2 Taming the External Force of High-stakes Language Testing – Identifying Conditions under Which Tests Work for Improving EFL Practices

Yoshinori Watanabe
(Sophia University)

INTRODUCTION

The Role of Testing in TEFL and Its Problems

A number of tests are carried out each and every year all over the world. Their purposes are many, and vary widely. Some tests are conducted to gather information for making an important decision, whereas other tests are used in an attempt to motivate learners. Still many other tests may be simply done for "test's for test's sake" without any good reason for other than just doing it. While doing so, we may be running a risk of making ourselves a servant of the test rather than using it as an instrument for accomplishing a specific purpose.

Tests in Asian countries are always a hot issue. Some people claim that it is because of the influence of Confucius ideas, but it does not seem that there is any element inherent in the Confucianism that induces the importance of testing in education. One interesting observation relating this is made by Zeng, that is: "in Japan, Taiwan and Korea, invoking scholarship deities for help in passing exams is a notable religious activity (Zeng, 1999, p. 13). Indeed, to be accepted to a certain university, particularly the one which is acknowledged as prestigious, students must go through a religious ceremony or obtain a "right of passage" (Madaus, Rusell & Higgins, 2009); thus, testing becomes a kind of ceremonial rituals. In fact, however, it seems to be a universal phenomenon across countries and ages (Spolsky, 1995).

Passing the test is so important for students, not only for its primary purpose of getting a sense of being selected to a certain higher educational institution, but also for developing a sense of identity or a sense of belongingness. In other words, whatever the quality of the method of testing, irrespective of its reliability and validity, an important thing is to take the same test that is offered uniquely at the institution that would accept the candidate. This is an irony indeed; the egalitarian purpose of employing examination system turned out to be accelerating competition.

Amongst many problems of blind application of tests in educational systems as pointed out by Shohamy (2001), particularly problematic with the attempt to innovate in education through testing is a failure on the part of teachers and learners to come to terms with the pressures they are likely to receive from external high-stakes examinations in their in-class formative assessment practices and students' achievements in them. Regrettably, the purpose of using such external tests is often likely to be pretty unspecific, not to mention its questionable qualities. It becomes normal, then, that teachers are in dilemma, in a way in which whereas they notice that 'teaching to the test' does not help students become able to use the target language in real life settings, an effort to teach students to become able to use the target language does not seem to help students pass the test either. A similar view may also be held by the students, though there is a difference in the degree of awareness.

Before going into the main part of the present investigation, one further remark is in order about the present purpose. The present paper does not purport to report on the state of testing in Japan. There are other publications that have been published to date, which would serve to that end. For example, Rohlen (1983) and White (1988) both give an ethnographic account of the life of teachers and students in the Japanese educational system with particular reference to primary and secondary levels, including an in-depth description of the examination system. Henrichsen (1989) investigates the effort the English teaching organization called ELEC made in innovating in the teaching of English in Japan. While so doing, he provides an interpretation of the role of testing in the framework of innovation theory. Amongst more recent contributions, Watanabe (2004) looks into the role of teachers in

inducing beneficial effects of university entrance examinations to EFL classrooms in Japan. Tanaka (2008) traces the history of EFL testing in Japan all the way back to the middle of the 19th century, and offers a set of useful suggestions for improving the status quo of EFL testing in the country.

Instead of giving an account of EFL testing in Japan, then, the present paper is intended to reveal the mechanism of using testing to improve EFL practices by referring to two cases that were observed in two distinct contexts which differed greatly, in the USA and in Japan. By doing so, it is expected to render several principles in the form of propositions. In other words, in the present paper an attempt will be made to seek for universal principles of using language testing for improving EFL, rather than seeking to establish the particularity of the principles of educational use of language testing in Asian countries per se, with special attention to Japan. Thereby, the present paper tries to generate hypotheses to be tested for their validity in the future empirical research and/or in the context where English is taught and learned as a foreign language. The effort I make in the present paper could be said to put forward my "personal knowledge" in a sense defined by Polanyi (1958); that is, by means of such knowing, I tried to "establish contact with a hidden reality; a contact that is defined as the condition for anticipating an indeterminate range of yet unknown … true implications" (pp.vii – viii).

The Primary and Secondary Purposes of Testing in EFL

The purpose of the present paper is to reconcile the two contradicting forces, teaching and learning to the test and improving proficiency in English as a foreign language, which may work negatively for classroom instruction. The ultimate purpose of the present study is to identify a range of conditions under which language tests may be used to help improve "the quality of life in the classroom" (Allwright, 2005; Allwrigtht & Hanks, 2009), which admittedly too long a distant goal that the present short paper is unable to reach. In order to derive suggestions for making the best use of language tests for pedagogical benefits, particular reference will be made to the theory of educational innovation, the theory of motivation and performance, and the findings

that have been made to date in the field exploring the issue of washback effects of language testing. On the basis of the analysis of a variety of actual cases of test uses, several practical recommendations will be provided to help teachers make the best use of high-stakes tests in the language classroom by taming its potentially negative forces. Several suggestions will also be offered for testers to produce tests the way of administering the test in a way in which that will be useful for teachers and test-takers alike. In so doing, the guiding principle adopted will be the first principle of Exploratory Practice, that is, "try to understand first, before you try to change anything, in case you discover that change is not necessary, or perhaps not desirable (or perhaps not actually possible)" (Exploratory Practice Centre, 2009).

I would argue that the primary purpose of using a language test may be pedagogical, in the sense that tests are used as its primary goal to change some aspects of education, be it the content of teaching, pacing the instruction, cultivating the students' study habit, or whatever. In this use of language testing, obtaining the psychometric data with reliability and validity to be established has to be sacrificed to some extent. I would further argue that though the test may serve two purposes, one being to provide an accurate set of information to help make a decision, and the second being to bring educationally beneficial practice in EFL classrooms, and the former psychometric use of language testing is normally taken as the primary purpose, the second pedagogic purpose of using the test is equally or often more important than the first purpose. The latter type of using tests is primarily concerned with the test per se, in the sense that taking and carrying out the test itself is a major concern of the test takers and test administrators, and it is assumed that the latter case may be allowed to involve subjectivity to a certain extent, though there will be no specific argument supporting this position in the present paper due to space constraint.

The present discussion is based on a very simple assumption and follows a very simple syllogism based on it. That is:

1) Language tests need to contribute to the well-being of students and teachers in the language classroom. In other words, language tests need to help learners to learn the target language rather than hinder it.

2) However, a test by its nature seems to have an undesirable element inherent in it.
3) Therefore, a deliberate and informed effort needs to be made to tame the force on the part of testers and test users.

The subsequent sections are so arranged to spell out each of these propositions. In so doing, it will be exemplified what sort of deliberate attempts need to be made.

IDENTIFYING PROBLEMS

The Basic Assumption

Contribution of Language Testing to the Well-being of Students and Teachers

The basic assumption for the present paper may appear to be too obvious to require any further argument. That is, language testing needs to contribute to the well-being of students and teachers in the classroom. Nevertheless, it is necessary to confirm that the readers of this paper share a common understanding. First and foremost, we may need to understand that there is a difference between the use of *test scores* and the use of *tests* to consider the importance of testing in education.

There are many ways in which a test serves its purpose. Any textbook on language testing lists various purposes, such as selection, placement, diagnosis, monitoring progress, grading, and so forth. However, it should be noted that these are the uses of test *scores* rather than the test itself. Thus, it is normal to take the following view in the research into language testing: "once we have made inferences about language ability, we may use these inferences for a variety of secondary purposes" (Bachman & Palmer, 1996, p. 96). These 'secondary' purposes vary widely, ranging from selection to assessing the effectiveness of teachers/teaching, assessing the effectiveness of programs. In the use of tests as a measurement device, it is certainly important to establish objectivity, in the sense that the measurement helps make an inference as to the ability each test taker possesses accurately.

Despite these undoubtedly crucial purposes of language testing,

however, the other purpose to generate positive effects to learner and teachers will be equally important in educational settings, particularly for teaching practitioners rather than researchers. In the latter case of using a test, teachers may use a test to motivate students, given them a chance to do a self-study at home, use testing as a preview of the content that will be covered in the upcoming lesson, and so forth. These are not secondary, but the primary purpose of using the test. In an extreme case, they may need to sacrifice the other purpose that is to obtain scores so they may make an accurate inference about the ability of the learner.

Shohamy (2001) argues that tests are often used to manipulate test users, and tests have built-in features that help this type of misuse. When she so claims, she at least in some cases seems to be assuming that using a test for the purpose other than obtaining test scores is not legitimate; "Tests are capable of dictating to test takers what they need to know, what they will learn and what they will be taught" (Shohamy, 2001, p. 17). In order to prevent such a misuse of tests, she proposes the notion of *critical language testing*, which "implies the need to develop critical strategies to examine the uses and consequences of tests, to monitor their power, minimize their detrimental force, reveal the misuses and empower the test takers" (p. 131). The present paper endorses her position, but differs in that it attempts to give a greater meaning to the use of tests in addition to obtaining accurate measurement, claiming that making use of a test and test scores to achieve an educational goal ought to be justified. Unlike the case of using tests to make high-stakes decisions as Shohamy and other researchers seem to assume when discussing the power of testing, using language tests for pedagogical purposes is done out of a naïve expectation on the part of teachers with pure benevolence of helping learners learn better than otherwise, and, I would argue, that such a use of tests ought to be justified. In the idea of using tests for educational purposes sacrificing its use for obtaining accurate measurements, the purpose of using language testing may include not only an attempt to motivate students, but also motivate teachers and innovate in education by innovating in testing, again with the good intention on the part of educational administrators.

Negative Features of Testing

Negative Features Inherent in Testing.

Having said that one of the important functions of language testing is to help learners learn and teachers teach in a better way than otherwise, it is quite common to find a number of critical comments on the negative aspects of examinations particularly in general public opinions. There is indeed apparently a negative assertion, such as "… unfortunately, most standardized tests are not intrinsically motivating; they promote competition and social comparison with an exclusive focus on outcome measures of achievement" (Paris, Lawton, & Turner, 1992, p. 232). Based on a piece of anecdotal evidence, I have been having a 90-minute session for teaching the basic notion of language testing to the students majoring foreign studies at university each year. The students will not become a teacher in the future, not to mention a researcher in the field of language testing. And yet they have taken a number of tests and will sit even a greater number of tests in the future for getting a job, getting a promotion, etc. At the beginning of each session, I always ask them whether there is anyone who 'loves' testing in a joking kind of way. Out of virtually 100 students in the lecture hall, only one student puts up his or her hand also in a joking way. I also ask them to write what comes to mind when they hear the word 'testing,' then their typical responses are "I hate it," "preparing for it overnight," "competition," "bothering me a lot," "gets me anxious," just to list a few (Watanabe, 2007).

It seems to be undoubtedly true that a test has something undesirable by its nature, in the sense that it is likely to cause a kind of anxiety on the part of learners and teachers. A perhaps a naïve view about the negative aspect of testing is observed not only in the comments of learners and teachers, but also in academic circles as well. One of the extreme cases of such a negative assertion is the comment made by Foucault; "The examination combines the techniques of an observing hierarchy and those of a normalizing judgment," and he continues: "it is a normalizing gaze, a surveillance that makes it possible to qualify, to classify and to punish" (Foucault, 1978, p. 184). Examining a variety of findings that are reported in the field of classroom research, it may strike us that what is undesirable for learners in the classroom may also characterize testing practices. For example, Allwright and Bailey (1990)

summarize the findings that are gained in classroom observation studies to date, where they report that learners tend to lower self-esteem, if the teacher puts pressure on them to monitor their own speech and correct themselves as they go along, if teacher constantly interrupts to correct students, and if teacher simply repeats the question in its original wording when students fail to respond (Allwright & Bailey, 1990). It is also illustrated that learners may constantly feel that they are representing themselves badly, showing only some of their real personality, only some of their intelligence (Allwright & Bailey, 1990). It may be that a test is having students lower their self-esteem by 'putting pressure on them to monitor their own' language and 'correct themselves as they go along,' or students may be frustrated because they are representing only part of their personality and intelligence.

Indeed there is a claim that tests have "the built-in features that allow their powerful uses" (Shohamy, 2001, p. 37). However, in the classroom, teachers do not necessarily intend to exercise power over students to manipulate them, but rather they may want to help students learn, though there is indeed a case where they may use it as a tool of punishment. When it comes to the issue of using the test for improving education, then, only if we understand the reason can we come up with concrete suggestions that are offered for improving education. If so, it is necessary to alleviate the negative effects on the part of teachers and test administrators.

Fear of Taking a Test

Another indication that there may be something undesirable inherent in testing comes from the research into educational innovation. Various attempts have been made to innovate in education by innovating in testing, and the outcomes seem to be endorsing the idea that any attempt to innovate in education by innovating in testing is doomed to failure. It seems as if tests were something that no one wants to be bothered by, but something that may be needed by someone for some purposes.

While there are many research studies reporting on the results of attempts to innovate in education by using tests with unexpectedly negative outcomes (e.g., Hillocks, 2002; House, 1998), one such a case that has been documented recently considers the No Child Left Behind (NCLB) program in the USA (e.g., Gallagher, 2007; Meier & Wood,

2004; Nichols & Berlinger, 2007; Perlstein, 2007). Due to space constraint, it is not possible to spell out the whole picture of the program, but it is important to notice that it poses an interesting but serious issue relating to the use of language testing for the benefit of language education in general, and the teaching of a foreign language in particular (Great Schools, TM). The program was first inspired by the report made by *A Nation at Risk* (National Commission of Excellence in Education, 1983), which reported on the demise of the then undesirable consequences of the educational programs in the USA, concluding that "if an unfriendly power had attempted to impose on America the mediocre educational performance that exists today, we might well have viewed it as an act of war." "As it stands," the report concludes, "we have allowed this to happen to ourselves …" (p. 5).

Recent publications reporting on the consequences of the NCLB has the title or its subheadings running with the words "collateral damage" (e.g., Murray, 2005), meaning "inadvertent casualties and destruction in civilian areas caused by military operations" (*Concise Oxford English Dictionary, 10th ed.*). This is a notion very similar to the one that has been used in the field of language testing as "unintended washback effects of language testing." Messick (1989) writes: "Judging validity in terms of whether a test does the job it is employed to do … requires evaluation of the intended or unintended social consequences of test interpretation and use. The appropriateness of the intended testing purpose and the possible occurrence of unintended outcomes and side effects are the major issues" (Messick, 1989, p. 84).

Many kinds of collateral damage or unintended washback effects of NCLB are reported in various professional publications as well as the media, and the findings include the ones which are relevant to the teaching of English as a foreign language as well as general education. For example, it is reported that there has been a marked decline in time spent on foreign language instruction in schools with largest populations of minority students and increased in class time focused on reading and English language arts and mathematics (Rosenbusch, 2005). More important findings, which are relevant to the present topic, are as follows, however: Teachers feel like scapegoats (Feller, 2003), testing consumes instructional time (Benett, 2002), students fear the exit exam (Tessler, 2003), and test success varies by students' socioeconomic

status (MacDonald, 2003). These findings appear as if they were echoing the reports that have been made to date: In appropriate curricula and repeated failures at school and on tests have resulted in discouragement, anxiety, fear, and diminished motivation to learn and work hard. As Eells and Davis (1951) noted, "to the average lower-class child … a test is just another place to be punished, to have one's weaknesses shown up, to be reminded that one is at the tail end of the procession" (Eells & Davis, 1951, p. 21, cited in Samuda, 1975, p. 87). Fullan makes an important comment regarding this:

> "Deutschman quotes Dr. Edward Miller, the dean of the medical school and the CEO of the hospital at John [sic] Hopkins University, who talks about patients with severe heart disease. Miller says, "If you look at people after coronary-artery bypass grafting, two years later, 90 percent of them have not changed their lifestyle. Even though they have a very bad disease and they know they should change their lifestyle, for whatever reason, they can't" (Deutschman, 2005, p. 2, cited in Fullan 2007, p. 42).

Fullan continues, "fear, as in fear of dying, turns out not to be a powerful motivator beyond an initial immediate effect. Similarly, in the United States, fear of not meeting "adequate yearly progress" in No Child Left Behind legislation, with its increasingly punitive consequences, is not much of a motivator – perhaps a little, but only in the very short run" (Fullan, 2007, p. 43). Though our knowledge is too limited to offer suggestions for practical purpose in education, it may be possible to surmise that in order to make test work for learning at school, several conditions need to be met. To find such conditions is one of the most important tasks that researchers have to do with practitioners.

Need for Alleviating the Fear

If there is something undesirable in testing, then, a deliberate attempt ought to be made on the part of test users to make the best use of its power beneficially. In order to minimize the potential negative use of a test, it is undoubtedly important as well as useful to promote the use of codes of practice, and it is putting in effect by organizations such as International Language Testing Association, but it is likewise important

to understand the principles and mechanism of how tests work in educational contexts to enhance the quality of life. This is particularly true for the teachers and learners who will not be directly involved in testing practice in a way in which the knowledge of the code of ethics is required.

Indeed, many research studies on the washback effect of language testing or the effect of testing to the classroom, to date report that the variance between teachers is greater than the variance between different target examinations (e.g., Wall & Alderson, 1993; Watanabe, 2004). Similar findings are also reported in the field of general education: "the differences among teachers [within a school] is [sic] substantial in comparison to the variance between schools (Nye et al., 2004, p. 247, as cited in Fullan, 2007, p. 53). An even more interesting part of the finding is that differences in teaching between teachers within the same school were bigger than differences in teaching between schools. These findings imply that it's left up to the test user to make it better or worse. These results strongly suggest that personal factors are crucial for improving education rather than simply putting teachers in different school systems. The key factor to making the best test use is then "how to help people feel and be better" (Fullan, 2007, p. 43), which is not dissimilar to the suggestion to be offered to any subject area.

In order to offer suggestions as to how best to use the test for the purpose, then, we need to understand the nature of examinations in educational settings and how testing works or do not work in the contexts as a device for improving teaching and learning. Alderson and Wall (1993) suggested that in order to explain why or why not washback is engineered, the theory of innovation and motivation theories provide a useful set of guidelines. The present paper is not intended to review the whole array of research in the field, which has been done in other brilliant works (e.g., Andrews, 2004; Cheng, 2005; Green, 2007; Wall, 1996; Wall, 2005). The purpose here is to relate some evidence based on the project that the present author has been personally involved, and draw some implications therefrom for solving the issue of taming the force of testing, which may likely be negative.

EXPLORING SOLUTIONS

The Issue of Motivation

Using Test to Motivate Students

While tests are deemed to possess negative features that might engender misuse to the detriment of teaching and learning, there are a number of suggestions that are offered to motivate learners by testing on an in-class level of instruction to date just to list a few of them.

- "…learning and thinking are better when students are involved in the [authentic] task because they find it interesting and challenging rather than when they are working for external rewards or merely to complete the job…" (Nicholls et al., 1989, cited in Paris et al., 1992).
- A task based on comparison among students reduced intrinsic interest, whereas a task based on achieving a predetermined goal increased intrinsic interest (Harackiewicz, Abrahamas & Wageman, 1987, cited in Stipek, 2002).
- Criteria for evaluation affect the information students use to evaluate themselves. Feelings of satisfaction in the competitive situation were based on whether they won or lost, not in the quality of their performance, whereas children in individual goal structure focused on their personal history with the task (i.e., whether they improved) (Stipek, 2002).
- When grades or other forms of evaluation are given, base them as much as possible on effort, improvement, and achieving a standard, rather than on relative performance (Stipek, 2002).
- Emphasize the information contained in grades (Stipek, 2002).
- Make grading criteria clear and fair (Stipek, 2002).
- Provide substantive, informative feedback, rather than grades or scores on assignments (Stipek, 2002).
- De-emphasize external evaluation (Harter, 1978).
- Use grades in a motivating manner, reducing as much as

possible their demotivating impact (Dörnyei, 2001).

- Make the assessment system completely transparent, and incorporate mechanisms by which the students and their peers can also express their views (Dörnyei, 2001).
- Make sure that grades also reflect effort and improvement and not just objective levels of achievement (Dörnyei, 2001).
- Apply continuous assessment that also relies on measurement tools other than pencil-and-paper tests (Dörnyei, 2001).
- Encourage accurate student self-assessment by providing various self-evaluation tools (Dörnyei, 2001).

Despite these recommendations, however, research to date indicates that there is very little guarantee that tests motivate learners, not to mention a powerful tool that enhances innovation in education (e.g., Alderson & Wall, 1993; Cheng, Watanabe & Curtis, 2004; Cheng, 2005; Green, 2007; Wall, 2005). In order to motivate students by testing, other factors than the test itself need to be taken into account, such as the purpose of enrolling in the class, the classroom climate, the affective factors, and so forth (e.g., Moeller & Reschke, 1993). To find a solution is the purpose of the next section.

Levels of Tasks Difficulty

The first implication of EFL practices that can be drawn from a research finding concerns the difficulty level of test tasks. In order to use tests to motivate learners, it seems that the level of *perceived* difficulty of the test task may have to be slightly more challenging than the *perceived* level of an individual's proficiency in the target language.

Very little research has been conducted to directly observe to identify the conditions under which students are motivated to learn. Watanabe (2001) conducted an interview study examining whether high-stakes university entrance examinations motivated those students who had been preparing for the examinations for a certain number of years. Based on the results of interviewing a group of students, it was concluded that the examinations did not motivate all the students to the same degree in the same way. If they perceived the test to be too

difficult, they simply gave up learning for it. Likewise if the test is too easy, they also gave up preparing for it, because they did not find it necessary to bother.

The students seemed to be motivated if they perceived the target examination slightly more challenging than their perceived level of their proficiency in English. This looks as if it were supporting the theory of flow Csikszentmihalyi (2000) reports when he suggests that the challenge level should be appropriate for the one to be motivated. It is important to note that it is not "only the 'real' challenges presented by the situation that count, but those that the person is aware of" and it is not "skills we actually have that determine how we feel, but the ones we think we have" (p. 75).

Stakes of Testing

A research study suggests that a test may be an effective tool for motivating proficient learners even further, but it does not seem to be appropriate for motivating weak learners. Watanabe (2005) carried out an interview study with approximately 90 students and 40 teachers teaching English as a foreign language in Japan for over three years. He asked students and teachers to recall as many incidents that were specifically related to the test that they took and which stays in mind (i.e., an antecedent event), and write what the incident caused them to do (i.e., consequential attempt). For example, one student wrote "I arranged a schedule and began test preparation three weeks before the test, and got 'good' scores" (an antecedent incident), and "I relaxed too much to prepare well enough for the next test" (consequence). Another student wrote: "I worked hard for the class every day, and scored very high" (an antecedent incident), and "I'd begun working harder for the class than before" (consequence). Results indicated that there were variations in the degree of the power of testing on students in terms of motivation. Some students seemed to take the exam as a chance to motivate themselves, whereas other students seemed to take an attitude towards the test detached from the outcome. Still other students seemed to have lost interest in English because of the incident that they reported having relating to the test. Based on the findings, the study concluded that the test motivates those students who are already motivated, but cannot motivate those students who are not motivated. An obvious

corollary of this is that if tests were used as a means of motivating students, it would run a risk of making those students who are not interested in learning the target language even more unwilling to learn.

The Case of Innovating in a High-Stakes EFL Examination in Japan

In this section, the role that a high-stakes test plays in improving EFL practices will be investigated yet from a different angle. That is, the topic will be to identify the conditions under which an examination should be innovated in so that it may be educationally beneficial, while in the previous section the issue was dealt with as to how to motivate students by innovating in testing. In so doing, a reference will be made the observation that I reported in Watanabe (2009) regarding my experience of serving as a leader of the committee of the National Center for University Entrance Examination. Before going into details, some background information regarding the Japanese educational system is in order below.

The Center Examination and Its Background

In Japan the school year begins on April 1 and ends at the end of March the following year. The entrance examinations of most universities are administered during the period of January to February, though the so-called 'recommendation exam' is held earlier. Each department of each university produces its own examination and offers it on its own campus, though there is a difference in the screening system between national/local public universities and private universities. The universities are divided into several types. In both two-year course junior colleges and four-year course universities, there are three different types according to the establishment basis; i.e., national, local public, and private. Out of approximately 600 four-year course universities, 20% are national, 10% are local public, and 70% are private.

The national and local-public universities on the one hand, and the private institutions on the other, employ different procedures to screen out entrants. In the former type, the applicants are required to go through two stages. First, they take the National Center for University

Entrance Examination (hereafter, the Center Examination). Unlike the national and local public institutions, private universities enjoy greater freedom in their screening procedures and examinations. Some institutions employ the Center Examination in the first stage, and administer interviews and/or essay tests in the second stage, while other universities employ the two- stage screening system, without using the Center Examination. However, most of the universities choose students by their own examination of paper and pencil type on their own campus.

The Center Examination is a very high-stakes test indeed, and carried out only once a year. In 2009, a total of 507,621 students took the test. After the administration of the test, a number of questions and criticisms are sent to the Center by the students, the teachers, and others who are interested in the examination. The content of the test is likely to be a national issue, and encounter many criticisms, mostly harsh rather than friendly, every year. Confidentiality, fairness, and other ethical concerns are very important and specially prepared guidelines are strictly followed by the committee members, who are selected from among the university faculty of various universities in different areas of the country. The length of service is two years. Throughout these years, the staff at the Center is likely to be very nervous. During the process of developing the test, items and tasks are constantly reviewed by various parties, including the education board, the previous committee members, as well as the present committee members. The listening test is carried out by an audio machine that is distributed individually to each test-taker, so it may avoid unfairness in the quality of recordings that may vary at the seating place. One of the things that make the news each year is how many machines did not work properly, and how many students had to take a make-up test, though the number is usually negligible statistically. Given the high stakes of the examination, the staff is also likely to be very nervous particularly about making changes in the content and tasks of the examination. What worries the committee the most each and every year is a criticism that is anticipated to be directed to the changes that will be made, especially when they are made without any pre-announcement.

Changing Test Tasks and Making it Public

When I served as a member of the committee at one time, one of the biggest issues was to innovate in the construction of the test. The Center Examination publishes an annual report, documenting the results of the examination based on the reviews that are made by three different parties, including high school teachers, an academic organization on EFL, and the test development committee. When I was on the committee, there was one test task type that had been constantly criticized for its validity. Though all the committee members agreed on this, we soon learned that it would be extremely difficult to change it because of time constraint on schedule. If changes would be made on the examination to any degree, an announcement had to be issued in advance to those who would be involved in the test, including all the stake-holders including parents, teachers, university admission officers, as well as test-takers. But we found that the schedule would be too tight to allow for such a pre-test announcement. This meant that even though a substantial number of committee members felt the necessity of changes in the test task, there was a concern that we would run out of time for making an announcement by following the schedule faithfully.

We were thus in dilemma. We were aware of the fact that there had long been criticisms about one of the task types, but we understood too well that it would be very difficult to change it under the tight schedule. The strategy that we ought to take was to cooperate with the committee members first. Next, we should convince the management staff at various levels of the Center, who would directly cope with it if something undesirable would happen. Solutions had to be made with several limitations. First, it would be too late to make a public announcement. Second, despite its awareness, it seemed to be counterproductive to continue with the test task, which comprised several sections of the whole test battery. Examining the past test papers revealed that minor changes had been made in some years in the past, but it was not sure exactly at what timing the change had been made, and nor was it clear how much changes would be sensible.

The solution that we finally reached among all the staff was as follows: we would change the problematic part of the test task anyway, though we understood that it would not be possible to make a public announcement in advance at least that year. But we promised that we

would make a minor change to the degree that that would not disturb test takers, that the change would be so made to render the task more congruent with the Ministry of Education Guidelines, and that we would prepare the fair and sound accountability documents in case we be required to provide it after the administration of the examination. By so doing, we tried not to make changes overly innovative, making the greatest effort to make new tasks, so they might appear to be friendlier to test takers than the current type.

Implications for Making the Examination Acceptable to Students and Teachers

The result of changing the test tasks turned out to be successful, in that that did not disturb test takers so much as we had been worried. The changes seemed to be rather well accepted by the general public as well as teachers not to mention test takers. The Center received surprisingly few criticisms after the administration, and there were even fewer questions about the content and the quality of the test. The mass media, including the newspapers, the internet, and TV broadcasts reported favorable comments, contrary to our concerns. Indeed some media reported that the post- examination interviews with test-takers revealed that the change had disturbed some students. But the number of such reports was surprisingly small. The comments that were summarized in the annual report of the Center included several comments about the changes, but they seemed to be requests for future tests rather than criticisms (translation mine: National Center for University Entrance Examinations, 2007):

- If changes have to be made in the examination in the future, it should be announced in advance (p. 362).
- This year, the test questions were appropriate in many ways. This said, though, there were a good many new types of tasks and items in this year's examination. It is expected that the committee gathers as much feedback as possible from various stake-holders and makes a greater effort to develop an examination which is fair in that those students who work hard at school will be rewarded (p. 366).
- There are still many sections that ought to be improved; for

> example there are the items which could be covered in the listening component more appropriately than the written component (p. 366).

The reservation needs to be made for these comments, however: that is, these are the 'formal' documents, which were collected from those teachers and researchers who were in the position of being directly in contact with the Center. But the voices of test-takers and the teachers who prepared them were not heard. With these reservations, however, the above illustration seems to give us a useful lesson for understanding how to make use of tests to improve education. The meaning of what happened after the introduction of a new test task type can be better understood by referring to a theory of innovation, which also makes it possible to understand what needs to be done in the future to improve it even further. In an innovation theory, the notion of innovation is usually defined as "an idea, practice, or object that is perceived as new by an individual or other unit of adoption" (Rogers, 2003, p.12). Though it is not possible to fully explicate the theory in the present paper (see Andrews, 2004, and Wall, 2005 for details), but suffice for the present purpose to understand the main elements that will be essential for rendering educational innovation by testing successful.

Five Attributes of Innovative Technologies

Rogers (2003) proposes that there are five main attributes of innovative technologies which influence acceptance. These are relative advantage, compatibility, complexity, trialability, and observability. *Relative advantage* is the degree to which an innovation is seen as superior to prior innovations to fulfill the same needs. *Compatibility* is the degree to which an innovation appears consistent with existing values, past experiences, habits and needs to the potential adopter. *Complexity* is the degree to which an innovation appears difficult to understand and use. *Trialability* is the perceived degree to which an innovation may be tried on a limited basis. Trialability can accelerate acceptance because small-scale testing reduces risk. *Observability* is the perceived degree to which results of innovating are visible to others and is positively related to acceptance (Rogers, 2003, pp. 15 - 16).

Rogers' version of innovation theory is helpful for us to draw

important implications from what is illustrated in the above report for innovating in testing in an attempt to improve EFL. First, three conditions seem to have been satisfied, in that 1) the task types that had been criticized were improved (relative advantage), 2) the changes that were made on the test task did not deviate from the Ministry of Education Guidelines, but rather became more relevant to them (compatibility), and 3) the tasks became less complex and thus easier for test takers to carry out (complexity). Despite improvement in these areas, however, the fourth and fifth conditions, testability and observability, could not be met. The confidentiality of the examination is very important, which makes it extremely difficult if not completely impossible to give test takers a chance to try it out prior to the main test. Though it is surely possible for them to get used to it by going over the past exam papers, it is not possible to do so for the part of the test, on which changes would be made. This means that it is not possible to examine how acceptable the change in the test would be to test takers. A solution would be made by making a public announcement to test takers on a regular basis, for example, every three years to have them informed as to whether there would be a change or not in the upcoming examination, and if there would be changes, sample items be provided. By so doing, it becomes possible for test takers to try them before taking the test. In summary, then, the following suggestions are offered in order to make the test acceptable to an educational setting.

- *Give test takers a chance to try it out prior to the main test.*
- *Give test takers a chance to 'see' how well or poorly they would do it.*
- *Make the test congruent with the skills and the knowledge that students are expected to have acquired.*
- *Make the test comparable to what they have been learning in the classroom.*
- *Avoid making the test overly complex, too complex to deal with.*

CONCLUSION

The purpose of the present paper was to identify conditions under

which tests be used to enhance the quality of life in EFL classrooms. It was based on the observation that the test by its very nature is likely to provoke a type of anxiety among test takers. The observation leads to the idea that a deliberate attempt needs to be made to make the best use of tests for educational benefits. In so doing, two of the relevant fields of research were referred to; one is the research into motivation and the other is a theory of innovation. Within the framework provided by these fields of research, various cases were examined, including the consequences of testing on students' motivation, and the consequence of having changed a high-stakes examination to test takers and various other stake holders. Both these cases seem to endorse the claim Fullan made regarding the relationship between change and assessment: "Change is personal" (Fullan, 1998, p. 255), in that a test itself is neutral, and it is testers and test users who make it educationally better or worse. This means that conscious efforts, which are informed by the result of empirical research studies, need to be made.

In order to motivate students by testing, several factors need to be taken into account other than those which may contribute to understanding the psychometric nature of language ability of the students. For example, in order to motivate students by testing, the difficulty level of the test task ought to be slightly more challenging than the proficiency level of the test taker. One other suggestion provided in the present article is to make the stakes of each administration to a test not too high. This implies that a test taker needs to be given multiple chances to exhibit his or her ability on different occasions by multiple means of testing.

The description of the present paper is admittedly sketchy, so it requires much more hard evidence to draw more useful implications for education. Nevertheless, it could be proposed that the research into language testing in the future needs to take account of affective as well as cognitive factors which are deemed to be involved in language assessment practice, because if the test is likely to cause fear on the part of test users in general and test takers in particular, then to know the source of fear or anxiety may help alleviate such a feeling. Recent years have witnessed a new movement towards exploring the social aspects of language testing (McNamara & Roever, 2006), which obviously involves the issue of interpersonal relationship between testers,

test-takers, teachers and other test users. Whatever relationship is focused, the most important stake-holder is the test-takers. This means that in order to make the best use of testing for TEFL, we ought to understand the nature of test takers. The role of socio-affective factors will become even more important if we wish to use tests for improving education. Asian countries may share something special in the way of dealing with affects in life in general, and in education in particular as well (e.g., Hawkins, 1994; Reagan, 2005). Perhaps there is a large area lying in front of us yet to be explored, and awaiting for a number of contributions that could be made from the new perspective to the field of language testing.

REFERENCES

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, *14* (2), 115-129.

Allwright, D. (2005). Developing principles for practitioner research: The case of exploratory practice. *The Modern Language Journal*, *89*(3), 353-366.

Allwright, D., & Bailey, C. (1990). *Focus on the language classroom*. Cambridge: Cambridge University Press.

Allwright, D., & Hanks, J. (2009). *The developing language learner: An introduction to exploratory practice*. London: Palgrave Macmillan.

Andrews, S. (2004). Washback and curriculum innovation. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 37-50). New Jersey: Lawrence Erlbaum.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bennett, V. (2002). Testing the schools, discussion. *Frontliner PBS Program*. Available online: http://www.pbs.org/wgbh/pages/frontline/shows/ schools/ talk/.

Cheng, L., Watanabe, Y., & Curtis, A. (2004). *Washback in language testing: Research contexts and methods*. New Jersey: Lawrence Erlbaum.

Cheng, L. (2005). *Changing language teaching through language testing: A washback study*. Cambridge: Cambridge University Press.

Csikszentmihalyi, M. (1990). *Flow: Psychology of happiness.* London: Ryder.

Deutshman, A. (2005, May). Change or die. *Fast company, 94,* 53-57.

Dörnyei, Z. (2001). *Motivational strategies in the language classroom*. Cambridge: Cambridge University Press.

Eells, K., & Davies, A. (1951). *Intelligence and cultural differences*. Chicago: The University of Chicago Press.

Exploratory Practice Centre. (2009). Basic principles for exploratory practice. Retrieved September 1, 2009, from the World Wide Web: http://www.letras.puc-rio.br/epcentre/background.htm.

Foucault, M. (1978). *Discipline and punish: The birth of the prison.* (A. Sheridan, Trans.). London: Allen Lange (original work published 1975).

Fullan, M. (1998). Linking change and assessment. In P. Rea-Dickins & K. P. Germaine (Eds.), *Managing evaluation and innovation in language teaching: Building bridges* (pp. 253-262). London: Longman.

Fullan, M. (2007). *The new meaning of educational change* (4th ed.). New York: Teachers College Press.

Gallagher, C. W. (2007). *Reclaiming assessment: A better alternative to the accountability agenda*. Protsmouth, NH: Heinemann.

Great Schools TM. What the No Child Left Behind Law means for your child. Retrieved July 17, 2009, from the World Wide Web http://www.greatschools.net/improvement/quality-teaching/no-child-left-behind.gs?content=61

Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education*. Cambridge: Cambridge University Press.

Harackiewicz, J., Abrahams, S., & Wageman, R. (1987). Performance evaluation and intrinsic motivation: The effects of evaluative focus, rewards, and achievement orientation. *Jouranl of Personality and Socail Psycholgoy, 53,* 1015-1023.

Harter, S. (1978). Effectance motivation reconsidered: Toward a developmental model. *Human Development, 21*(1), 34-78.

Hawkins, J. N. (1994). Issues of motivation in Asian education. In H. R. O'Neil, Jr., & M. Drilings (Eds.), *Motivation: theory and research* (pp.101-115). Hillsdale, NJ: Lawrence Erlbaum.

Henrichsen, L. E. (1989). *Diffusion of innovations in English language teaching: The ELEC effort in Japan, 1956-1968.* New York: Greenwood.

Hillocks, G., Jr. (2002). *The testing trap: How state writing assessments control learning*. New York: Teachers College Press.

House, E. (1998). *Schools for sale*. New York: Teachers College Press.

MacDonald, M. (2003, October 9). County does well on state curriculum test. *The Atlanta Journal-Constitution*. Retrieve at June 21, 2009, from http://www.ajc.com/metro/content/metro/gwinnett/1003/09crct.html.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.

Madaus, G. , Russell, M., & Higgins, J. (2009). *The paradoxes of high stakes testing: How they affect students, their parents, teachers, principals, schools, and society*. Charlotte, NC: Information Age Pub Inc.

Meier, D., & Wood, G. (2004). *Many children left behind: How the No Child Left Behind Act is damaging our children and our schools.* Boston: Beacon Press.

Messick, (1989). Validity. In R. L. Linn (Ed.), *Educational measurement, 3rd ed.* (pp.13-103). Washington DC: American Council on Education and National Council on Measurement in Education.

Moeller, A. J., & Reschke, C. (1993). A second look at grading and classroom performance: Report of a research study. *The Modern Language Journal*, *77*(2), 163-169.

Murray, R. T. (2005). *High-stakes testing: Coping with collateral damage.* Mahwah, NJ: Lawrence Earlbaum.

National Center for University Entrance Examination (2007). *Daigaku nyushi center shiken: shiken mondai hyokaiinkai hokokusho* (Annual report of the Center Examination.) Tokyo: National Center for University Entrance Examination.

National Commission of Excellence in Education. (1983). *A nation at risk.* Retrieved at June 21, 2009, from the World Wide Web: http://www.ed.gov/pubs/ NatAtRisk/index.html.

Nicholls, J. G., Cheung, P. C., Lauer, J., & Patashnick, M. (1989). Individual differences in academic motivation: Perceived ability, goals, beliefs, and values. *Learning and Individual Differences*, 1(*1*), 63-84.

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Cambridge, MA: Harvard Education Press.

Nye, B., Konstantopoulos, S., & Hedges, L. (2004). How large are teacher effects? *Educational Evaluation And Policy Analysis*, *26*, 237-257.

Paris, S. G., Lawton, T. A., & Turner, J. C. (1992). Reforming achievement testing to promote students' learning. In C. Collins, & J. N. Magnieri (Eds.), *Teaching thinking: An agenda for the twenty-first century* (pp. 223-241). Hillsdale, NJ: Lawrence Erlbaum.

Perlstein, L. (2007). *Tested: One American school struggles to make the grade*. New York: Holt.

Polanyi, M. (1958). *Personal knowledge: Towards a post-critical philosophy*. Chicago: the University of Chicago Press.

Reagan, T. (2005). *Non-western educational traditions: Indigenous approaches to educational thought and practice* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York: Free Press.

Rohlen, T. (1983). *Japan's high schools*. Berkeley: Center for Japanese Studies.

Rosenbsch, M. H. (2005). The No Child Left Behdin Act and teaching and learning languages in U.S. schools. *The Modern Language Journal, 89* (2), 250-261.

Shohamy, E. (2001). *The power of tests*. London: Longman.

Samuda, R. J. (1957). *Psychological testing of American minorities: Issues and consequences*. New York: Dodd, Mead & Company.

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

Stipek, D. (2002). *Motivation to learn: Integrating theory and practice* (4th ed.). Boston: Allyn and Bacon.

Tanaka, M. (2008). *A history of English language testing in Japan*. Hiroshima: Keisuisha.

Tessler, J. (2003, July 10). State board delays exit exam requirement. *San Luis Obispo County Tribune*, p. B5.

Wall, D., & Alderson, J. C. (1993). Examining washback: the Sri Lankan impact study. *Language Testing, 10(1)*, 41-69.

Wall, D. (1996). Introducing new tests into tranditonal systems: insights from general education and from innovation theory. *Language Testing*, *13*(3), 334-354.

Wall, D. (2005). *The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory*. Cambridge: Cambridge University Press.

Watanabe, Y. (2001). Does the university entrance examination motivate learners?: A case study of learner interviews. *Trans-Equator Exchanges*: *A Collection of Academic Papers in Honour of Professor David E. Ingram*. Faculty of Education and Human

Studies, Akita University, 100-110.
Watanabe, Y. (2004). Teacher factors mediating washback. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp.129-146). New Jersey: Lawrence Erlbaum.
Watanabe, Y. (2005). *Eigokyouiku ni okeru mokuhyo junkyo hyouka (zettaihyouka) no dokizuke kokano kensho*. (Exploring the effectiveness of criterion-referenced assessment on motivation. Grants-in Aid Scientific Research, Scientific Research (C) [15520346].
Watanabe, Y. (2007). Assessment and organization: the role of assessment in teaching English as a second or foreign language. A paper delivered at ELT Career and Professional Development Conference, The Yomiuri Shimbun Building.
Watanabe, Y. (2009). Daigaku nyushi center shaken no henko to shakaiteki eikyo nit suite – yoriyoi kyoiku koka wo motarasutameni (Changes in the national center examination and its social consequences – To produce a better effect). *Daigaku nyushi forum*, *31*, 39-45.
White, M. (1988). *The Japanese educational challenge: A commitment to children*. Clearwater: Touchstone books.
Zeng, K. (1999). *Dragon gate: Competitive examinations and their consequences.* London: Cassel.