Guangdong University of Foreign Studies, China

Published online: 24 Apr 2007.

To cite this article: Qi Luxia (2007) Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China, Assessment in Education: Principles, Policy & Practice, 14:1, 51-74, DOI: 10.1080/09695940701272856

To link to this article: http://dx.doi.org/10.1080/09695940701272856

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
Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China

Qi Luxia*
Guangdong University of Foreign Studies, China

High-stakes tests have been employed widely to engineer curriculum innovation, or achieve intended washback in education. But our understanding of the role of high-stakes tests as an agent for change is limited due to the small number of empirical studies available on this issue. This paper reports on a washback study which focuses on the writing task in the National Matriculation English Test (NMET) in China and describes both its actual and its anticipated effects on secondary school teaching. Participants in the study consisted of an extensive sample of the NMET stake-holders, comprising test constructors, teachers, and students. Data collection instruments included interviews, classroom observations, and questionnaires. The results show that although writing was practised in schools, it was not taught in the way intended by the test constructors. The communicative features of the NMET writing task were not observed in the school practice that prepared students for this task. Both teachers and learners neglected the communicative context of writing while emphasizing the testing situation and the assumed preferences of the markers. This urge to raise scores in a real test situation suggests that high-stakes tests are not an efficient agent for change despite their crucial role in education.

Introduction

In the field of education there is general consensus that the testing process has an influence on teaching and learning. This influence is termed ‘washback’ in language education (Alderson & Wall, 1993) and is regarded as neutral with the potentiality of becoming either positive or negative (Hamp-Lyons, 1997, p. 295). When the influence is manipulated to control a specific curriculum or to induce desirable pedagogical changes, it is regarded as the intended washback of testing (Wall & Alderson, 1993). In this paper, ‘washback’ is used as a neutral term and ‘intended washback’
refers to the teaching-learning approach which the test constructors and policy-makers try to introduce into classroom teaching through testing, particularly through high-stakes public tests.

High-stakes tests, namely, tests which are believed to bring about important consequences to test-takers and other people concerned (Madaus, 1988) have been employed to engineer curriculum innovation, or achieve intended washback, in both general education and language education (e.g., Pearson, 1988; Li, 1990; Shohamy, 1993; Wall & Alderson, 1993; Shohamy et al., 1996; Cheng, 1997; James, 2000; Linn, 2000; Andrews & Wong, 2002; Stecher et al., 2004). Underlying this power-coercive, top-down approach to educational reform is the assumption that high-stakes tests possess the power to exert an expected influence on teaching and learning because of the consequences they bring about.

Unfortunately, however, the results of the present study do not support this assumption. It has been found that the intended practice, namely, writing with a communicative purpose, has not occurred in schools in spite of the efforts of the designers of a high-stakes language test (see results and discussion below).

Varied views of the intended washback of tests

The widespread practice of manipulating testing for the sake of curriculum innovation has aroused controversy. In general education, scholars advocating measurement-driven instruction (MDI) argue that it is a quick and most cost-effective way to improve education (e.g., Popham, 1987). Those against MDI criticize it for creating adverse effects on the curriculum in the long run (Bracey, 1987; Madaus, 1988).

Scholars in language education tend to adopt an empirical approach to the issue by conducting research studies to explore the complex nature of washback (Cheng, 1998; Hughes, 1988; Li, 1990; Shohamy, 1993; Wall & Alderson, 1993). Their findings have been mixed and far from conclusive. Hughes (1988) reported that a new English test had achieved its intended washback effects in that students’ English proficiency improved greatly a year after the test was implemented. Wall and Alderson (1993) found the new Sri Lankan English test employed to reinforce a set of new textbooks only managed to produce some changes, both intended and unintended, in teaching content, but not in the methods of teaching. Cheng (1998) investigated the washback of the new Hong Kong Certificate of Education Examination in English (HKCEE) and came to the conclusion that the test had strong effects on textbook publishers and materials writers as well as on teachers’ and students’ perceptions of classroom activities, but weak effects on teaching methods. Other studies have also found that language tests affected teaching content, but few changes were observed in teaching methods (Li, 1990; Shohamy 1993; Andrews, 1995).

The need for further research

Given the current status of research in this area, further studies on the intended washback of high-stakes tests are highly desirable. Thus, a research study on the intended
Examining intended washback

53

washback effects of the writing task in the National Matriculation English Test in China (NMET) was carried out to shed new light on the issue. The study was so designed that it included not only investigation into the pattern of teaching and learning under the pressure of the high-stakes test, but also an in-depth examination of the test constructors’ intentions and measures employed to realize their intentions. This design was based on the consideration that insufficient investigation has been made concerning the intentions of test designers and policy-makers in washback studies. Wall and Alderson (1993) in their study in Sri Lanka listed the expected washback effects according to their understanding of the official statements of the goals for the test. However, they did not document in detail how the test was designed to facilitate realization of the goals. Andrews (1995) found that test developers’ views on expected forms of washback were diverse. Interesting as this finding is, he did not attempt to explain why this was the case. Cheng (1998) found that ‘the HKEA failed to specify areas in teaching and learning on which this new examination may have an impact, nor how such an intended washback effect could be brought about in teaching’ (p. 311). But like the other researchers, she did not carry out in-depth studies of the test constructors’ intentions.

Our understanding of the specific expectations of the intended washback of high-stakes tests by the constructors and what they do to facilitate the occurrence of the intended washback tends to be hazy and superficial. Shohamy (2001) proposed that the intentions of decision-makers should be one focus of washback studies. In-depth examinations of test constructors’ intentions, in my view, are essential because the washback process involves both the ‘influencer’ (test constructors and their tests) and ‘the influenced’ (test users including teachers and learners). Neglect of either party will result in a partial picture of the issue being studied.

The writing task in the NMET appears to be an appropriate target for research on the intended washback of high-stakes tests. This is because it was considered by the test constructors to be a better manifestation of their intentions to encourage development of students’ language-use ability instead of mere linguistic knowledge since it was believed to be more direct and communicative than the other tasks or items in the same test, most of which adopt the multiple-choice format. Linguistic knowledge refers to knowledge of grammatical rules and the dictionary meaning of words, whereas language-use ability is believed to consist of linguistic knowledge, pragmatic and sociolinguistic knowledge, and the ability to use the knowledge for realistic communicative purposes (Li, 1988). This concept of language use is consistent with the language-ability model proposed by Bachman & Palmer (1996).

Context of the study

The role of testing in the Chinese educational system

The Chinese educational system is characterized as an examination-oriented system. Large-scale high-stakes tests serve as gatekeepers at every rung of the educational
ladder: primary, junior secondary, senior secondary, and tertiary. The two most important sets of examinations are the senior secondary school entrance examinations at the municipal level and the university entrance examinations at the national level. The latter tests nine subjects dictated by the national curricula. A student has to take tests in five or six subjects depending on the requirements of the type of university applied to. Chinese, mathematics, and English are three compulsory subjects for all candidates regardless of their choice of university. Only those candidates whose scores rank high in the results of these competitive norm-referenced tests have a chance of enrolling at a university.

Moreover, access to the next level of education is not the only goal of education at a lower level. In effect, enrolling at a key school or prestigious university is the dream of most Chinese youngsters and their parents. This dream can be fulfilled only through outstanding performance in the relevant examinations. Given this crucial role, examinations and tests are perceived to be powerful tools that can be manipulated to control and improve education (Spolsky, 1994; Guo, 1997).

The National Matriculation English Test

The National Matriculation English Test (NMET) in China is a standardized norm-referenced proficiency test. It is one of the three compulsory examinations for entrance into a Chinese university or college. It is developed by the National Education Examinations Authority directly under the Chinese Ministry of Education. The purpose of this test is to select the best secondary school graduates who demonstrate the highest level of English proficiency among their counterparts to compete for the limited spaces in Chinese universities and colleges. It is taken annually by millions of secondary-school graduates (Li, 1990; Anonymous, 2002).

The NMET was introduced in 1985 as a replacement for the old national university entrance English test which was criticized for its low validity and reliability (Li, 1990). The new test was first implemented in Guangdong province and was gradually adopted in other provinces throughout the country in subsequent years. Having been modified several times, the current version of the NMET tests listening, reading, writing, and knowledge of grammar and vocabulary (see Table 1).

Method

The research question

The study addresses the question ‘Does the NMET writing task affect teaching and learning in the way intended by the test constructors?’

To answer this, it is necessary to (i) understand the test constructors’ intentions and establish whether these are manifested in the test papers, (ii) investigate school practices to see whether writing is taught and how it is taught, and (iii) compare test constructors’ intentions and school practices to identify matches and mismatches.
Examining intended washback

between the two to determine if the intended washback has occurred, that is, if testing is a good agent for change.

Participants

Altogether 1382 participants took part in the study. They were 388 Senior III middle school English teachers, 986 Senior III students and eight NMET test constructors. The teachers and students were from approximately 130 schools in four districts of Guangdong province and 50 schools in one district of Sichuan province. The eight test constructors included two test designers, one subject officer, and five item writers. Item writers are considered to be test constructors because they not only write items but are also involved in redesigning the test when there is a need to make changes in its composition. There are altogether six item writers on the NMET production team. The author is one of the item writers and was thus both the researcher and a participant in the study.

Data collection

Three main instruments were employed to collect data: an interview, a questionnaire, and classroom observation. Interview data were obtained through in-depth semi-structured individual interviews with the 8 test constructors, 10 teachers, and 10 students; a further group interview with the test constructors; and follow-up contacts with most of the other interviewees. Altogether 29 interviews were conducted. The 28 individual interviews were audio-recorded, transcribed, and translated into English by the author. The group interview was audio-recorded but not transcribed because it was primarily intended to confirm the summarized results of the individual interviews with the test constructors and to clarify some points. Notes were kept of the follow-up contacts.

Based on findings from the interviews, two versions of a questionnaire (one for teachers and one for students) were constructed in order to cross-check the findings
and also to find out how far they could be applied to a larger group of participants (see Appendix 1). The teacher questionnaire was completed at four teachers’ meetings and the student questionnaire was administered in the students’ classrooms with the help of their teachers. After discarding those that were unusable (in which 50% of the items were unanswered), completed questionnaires from 378 teachers and 976 students were used for data analysis. The return rate for the teachers’ version was 81% and for the students’ version, 95%.

Four teachers’ classes were observed and audio-recorded. These teachers were among the 10 who acted as the interview informants. All 10 teachers were asked for permission to observe their classes. Seven agreed but three declined. Among the seven teachers, four taught writing in the classes observed. Since data from only four teachers’ classes were analyzed, some bias may have occurred. For example, other teachers might teach students to write with a communicative purpose and our conclusion about classroom practice based on the limited classroom observation data might not be valid. Thus, it is necessary to cross-check data using different instruments such as interviews and questionnaires.

NMET-related documents and teaching materials were collected for analysis. These included testing syllabi, journal articles by NMET designers, past test papers, marking schemes, textbooks, and commercial test preparation materials.

Data analysis

Analysis of the individual interview data was undertaken mainly through descriptive coding using WinMax (Kuckartz, 1998), a software package for qualitative data analysis. The data were initially coded with a ‘start list’ (Miles & Huberman, 1994, p. 58) with four categories: content, organization, accuracy, and mechanics, in line with conventional views about writing (e.g. Jacobs & Jacobs, 1973). This list grew with the adoption of more new codes and sub-codes during the process of data analysis.

NMET-related documents and teaching materials were scrutinized to confirm participants’ declared intentions and actions. A check-list was used to establish whether the specific features of the NMET writing task were manifested in past papers. Classroom observation data were analyzed through coding the recorded lessons using the observation scheme (Appendix 2) developed on the basis of the interview results. The questionnaire data were analyzed using SPSS 10 to extract frequencies and means as a cross-check on the validity of the findings from the interview data and classroom observation results.

Results and discussion

To answer the research question ‘Does the NMET writing task affect teaching and learning in the way intended by the test constructors?’, the results and discussion are organized thematically to address (i) the test constructors’ intentions and their realization in the test papers; (ii) writing practice in schools; (iii) comparison of views
on writing held by the test constructors, teachers, and students; and (iv) various reasons for the non-occurrence of intended washback.

**Test constructors’ intentions: write in a communicative way**

The interview data and the NMET-related documents demonstrate that the positive washback effect the test constructors anticipated in ELT at secondary schools in China was ‘a shift from formal linguistic knowledge to practice and use of the language’ (Li, 1990, p. 402). This intention arose from the test constructors’ belief that ELT in Chinese schools in the 1980s focused on linguistic knowledge and neglected language use (Gui et al., 1988; Li, 1988; Li et al., 1990), and students ‘never do any “practical” writing’ (Li, 1990, p. 399). This belief was shared by other scholars who described the Chinese approach to foreign language learning as concentration on intensive learning of grammatical structures, rote learning of vocabulary, and lack of attention to more communicative skills (Dzau, 1990; Harvey, 1990; Cortazzi & Jin, 1996; Yu, 2001).

Their beliefs, especially concerning the lack of writing practice, are supported by an examination of the Senior III English Textbook used at the time (Hu & Chen, 1982). There were no writing exercises in the textbook, the exercise closest to actually writing something in English being translating isolated Chinese sentences into English, and two retired English teachers confirmed that they did not teach writing in English before 1985 (Fang Xinying, personal communication, 15 February 2006; Peng Zhancai, personal communication, 18 February 2006). Thus, it is reasonable to believe that English writing was neglected in Chinese schools in the early 1980s.

**Embodiment of constructors’ intentions in the NMET writing task**

The NMET writing task provides a hypothetical situation which simulates a real-life language-use context for Chinese learners and asks the candidates to write a short text of about 100 words according to the content given in the input (see Appendix 3). This task reflects the principles in designing and producing the NMET, which include contextualizing the items in a communicative situation whenever plausible and making test items and tasks as close to real-life language use as possible (Li et al., 1990). Specifically, guidelines for setting the writing task include the following: (a) there should be a clear description of the purpose for writing, the intended reader, the role of the writer, and the relationship between the writer and the reader; and (b) the task should be one that the Chinese students might realistically be expected to meet in real life (Li et al., 1990; NMET Item Writers’ Guidebook, 1997).

The marking of the NMET writing scripts adopted an analytical method of scoring in the early years, which requires a separate score for content and language. The total score was 25, with 12 marks for content and 13 for language. The requirement for language concerns fluency, accuracy and appropriateness. In 1992 there was a
change to holistic marking, with a scoring system of seven levels. The change was made because the markers complained about the complexity of the analytical marking scheme and the NMET designers found that markers in some provinces did not in fact follow the official analytical marking scheme.

This change might have affected the intended washback. An examination of the holistic marking schemes from 1992 to 1998 revealed that the requirement for appropriateness was not included in them, although those used since 1999 have a statement added to the effect that the writing script should ‘achieve the desired effect on the target reader’ (for example, *NMET Writing Scheme*, 2003), which implies that the language of the script should be appropriate.

Nevertheless, the test constructors’ intention to make students practise writing in a more communicative way than previously is reflected in the input of the writing task, as is illustrated by an examination of the NMET past papers using a checklist (see Table 2).

It became clear that the guidelines for designing the writing task were followed in all years except in 1992 and 1998. In eight of the ten NMET papers, the context, consisting of writing purpose and the imaginary writer and audience, is explicitly described. It is reasonable to believe that the test constructors’ intentions concerning practice writing for communication is illustrated in the input of the writing task, though the requirement for appropriateness has not been consistent in the marking schemes as discussed above.

The underlying assumption of the test constructors was that if writing were tested in the NMET it would be taught in schools. Further, through testing writing in a more communicative way, teachers would be encouraged to teach writing communicatively and students would learn to write for communicative purposes.

The next question is, ‘Have the test constructors’ intentions been realized in school ELT?’

<table>
<thead>
<tr>
<th>NMET Paper</th>
<th>Purpose</th>
<th>Writer</th>
<th>Reader</th>
<th>Realistic situations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>1992</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>1993</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1994</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>1995</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>1996</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>1997</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>1998</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>2000</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

*Note: A tick indicates that the relevant feature is manifested in that year’s NMET test paper.*
Most teachers and students who completed the questionnaire chose ‘occasionally’, ‘sometimes’, or ‘often’ for the questions asking about the frequency of writing practice. Ninety-two percent of teachers indicated that writing was practised during the whole year of Senior III. The average number of passages written was 39 (Appendix 4). This result provides one more example of test influence on the content of teaching.

Writing practice, however, is a vague term in China’s ELT circle. It can mean constructing sentences, re-writing what one has read, or writing an essay. Even copying two simple sentences and joining them with an adverbial clause to make a complex sentence is treated as writing practice (Anonymous, 1986). Thus, it is important to be clear how, not just whether, writing was practised in the Senior III English course.

With respect to how writing was taught, it was found that although writing practice in schools imitated the NMET writing task, the similarity between the two is only superficial. The writing tasks used in schools are like the NMET writing task in that they also require students to write a short text of about 100 words based on the content provided, but unlike the NMET writing task there are no instructions which specify the hypothesized purpose of writing, the writer and the audience. Thus, it is risky to draw a conclusion that the NMET writing task has achieved the intended washback, despite the fact that teachers claimed in the interviews that writing was practised in the way it was tested in the NMET and 89.7% of the teacher respondents answered ‘yes’ to the question ‘Do most of your writing exercises resemble the NMET writing task?’ (see Appendix 1). Therefore, participants’ views on writing as expressed in the interviews should be considered, reflecting the notion that their way of thinking could affect how they test, teach, or learn.

Both matches and divergences have been detected between views on writing held by test constructors, teachers and students (see Table 3).

Matches were found in content, organization, and accuracy. Both parties thought that these were important components of a piece of good writing. This is not surprising because these aspects of writing have been accentuated traditionally (e.g., Lawrence, 1975; Ding & Wu, 1984; McKay, 1984; Ferris & Hedgcock, 1998). It would be far-fetched to give credit to the NMET writing task for drawing teachers’ and students’ attention to these areas. Nevertheless, there are minor differences even in these shared areas. For example, the teachers and students placed considerable emphasis on accurate use of punctuation and capital letters whereas none of the test constructors touched upon these points.
One mismatch occurs in the category of mechanics. Five teachers and three students emphasized the mechanics while the test constructors did not bring up these points. Teacher E\textsuperscript{11} said, ‘The appearance of the answer sheet is important... A student might lose some marks because his answer sheet doesn’t look good. That’ll be the marker’s first impression of the piece of writing...very important’.

Another mismatch between the views of the two parties is found in the communicative context of writing, which was supposed to determine the content of writing and appropriateness of language use. The test constructors attached much importance to the context whereas none of the teachers and students in the interviews said anything about it. To quote Constructor A, ‘Another point is appropriateness. Here, we tell the test-takers who the letter writer and receiver are and what their relationship is. If a teacher doesn’t train his or her students in this respect, they will have a hard time’.

The emphasis on writing in a communicative context reflects the test constructors’ beliefs in communicative language testing, teaching, and learning (Li, 1988). Three
test constructors mentioned in the interviews that the ability to use language effectively comprised not only linguistic knowledge but also pragmatic and sociolinguistic knowledge and the ability to use this knowledge in communication. This confirms the test designers’ intentions to introduce teaching for language use in schools as it was stated explicitly in the journal articles on the NMET (see Li, 1990; Li et al., 1990).

The concept of communicative context was rather new in China when the NMET was first introduced in 1985 (Li, 1987). It forms the basis of the writing task and is one of the most important intended influences in terms of washback effect. Why did the teachers and students make no mention of the communicative context of writing in the interviews? To clarify this issue, a questionnaire was used to cross-check the interview data, the results of which are discussed below.

Neglect of the communicative context of writing

The interview data were drawn on in the construction of items for the questionnaire. Each item was designed so that a 5-point Likert scale was juxtaposed beside a statement with five numbers denoting varying degrees of importance (see Appendix 1). A mean score above 3 suggests that the questionnaire respondents as a group thought the statement important and a score below 3 suggests they thought it unimportant.

The results of the questionnaire data, obtained from 976 students and 378 teachers, confirm what was found from analysis of the interview data, that is, the communicative context of writing was neglected by teachers and students (see Table 4).

Table 4 demonstrates that the ratings for all the 13 items are above 3, suggesting that all the aspects of writing described by the interviewees were believed to be important by this group of teachers and learners. However, self-report data such as

<table>
<thead>
<tr>
<th>Items</th>
<th>Cases</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Correct grammar and words</td>
<td>1352</td>
<td>4.55</td>
</tr>
<tr>
<td>8</td>
<td>Opening and ending</td>
<td>1347</td>
<td>4.42</td>
</tr>
<tr>
<td>12</td>
<td>Idiomatic English</td>
<td>1335</td>
<td>4.38</td>
</tr>
<tr>
<td>7</td>
<td>Neatness of paper</td>
<td>1351</td>
<td>4.34</td>
</tr>
<tr>
<td>4</td>
<td>Cohesion</td>
<td>1349</td>
<td>4.31</td>
</tr>
<tr>
<td>6</td>
<td>Appropriateness</td>
<td>1345</td>
<td>4.25</td>
</tr>
<tr>
<td>9</td>
<td>Including all key points</td>
<td>1335</td>
<td>4.04</td>
</tr>
<tr>
<td>13</td>
<td>Including necessary details</td>
<td>1327</td>
<td>4.00</td>
</tr>
<tr>
<td>5</td>
<td>Correct capitals and punctuation</td>
<td>1348</td>
<td>3.91</td>
</tr>
<tr>
<td>1</td>
<td>Keeping to word limit</td>
<td>1341</td>
<td>3.70</td>
</tr>
<tr>
<td>11</td>
<td>Purpose of writing</td>
<td>1306</td>
<td>3.57</td>
</tr>
<tr>
<td>2</td>
<td>Supposed writer</td>
<td>1309</td>
<td>3.39</td>
</tr>
<tr>
<td>10</td>
<td>Supposed audience</td>
<td>1291</td>
<td>3.07</td>
</tr>
</tbody>
</table>

*Note Column 1 contains short forms of the items in the questionnaire (see Appendix 1) and the means are arranged in descending order.*
questionnaire ratings should be interpreted with some caution. A close examination of the figures in Table 4 shows that the nine items relevant to the aspects of writing discussed by the teachers and students in the interviews have higher mean scores than the items (Items 11, 2 and 10) that represent the communicative features of writing emphasized by the test constructors. Item 6 (appropriateness), with a mean score of 4.25, is an exception. This was unexpected because the point was not made by any respondents in the interviews. A possible explanation can be found in a follow-up telephone contact. When asked about their understanding of this item, two teacher interviewees did not seem to have a clear idea about it. Instead, they referred the researcher to the NMET marking scheme, in which there was a requirement for appropriateness (in Chinese) without any explanation of what that meant and how it should be assessed. It seems that teachers and students, at least some of them, might have taken the face value of the Chinese word without understanding its connection with the communicative context denoted by the other three items (Items 11, 2, and 10). That could account for the unexpectedly high mean score for this item and comparatively low scores for the other three items in the same category.

The respondents' ratings on the three context items vary more than those on the other items as the standard deviation values of these three items are the highest (see Items 11, 2, 10 in Table 4). This is probably due to some respondents' uncertainty about the role of the communicative context. Evidence for this interpretation comes from the counting of the choice of the 'I don't know' option or item non-response (INR), a term often used in survey studies (see Table 5).

It can be seen that the items focusing on the communicative context (Items 2, 11, 10) receive more INR in comparison with the other items.

The fact that these three items representing the communicative context of writing have the lowest means, the highest standard deviations, and the largest percentages

<table>
<thead>
<tr>
<th>Items</th>
<th>'I Don’t Know'. Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Correct grammar and words</td>
<td>2</td>
<td>0.1</td>
</tr>
<tr>
<td>7. Neatness of paper</td>
<td>3</td>
<td>0.2</td>
</tr>
<tr>
<td>4. Cohesion</td>
<td>5</td>
<td>0.4</td>
</tr>
<tr>
<td>5. Correct capitals and punctuation</td>
<td>6</td>
<td>0.4</td>
</tr>
<tr>
<td>8. Opening and ending</td>
<td>7</td>
<td>0.5</td>
</tr>
<tr>
<td>6. Appropriateness</td>
<td>9</td>
<td>0.7</td>
</tr>
<tr>
<td>1. Keeping to word limit</td>
<td>13</td>
<td>1.0</td>
</tr>
<tr>
<td>9. Including all key points</td>
<td>19</td>
<td>1.4</td>
</tr>
<tr>
<td>12. Idiomatic English</td>
<td>19</td>
<td>1.4</td>
</tr>
<tr>
<td>13. Including necessary details</td>
<td>27</td>
<td>2.0</td>
</tr>
<tr>
<td>2. Supposed writer</td>
<td>45</td>
<td>3.3</td>
</tr>
<tr>
<td>11. Purpose of writing</td>
<td>48</td>
<td>3.5</td>
</tr>
<tr>
<td>10. Supposed audience</td>
<td>63</td>
<td>4.7</td>
</tr>
</tbody>
</table>

Note: The items are arranged in an ascending order according to the number of INR they received. Case = 1354
of INR implies that the teachers and students as a group did not attach sufficient importance to the communicative context of writing and that the group opinion on its significance varied more than it did concerning other aspects of writing.

However, both the interview and the questionnaire instruments yield only self-report data that might be of low validity (Allan, 1995). Moreover, from these data we can only learn about the informants’ reported views and their behaviour by inference but not direct observation. We do not know, for example, if they actually overlooked context and emphasized the other aspects when practising writing. Alderson and Wall (1993) pointed to the need to conduct classroom observations to see ‘whether what teachers and learners say they do is reflected in their behaviour’ (p. 127). For this, classroom observations were conducted to see how writing was practised in class.

**Focus of teaching demonstrated by overt attention to different aspects of writing**

Classroom observation results indicate that in writing practice the communicative feature of the NMET writing task was largely missing, whereas linguistic accuracy was highlighted. This is evident from an analysis of the data using Part B of the classroom observation scheme (Appendix 2) which reveals teachers’ and students’ overt attention to different aspects of writing, as indicated by what students and teachers say or talk about. For example, when copying a piece of writing on the blackboard as a model for the students, a teacher asked the students to draw some lines on a white sheet before they started to copy the model so that their copy would look neat. This was recorded under ‘neatness of paper’ in the observation scheme. However, when the students were copying the model silently, it was impossible to find out what they paid attention to and therefore nothing was recorded.

It is clear from the results of the classroom observation relating to Part B of the observation scheme (see Table 6) that all the five aspects of writing gained attention from teachers and students, with accuracy receiving the largest proportion (59.6 %) and the communicative context the smallest (6.8%).

In fact, only one of the four teachers paid some explicit attention to the context of writing when commenting on students’ work done in a mock test. A close examination, however, shows that the teacher’s attempt to teach her students how to write appropriately in relation to the context was not particularly successful, as her suggestions did not seem to be appropriate at all. The instructions of the relevant writing task are presented below.

<table>
<thead>
<tr>
<th>Table 6. Overt attention to different aspects of writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Writing aspects</td>
</tr>
<tr>
<td>Overt attention</td>
</tr>
<tr>
<td>Counts</td>
</tr>
<tr>
<td>Percentage</td>
</tr>
</tbody>
</table>
Instructions: Suppose you came to Shanghai from Beijing and the suitcase you had checked in was missing. Write a letter of about 100 words to the Shanghai railway station to ask about it. (Mimeographed mock test paper)

According to Teacher I, the students who did the above task in the mock test did not write in appropriate English. So in class she commented on the students’ writing and suggested some sentences that she thought were appropriate. To quote Teacher I:

Some of you didn’t pay any attention to the tone. It didn’t sound polite. You write to ask people to help you… You can begin with ‘I’m sorry to trouble you… At the end, you should at least say ‘Thank you’. It’s better if you write ‘I would be very much obliged if you could kindly help me to find it’… The teacher who rates your writing script will be very pleased to see this sentence, thinking ‘This student is really good. He or she is so polite and can express the idea so naturally’.

The sentences proposed by Teacher I to start and end the letter sound unnatural and inappropriate. The purpose of the letter in such a situation is to make the railway people do their job, that is, to find the suitcase that got lost. By logic, this letter has an element of complaint although the writer may sound polite. The suggested sentences would actually sound sarcastic rather than polite to native speakers of English (Alastair Allan, personal communication, March, 2001). Thus, what was coded as occasions for attention to appropriateness (Table 6) and communicative context ended up in an attempt to raise the students’ awareness of it, not a successful case of teaching students how to write appropriately in accordance with the communicative context. Furthermore, since politeness, emphasized by Teacher I, was not really addressed to the supposed audience (the railway station staff) but to the rater marker who was assumed to prefer polite English, the true motive for advocating politeness as appropriateness in the letter was to please the rater for higher test scores.

But why is it that a case of teaching appropriateness turns out to be unsuccessful? It seems that the teacher did not know what was appropriate in the context provided by the mock task. This is not surprising at all, given that she taught English in an English-as-a-foreign-language situation. The interview data revealed that the teacher had never been in an English-speaking country and had never had an opportunity to use the language for authentic communication herself. Such being the case, it is very difficult to make teachers use tasks that resemble real-life communication despite the effort of the NMET constructors to build communicative features into the task.

The above case is not unique. In fact, language teaching professionals in other English-as-a-foreign language situations often express concern about the difficulty and appropriateness of teaching English communicatively (Medgyes 1994).

To sum up, three of the four teachers observed did not teach anything approximating the communicative context or appropriateness. Only one teacher was seen to try and teach appropriateness, but with little success. In class, teachers emphasized linguistic accuracy in writing (see Table 6) while neglecting communicative features which were stressed by the test constructors.
Reasons for not practising in the way intended by the test constructors

It appears that it is because of the pressure involved in preparing for and taking a high-stakes test that the communicative context was neglected during writing practice. This is illustrated by the fact that teachers and students focused only on the aspects of writing that they believed would help to achieve higher test scores, while completely ignoring the need to be able to write communicatively in real-life situations.

The motive of gaining higher scores accounts for teachers’ and students’ interest in the mechanics of writing (see Table 3) and Teacher I’s emphasis on ‘sounding polite’ in the writing scripts. Both mechanics and ‘politeness’ were assumed to affect the raters’ impression of a piece of writing and in turn the score awarded.

Teaching to raise scores is consistent with the widespread instrumental motivation for English learning in China. English is perceived by many Chinese learners as the fabled Aladdin’s lamp that provides them with much needed social and economic mobility. ‘Many secondary students work hard on English, not to communicate with foreigners but to go to college’ (Zhao & Campbell, 1995, p. 384; Qi, 2003). High-stakes tests have done much to reinforce such motivation for learning English. Therefore, it is not surprising that they have succeeded in focusing school ELT on what is tested and on the aspects that are assumed to bring about higher scores in the test. Test constructors’ intended innovations will not occur if they are not closely connected with raising scores in the coming test.

Furthermore, it is doubtful that writing would be taught effectively with an emphasis on communicative context even if appropriateness were given sufficient weighting in the marking scheme. As can be seen from Teacher I’s attempt at teaching appropriateness within the communicative context discussed above, the teacher might not possess the ability to write in English for authentic communication herself. In fact, teachers’ educational background and teaching beliefs have been found to play an important role in what happens in the classroom even when teaching is planned specifically to prepare students for examinations and tests (Wall & Alderson, 1993; Alderson & Hamp-Lyons, 1996). In certain contexts, teacher factors have been found to outweigh the high-stakes test in shaping classroom teaching and learning (Watanabe, 1996). Alderson and Hamp-Lyons (1996, p. 296) point out that ‘the amount and type of washback will vary according to the extent to which teachers and textbook writers are willing and able to innovate…’

Conclusion

Pulling all the lines of evidence together, we can answer the research question ‘Does the NMET writing task affect teaching and learning in the way intended by the test constructors?’ The answer is Yes and No. Yes, in terms of the amount of time and attention given to writing practice: the NMET had an impact on what was taught or teaching content. Test influence on teaching content is well documented in empirical studies on washback (e.g., Shohamy, 1993; Wall & Alderson, 1993; Andrews, 1995; Cheng, 1998) to which this study contributes further evidence.
With respect to whether writing is practised in the way intended by the test constructors, the answer is mostly No. In spite of the superficial resemblance to the NMET writing task, writing practice in Senior III was not done for communicative purposes in a simulated real-life context as advocated by the test constructors. Rather, the aspect of writing most emphasized was grammatical accuracy, which indicates that the traditional approach to teaching still prevails. The NMET writing task did not succeed in inducing the intended innovations in the teaching of writing. Therefore, high-stakes tests alone do not seem to be efficient agents for the type of changes expected by the test constructors and policy-makers. Other measures should also be sought in order to make the intended washback of the NMET writing task occur.

First, teachers should be given some, or more, opportunities to use English in authentic contexts themselves to improve their own language-use ability and awareness of what constitutes authenticity and real-life situations, for example by being sent to study in English-speaking countries as part of their in-service training. Second, training and support for teachers should be provided to help them understand better the intentions of the NMET constructors and find more effective and appropriate ways to prepare their students for the test. Third, linked to training and support, better communication between the test constructors and teachers would be beneficial. This could be accomplished through the use of electronic resources, such as those employed by the developers of IELTS (http://www.ielts.org/) and TOEFL (http://www.ets.org/toefl/), as well as face-to-face interaction. Workshops between test constructors and teachers could provide opportunities for test constructors to inform teachers of their intentions in relation to washback; equally, teachers might be invited to inform the test constructors about their teaching practices and to air their opinions about how the NMET should be reformed. With greater support for teachers and improved communication between teachers and test constructors, the intended washback of the NMET might be realized.

Acknowledgements

The study reported in this article has been supported by a Ministry of Education (MOE) project of the Centre for Linguistics and Applied Linguistics of Guangdong University of Foreign Studies. I am grateful to Dr. Alastair Allan, Dr. Carol MacLennan, and the anonymous reviewers for their constructive comments on the earlier drafts of the article.

Notes

1. This is part of a larger washback project which investigated the intended washback effect of the NMET as a whole.
2. HKEA stands for Hong Kong Examination Authorities, the institution that produces the test under study.
3. In China, schools at all levels, as well as universities, fall into two categories: key schools and ordinary schools. Key schools enjoy a number of privileges in terms of a larger share of the educational budget, better-qualified teachers, and permission to enroll students before ordinary schools (Liu, 1993; Zhou & Campbell, 1995).

4. Recently, a top-down reform was proposed by the Ministry of Education. Starting from February 2004, the NMET test paper should be developed both by the National Education Examinations Authority to be used by the majority of provinces and locally in some provinces and municipalities for local use, but the different versions of the NMET should follow the same testing syllabus and take the same test formats (Muo, 2004).

5. Although the NMET has undergone some changes, the writing task has remained the same since 1985. In this study, therefore, it was assumed that the influence on English writing practice in schools has remained unchanged.

6. Senior III teachers, not teachers from lower grades, were selected for the study on the consideration that, since they have to help their students prepare for the NMET to be taken at the end of the school year, they would be more concerned with and have a better knowledge of the philosophy of the test than teachers teaching the other grades.

7. These authors are the NMET designers.

8. Appropriateness refers to the suitability of the language for the particular situation and purpose. For example, addressing strangers in a very formal situation by their given names is considered inappropriate.

9. NMET Writing Schemes are unpublished documents meant for the raters. They are usually released after the marking has been finished each year with the used test paper and teachers would study them carefully in order to be able to help the next cohort of students prepare for the test.

10. Students were not asked to answer the same question because when the data were collected they were not in a position to do so, being only halfway through that school year.

11. The participants are identified by their profession and a letter to preserve anonymity.

12. Part A of the scheme is an open category which records the activities carried out in a class.

13. The instructions are given in Chinese. The English version is a translation by the present author.

Notes on contributor

Qi Luxia is associate professor of English, Faculty of English Language and Culture, Guangdong University of Foreign Studies, China. Her research interests are language testing and assessment, the impact of testing on teaching and learning, and the teaching of reading as a foreign language.

References


Appendix 1. Questionnaire (Teacher version)\textsuperscript{a}

Dear colleagues:

We are engaged in a study on English teaching and learning in Senior III. Could you please help us by completing this questionnaire? All information will be treated in the strictest confidence and please do NOT write down your name on the questionnaire. Thank you very much!

Part Two. This section investigates the general practices in the Senior III English course. There are no ‘right’ or ‘wrong’ answers. Please tell us what you actually do but not what you think should be done. Please circle your choice or fill in the blanks as required.

2. In the school year of Senior III most of the time students:
   • practise writing

6. In Senior III how many short passages do your students write?
   (1) 0–19 (2) 20–39 (3) 40–59 (4) 60–79 (5) 80–99 (6) 100 or more

7. How often do you do the following in class?
   • Make students practise writing
     Never Occasionally Sometimes Often Always
     1 2 3 4 5

8. How often do you make your students do the following after class as homework?
   • Writing
     Never Occasionally Sometimes Often Always
     1 2 3 4 5

9. How do your students practise the following?
   • Do most of your writing exercises resemble the NMET writing task?
     Yes No
     1 2

11. How important do you think the following suggestions are concerning writing practice?
   1 = not important at all  2 = not important  3 = neutral
   4 = important  5 = very important  ? = don’t know
Examining intended washback

- Write about 100 words, no more no less
- Think about the supposed writer before starting the task
- Write accurately. Try not to make grammatical and vocabulary errors
- Use link words to make the writing cohesive
- Get capitals and punctuation right
- Write in appropriate language
- Make handwriting look nice and keep answer sheets clean and tidy
- Make the writing complete with an opening and an ending
- Include all the necessary points. Do not add or leave out any
- Think about the supposed audience before starting the task
- Think about the supposed purpose of writing before starting the task
- Use idiomatic English. Try not to write in Chinglish
- Add necessary details to make the content complete

Please check to see whether you have answered all the questions.
Thank you very much for your help!

^aTo save space only the relevant items in the original questionnaire are given here
## Appendix 2. Observation scheme for writing lessons

### Part A and B

<table>
<thead>
<tr>
<th>Activity</th>
<th>Time Points</th>
<th>Details</th>
<th>Opening &amp; ending</th>
<th>Cohesion</th>
<th>Grammars</th>
<th>Vocabulary</th>
<th>Punctuation</th>
<th>Capitals</th>
<th>No Chinglish</th>
<th>Neatness of paper</th>
<th>Word limit</th>
<th>Handwriting</th>
<th>Pur - pose</th>
<th>Writer</th>
<th>Audience</th>
<th>Appropriateness</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of counts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix 3.  Writing task (NMET, 2003)*

*The English translation of the instructions is as follows:

Suppose you are Li Hua. Your British pen-friend Bob will come to your city to take a Chinese course in the Jianxin Chinese School. You have helped him find a small flat to stay in, which is 25 square meters and the rent is RMB 500 a month. Use information given in the following drawings to write him a letter to tell him about the flat.
## Appendix 4. Frequency of writing practice in Senior III

<table>
<thead>
<tr>
<th>Items</th>
<th>Cases</th>
<th>Response frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. How often is writing practised in class?</td>
<td>1351</td>
<td>10.6 30.3 41.4 16.8 0.7</td>
</tr>
<tr>
<td>2. How often is writing practised after class?</td>
<td>1348</td>
<td>7.4 27.9 43.1 19.4 1.9</td>
</tr>
<tr>
<td>3. Is writing practised during the whole year of Senior III?</td>
<td>378</td>
<td>Yes 92.6 No 7.4</td>
</tr>
<tr>
<td>4. In Senior III how many short passages do your students write?</td>
<td></td>
<td>Average 39 Range 15–105</td>
</tr>
</tbody>
</table>

*T stands for teachers and S for students.*