The impact of high-stakes testing on teaching and learning: can this be predicted or controlled?

Dianne Wall *

Institute for English Language Education, Lancaster University, Lancaster LA1 4YJ, UK

Received 2 February 2000; received in revised form 28 March 2000; accepted 4 April 2000

Abstract

One of the issues which attracted the attention of language testers in the 1990s was the impact of high-stakes tests on the classroom: what kind of influence did such tests have on teaching and learning and what could educators do to ensure that this was beneficial rather than harmful? Some progress was made in defining notions such as ‘impact’ and ‘washback’, and a number of studies appeared which analysed the relationship between tests and teachers’ and learners’ attitudes and behaviour. There was a growing awareness of the importance of factors other than test design in determining whether tests would have the impact that was desired. These factors also appear in the literature of educational innovation, and it is to this field that some testers turned for guidance on whether test impact could be predicted or controlled. This paper summarises what language testers have learned about test impact in the last decade and discusses what one model of educational innovation has revealed about how tests interact with other factors in the testing situation. It concludes with a set of recommendations about the steps future test developers might take in order to assess the amount of risk involved in attempting to create change through testing. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: High-stakes tests; Teaching; Learning; Impact

Tests serve a number of functions in society, ranging from ‘encouraging higher levels of competence and knowledge’ and ‘checking patronage and corruption’ to ‘allocating sparse places in higher education’ and ‘measuring and improving the effectiveness of teachers and schools’ (Eckstein and Noah, 1993). While the first two functions can be seen as beneficial to all, the last two involve loss as well as gain.
Tests are a ‘differentiating ritual’ for students: for every one who advances there will be some who stay behind (Eggleston, 1984, p. 20). They serve much the same function for teachers and schools since their reputations can be strengthened or weakened with the publication of test results. This may lead to anxiety or other negative effects in the classroom. Tests which are helpful to decision makers (admissions officers, educational administrators) are not necessarily helpful to teachers and students.

Although this is a relatively new topic in language education, it has long been discussed in general education. Wiseman (1961) argued that tests should be judged not only by their ‘technical efficiency’ but also by whether they were ‘educationally profitable’. Tests were ‘profitable’ if they produced more ‘credits’ than ‘debits’ in the classroom. Wiseman illustrated these concepts by referring to what he called “a forgotten classic”: the 1911 Report of the Consultative Committee on Examinations in Secondary Schools. This report listed the effects that tests might have on teachers and on pupils. The good effects on teachers included inducing them to cover their subjects thoroughly, forcing them to complete their syllabuses within the prescribed time limits, compelling them to pay as much attention to weak pupils as to strong ones, and making them familiar with the standards which other teachers and schools were able to achieve. The possible negative effects included encouraging teachers to ‘watch the examiner’s foibles and to note his idiosyncrasies’ in order to prepare pupils for questions that were likely to appear, limiting the teachers’ freedom to teach subjects in their own way, encouraging them to do the work that the pupils should be doing, tempting them to overvalue the type of skills that led to successful examination performance, and convincing them to pay attention to the ‘purely examinable side’ of their professional work and to neglect the side which would not be tested (Wiseman, pp. 159–161).

There have been many discussions of test impact since then, some focusing on specific tests in specific situations (Vernon, 1956; Kellaghan et al., 1982; Smith, 1991), others surveying the literature for evidence of general tendencies (Kirkland, 1971) or presenting issues of relevance to the wider educational community (Popham, 1987; Frederiksen and Collins, 1989; Airasian, 1988; Madaus, 1988; Haladyna et al., 1991; Herman and Golan, 1993). Popham was perhaps the best known advocate of using high-stakes tests to influence the curriculum (‘measurement-driven instruction’). He argued that if tests were “properly conceived and implemented”, then focusing teaching on what they assessed was a positive activity. Several conditions had to be met, however: criterion-referencing; defensible content; a manageable number of targets; ‘instructional illumination’; and adequate support for teachers, in the form of descriptions of the skills to be tested, sample test items, and advice on beneficial teaching activities.

Madaus (1988) was one of the fiercest critics of measurement-driven instruction, claiming that it was “nothing more than psychometric imperialism” and that it

...invariably leads to cramming; narrows the curriculum; concentrates attention on those skills most amenable to testing...constrains the creativity and spontaneity of teachers and students and finally demeans the professional judgement of teachers. (Madaus, p. 95)
Madaus reviewed a number of studies on test impact and presented a set of ‘principles’ to summarise his own position. These included the notions that teachers will always ‘teach to the test’ and that, ominously, the high-stakes test “transfers control over the curriculum to the agency which sets or controls the exam”. Madaus claimed that this was particularly serious in situations where those who developed and validated tests had commercial rather than educational interests at heart (Madaus, p. 98).

A separate literature was also growing on the effects of high-stakes testing in developing countries (Eisemon, 1990; Heyneman and Ransom, 1990; Kellaghan and Greaney, 1992). The functions of tests in these countries are often similar to those mentioned above, but the fact that there are far fewer places available in the upper levels of education makes the stakes much higher (Mathews, 1985, p. 23; Foster, 1992, p. 123). While these researchers recognised the negative effects that tests could have in these settings, they felt confident about giving guidelines for improving the situation, focusing on improving test construction and analysis, carrying out research into the quality of tests and their impact; and building closer co-operation between testing agencies and other important parties (e.g. curriculum designers) in the educational system. However, other researchers, such as Little (1992), criticise the notion that changing a test is a sufficient condition for changing the curriculum: just because tests which are out of line with the curriculum can constrain, it does not mean that tests that are in line with it will ensure its implementation.

There was little discussion of the impact of testing in language education before the 1990s. Much of what was published earlier was based on anecdote or assertion, or expressed intentions to create positive impact without providing evidence that it had occurred. The few empirical studies which existed relied on survey data or on test results rather than on direct contact with the classroom (e.g. Wesdorp, 1983; Hughes, 1988; Khaniya, 1990; Li, 1990). Shohamy (1993) used a more ambitious research design, which included a focus on three different tests and several different methods for gathering data. Her work suggested that test impact was more complicated than had previously been assumed, and that many factors needed to be taken into account before it was possible to say that a specific test would have a specific effect in a specific context.

Alderson and Wall (1993) took a critical look at the concept of test ‘washback’ (in simple terms, the effect of tests on teaching and learning — also known as ‘backwash’ in general education) as it had been presented in the language testing literature up to that time. They argued that those who wished to create positive washback needed to be more explicit when they described what they hoped to achieve and those who wished to investigate the phenomenon needed to use tighter definitions (the ‘Washback Hypotheses’), more varied research techniques (especially observation), and insights from other fields such as educational innovation. Hughes (1994) added to the theoretical discussion by making a distinction between washback on the participants, the processes and the products of an educational system. Among the participants who could be affected by tests were teachers, learners, educational administrators, materials writers and publishers. Tests could affect their attitudes and the activities they engaged in, and they could also affect
the amount and quality of learning. Bailey (1996) made a further distinction between ‘washback to the learners’, which was the result of supplying ‘test-derived information’ to the test-takers, and ‘washback to the programme’, which was the result of supplying information to all of the other participants in the education system. Bailey also discussed the difficulties of investigating washback, which included working in ‘naturally occurring settings’, using a ‘non-random sample of subjects’ and attempting to sort out how much of what happens in classroom can be ‘evidentially linked to the introduction and use of the test’ (Messick, 1996, p. 242; also Alderson and Wall 1993, p. 117).

These discussions were complemented by empirical work on the impact of different types of tests in different areas of the world. The types of tests include national school examinations in Sri Lanka (Wall and Alderson, 1993; Wall, 1996), Israel (Shohamy et al., 1996), and Hong Kong (Cheng, 1997, 1998); university entrance examinations in Japan (Watanabe, 1996); and international proficiency tests (Alderson and Hamp-Lyons, 1996). Much of what was revealed by these studies had to do with what Hughes (1994) would call the “processes” of teaching: the selection of content (skills, teaching materials, exam preparation materials), the methodology teachers used and the ways in which they assessed their own students. The findings relating to ‘participants’ often had to do with the stress and anxiety felt by teachers and learners. Interestingly, little attention has been paid to the ‘products’ of learning. Wall (1999) analyses teachers’ impressions of what their students are learning and how this compares with students who took earlier examinations and Cheng (1998) surveys students directly. What is missing, however, are analyses of test results which indicate whether students have learned more or learned better because they have studied for a particular test. Two of the problems involved in investigating this area are the need to compare groups before and after the introduction of a new test, and the need for an independent test which measures the ‘right’ things (the aims of the curriculum) — both of which were difficult to provide in the situations that have been studied so far.

These studies have attempted not only to describe test impact (or lack of impact) but also to account for the form it takes. Whereas some of the early commentators made general statements about the relationship between tests and their impact (e.g. Heaton, 1990, p. 16 — “If it is a good examination, it will have a useful effect on teaching; if bad, then it will have a damaging effect on teaching”), later studies have recognised that test design is only one of the components in a quite complicated equation. Wall and Alderson (1993) identified factors such as teacher ability, teacher understanding of the test and the approach it was based on, classroom conditions, lack of resources, and management practices within the school. Other factors that have been identified include the status of the subject in the curriculum (Shohamy et al., 1996); feedback mechanisms between the testing agency and the schools (Shohamy et al.); teacher style, commitment and willingness to innovate (Alderson and Hamp-Lyons, 1996); teacher background (Watanabe, 1996); the general social and political context (Wall, 1996); the amount of time that has passed since the introduction of the examination (Shohamy et al.); and the role of publishers in materials design and teacher training (Cheng, 1997).
The challenge facing testers now is how to categorise these findings: is there a model available or can one be constructed which takes all of these variables into consideration while still remaining coherent and user-friendly? Alderson and Wall (1993, p. 127) suggested seeking insights in the field of educational innovation. Testing, like other areas of language education, had not by that time taken much account of findings in this area. White (1993, p. 45) claimed that the applied linguistics and language teaching literature “tends to skim over the issues of innovation”. Beretta (1990) claimed that while some attention was paid to the theory underlying innovative programmes, little was paid to questions of implementation. Bowers (1980) argued for a common descriptive framework so that the lessons learned in major educational projects could be passed on to future innovators.

A small but informative literature has now appeared which explains some of the ideas of innovation specialists such as Fullan (1991) and Rogers (1995) and shows how they are relevant to language education (e.g. Kennedy, 1988; Smith, 1989; Beretta, 1990; Stoller, 1994; Markee, 1997). It is beyond the scope of this article to explain these ideas in detail, but it is possible to list some of the most important observations (Wall, 1999, pp. 246–248):

1. Innovation is different from other sorts of change in that it is deliberate, it is planned, and it is designed to bring about improvements in the system it is introduced into.

2. The process of innovation is long and complex, consisting of many stages, and innovators should ask different sorts of questions at every stage to make sure that the innovation is going the way they want it to.

3. There are many participants involved in the process of innovation, each with their own needs and limitations. There must be good co-ordination and communication amongst all the participants if the innovation is not to founder.

4. The meaning of an innovation will be different for every individual involved in the process.

5. An innovation may require change on three different levels: content, methodology and attitudes. It is easier for teachers to change the content of their teaching than to change their behaviour and easier for them to change their behaviour than to change their attitudes or values.

6. The users of an innovation will reach different ‘levels of implementation’ (Beretta). Some will take full advantage of the innovation and be able to build on it, while others will never be able to respond to it in the way that was intended.

7. It is difficult to measure some kinds of changes, especially changes in awareness or changes which are open-ended.

8. Every innovation has a number of characteristics, some of which may facilitate its adoption and some of which may hinder it. Innovators may need to emphasise different characteristics when trying to appeal to different potential adopters.

9. It is necessary to analyse the context of an innovation in order to judge whether it is likely to be adopted. The context involves not only the classroom and...
the schools, but also the educational system, the political system, the cultural system, etc.

10. The rate of adoption of an innovation is determined by many factors, including the characteristics of the innovation, the context and the potential users.

11. There are a variety of models for introducing change, each accompanied by specific strategies and different leadership styles. Some approaches are more suitable for certain types of innovations than for others.

12. It takes time before an innovation can bring about fundamental changes, so innovators should not be too hasty in their judgement of whether innovations have succeeded or failed.

Henrichsen (1989) provides a useful synthesis of many of these ideas in his Hybrid Model of the Diffusion/Implementation Process (Fig. 1). Here, Henrichsen breaks the process of innovation down into three different components. The Antecedents component includes the conditions in place in the educational context or environment before the introduction of an innovation. The Process component includes factors which are in operation when the innovation is being implemented, some of which may facilitate the implementation and some of which may hinder it. The Consequences component includes the types of outcomes that may occur as a result of the interaction between the factors in the Antecedent situation and the Process. Wall (1999) uses various aspects of this model in a detailed analysis of an attempt to use new tests to influence teaching in Sri Lanka.

What this analysis showed was that it would have been difficult for any new test to have much influence on the type of teaching that was taking place in that particular

Fig. 1. Henrichsen’s Hybrid Model of the Diffusion/Implementation Process. Figure taken from “Diffusion of Innovations in English Language Teaching: The ELEC Effect in Japan, 1956–1968.” by Henrichsen, L.E. Copyright © 1989 by Greenwood Press. Reproduced with permission of Greenwood Publishing Group, Inc., Westport, CT.
setting, and that there were good reasons why the test that was introduced had more influence on some aspects of teaching than on others. The analysis of the Antecedent conditions revealed that there were traditional ways of teaching which were not likely to be affected without a great deal of teacher training and monitoring, and that even well-trained teachers would have difficulty introducing some of the new teaching ideas in the classroom and school conditions they found themselves in. The analysis of the Process showed that there were factors in the innovation itself (the test and the materials that it was based on) that made it difficult to understand and to work with. The teachers recognised that the test was more relevant to their students’ needs than the previous test, and that it had many redeeming features; however, many of them did not understand the principles underlying the test or the kinds of changes they needed to make in their teaching to help their students to perform better. There were factors in the ‘User System’ which also made it difficult for teachers to cope with the new test easily. These included poor classroom conditions, the limited time available for English lessons, teacher shortages, problems in communication within schools and between schools and the agency responsible for the test, the influence of institutions offering after-school tuition, and geographic, economic and political factors. The ability and commitment of the teachers were also very important: their own level of English, the amount and type of training they had received, the amount of teaching experience they had, and their own beliefs about language and education.

Once all of these factors had been identified it was necessary to examine the interaction between them, to determine whether the innovation (the new test) was compatible with the user system it was being introduced into. The major finding here was that it was difficult for the user system to respond adequately to the demands the test made of it: there was not enough teacher support material, there was too little teacher training, and there was not enough time for most teachers to cover the syllabus. There were too few links within the user system (between teaching associations, training colleges, universities, teacher, publications, etc.) to provide for effective exchanges of information and expertise. There were few rewards for teachers who tried to change their approach to teaching and they were pressured by students and their families to revert back to the traditional model of what teaching should be.

An analysis of the Consequences suggested that the impact of the test was partial rather than complete: it affected the way that teachers selected the content of their lessons but it did not affect their methodology. Although they spoke in favour of both the test and the materials it was based on, the attitudes that they professed did not always translate into the kind of behaviour that the policy-makers had intended and did not necessarily indicate a change in their basic beliefs about teaching. For the innovation specialist this would count as superficial change only (Fullan, 1991, p. 37). In Shohamy’s (1993) terms the impact that had been achieved was ‘instrumental’ rather than ‘conceptual’.

This study has confirmed the notion that test impact is a complex phenomenon, which should not be seen as a natural or inevitable result of introducing a new test into an educational setting. It should not be assumed that a ‘good’ test will automatically produce good effects in the classroom, or that a ‘bad’ test will neces-
sarily produce negative ones. All of the factors mentioned above will have some part to play in the process of trying to create change. It would not be reasonable to assume that all the factors will work the same way in every context, or that they will lead to the same outcome. It should be possible, however, with the help of models such as Henrichsen (1989) or the Wall (1999) adaptation to make more informed judgements about the amount of risk that will be involved in any new testing venture, and to help educators to think about the kinds of characteristics that tests need to display if they are to create positive impact in difficult circumstances.

The study of test impact is still relatively new in language education and there are clearly aspects which should be investigated further in the next decade. While it will be useful to continue to study the effects of tests on teaching, it is also important to investigate the effects they have on learning. We know very little about students’ perceptions of tests (as opposed to their teachers’ impressions of their perceptions) and even less about how new tests influence what students know and can do. It is also important to continue investigating how tests interact with other factors in the educational setting. This will require a thorough analysis of the context that the tests are going to be introduced into (the Antecedent situation) and dedicated monitoring of the process once the test is in place. What follows are guidelines which are intended to help researchers who wish to carry out such investigations:

1. It is valuable to use a framework such as Henrichsen (1989) or Wall (1999) when carrying out investigations into the feasibility of testing reform. It is especially important to analyse the antecedent conditions, in order to determine whether change is desirable and whether the education system is ready and able to take on the burden of implementation. This type of investigation is referred to as a ‘baseline study’ (Weir and Roberts, 1994; for examples, see the baseline studies produced as part of the St. Petersburg Examination Project [1996] and the Hungarian Examination Reform [Nikolov et al., 1999]).

2. It is crucial to involve teachers in these initial investigations and in all stages of planning; however, it is equally important to involve representatives of the key stakeholders (the institutions that are likely to be affected by the reform), including policy-makers, administrators, pre-service and in-service teacher trainers and students. It is important to talk to students directly rather than rely on their teachers’ perception of what they think and feel.

3. It is important that the reform management team include representatives of all of the key stakeholders, to facilitate communication and co-operation amongst these parties and to ensure that the new test and associated procedures respect not only the traditional criteria for judging assessment systems (validity, reliability and practicality) but are also comprehensible to teachers and students, acceptable to other relevant parties in the system, and fit in with the available training and development structures.

4. It is important to draw up draft test specifications which can be scrutinised by representatives of all the key stakeholders and be revised before being adopted as the official examination blue-print. The specifications should be accompanied
by sample test items, tasks and marking criteria so that those who review them will have a clear idea of what the demands will be on teachers and students. The specifications should also explain the purpose of the test, the underlying constructs, the levels embodied in the test, etc. (See Alderson et al., 1995, for a checklist of features to include in a test specification.) It will be easier for future users and evaluators of the test if this information is contained in a single document rather than being spread around in several different documents.

5. It is important to pilot new tests before introducing them full scale, both to try out the feasibility of the new ideas and procedures and to gather feedback from teachers and students. This is especially important when the components of the new test are very different from the test it replaces or when there is additional complexity in its structure.

6. Evaluation should take place from the earliest stages of a test reform project and at regular intervals, and it should cover not only the test design and procedures, but also the attitudes of teachers and students and the impact of the test on teaching and learning. The results should be disseminated to key parties within the education system so that decisions can be made about the match between the test and the curriculum, the appropriateness of the design and procedures, the adequacy of communication, the allocation of resources, the planning of further training, and changes in overall and specific policy.

7. Policy-makers and test designers should not expect significant impact to occur immediately or in the form they intend. They should be aware that tests on their own will not have positive impact if the materials and practices they are based on have not been effective. They may, however, have negative impact and the situation must be monitored continuously to allow early intervention if it takes an undesirable turn.

These guidelines have emerged from research and development work carried out on national examinations in the 1990s. The first decade of the new century will open up new possibilities, given developments such as the revision of the Council of Europe Framework, the implementation of testing systems across languages and the ever-increasing sophistication of internet test delivery (Alderson, 2000). It will be interesting to see whether testers working at regional and international levels and through new media will be as concerned with the impact that their tests have as policy-makers at national levels need to be, and whether guidelines such as the ones given above will be relevant or adequate in these new situations.

It will also be interesting to see whether the language testing community will continue to ‘plough their own furrow’ or whether they will take more account of work done in other areas of the curriculum (see, for example, Ridgeway, 1988, on assessment in mathematics) and by researchers in disciplines other than education.

References


St. Petersburg Examination Project, 1996.


