Accountability and External Testing Agencies

EDWARD SARICH
Shizuoka University, Japan

Bio Data:
Edward Sarich has been working in the field of language education for more than 15 years. He taught junior and senior high school for 7 years in Hamamatsu Japan. While completing an MA in Applied Linguistics from the University of Birmingham, in 2010, Edward began working as a language instructor at Shizuoka University. He is especially interested in issues concerning language pedagogy in Japan, particularly regarding language planning policy, standardized testing, evaluation, and communicative language teaching.

Abstract
Standardized testing is ubiquitous in Japan. Inexpensive and easily mass distributed, their use has been encouraged at every level of the education system. Over the past thirty years, external testing agencies have been increasingly relied upon to make standardized tests for use as benchmarks in the education system and in the private sector. However, while great trust has been placed in these agencies that create these tests, many of them operate with very little supervision. This article will review the practices of some of the commonly used external testing agencies in Japan and discuss how greater accountability from these agencies might not only improve test validity, but make them more useful for score users and test takers.
Keywords: language, testing, standardized tests, external agencies, TOEIC, EIKEN

Introduction

The use of standardized tests in the evaluation of language proficiency is a much debated topic. Although in general great faith has been placed in them as objective and consistent measures of assessment, they have recently faced mounting criticism due to the negative impact that they can exert on language education. In Japan,

however, standardized language testing has become increasingly commonplace. Policy planners have begun recommending their use as benchmarks in the secondary education system, and, in the private sector, standardized language test scores are more and more being linked with promotion and advancement. However, despite their popularity, little is actually known about how these tests are made. One of the hallmarks of standardized tests is that they are produced not by teachers but by external testing agencies. Although this practice brings with it many advantages, mostly by way of offering objectivity and reliability, it is not without its issues. This article will explore the use of standardized tests throughout Japan, examining why they are commonly relied upon, how they have .exerted an influence on English education in Japan, and some recommendations for how the use of standardized tests in Japan might be more effectively utilized.

What is a Standardized Test?

Bachman (1990, p. 74) defines standardized tests as those that are made from fixed content, that are administered according to uniform procedures, and whose validity and reliability are thoroughly researched. However, there are other similarities as well. First, they are mostly norm-referenced, meaning that unlike classroom tests which measure student progress against a predetermined set of criteria, standardized tests only provide an assessment of the test takers' general proficiency as compared with others who take the same test (Hughes, 2003, p.19). Second, standardized tests are constructed and scored by external testing agencies. Third, the test questions are made up of discrete-point items, which assess isolated language components that are intended to be inferential of language ability. Fourth, the most commonly used format for these tests is multiple-choice. Like any measure of assessment, the qualities specific to standardized tests bring with them many advantages and disadvantages. However, it has been noted that the advantages that standardized tests offer are more likely more beneficial for test makers and score users than they are for test takers (McNamara & Roever, 2006, p.136).

Positive Aspects of Standardized Tests

It is undeniable that standardized tests have been beneficial in many ways. They are inexpensive and easily mass produced, which allows them to serve as a unifying benchmark for students from a wide range of language instruction programs. They can serve as a short-term motivating goal for students, allowing them to compare their scores against age-based recommended benchmarks. They have also proven very convenient for score users, producing unambiguous and replicable results, the statistics of which not only say much about individual against the average but about

how groups of test takers are faring against the norm (McNamara and Roever, 2006, p. 136). Policy planners can look at scores to confirm or deny the effectiveness of their policy initiatives. School administrations and boards of education have a means by which they might gauge the effectiveness of their language programs. And students have a way of comparing their ranking against the regional and national mean.

However, while these tests might offer an extremely convenient measure of language proficiency, they do not offer the most accurate measure, and not enough has been said about their limitations. In many ways, the aspects of test design that are attractive to the more powerful stakeholders can also negatively affect the test's validity, the quality that is most relevant to test takers (Hughes, 2003, p. 50).

Negative Aspects of Standardized Tests

While the common format of standardized tests offers gains in reliability, test validity, or how accurately a test measures the constructs it is designed to measure, can be affected in several ways. First, although multiple-choice format allows perfect marker reliability, circling the correct answer from a list of choices is not a real task and therefore "there is considerable doubt" about its validity as a measure of language ability (Weir, 1990: 47).

Second standardized tests contain discrete-point questions, or indirect questions designed to be inferential of language ability without being authentic instances of language use. Generalizations made about language ability from discrete-point items, as seen in standardized testing, however, must be made with caution because they "give a limited picture of the knowledge and proficiency of the person tested" (Spolsky, 1985, p.182). Moreover, the convenience by which standardized tests reduce the complex and multi-faceted concept of language ability into a single score can encourage test scores to be used in ways for which they were not intended, for example, as direct evidence of writing or communicative ability, or as the justification for a school language curriculum. It is when these 'unintended' uses of test scores become commonplace that they become a threat to the test's validity.

Third, standardized language tests are norm-referenced. They do not comment on how well one knows any specific material, they can only tell us how test takers do in comparison with other test takers (Hughes, 2003, p.22). Bachman and Palmer discuss how construct validity must be established through a comparison with a "specific domain of generalization" (1996, p.21). However, as standardized tests are not based on any criteria, there can be no way for the test taker to anticipate the test content, particularly the target language usage. Thus, it is

difficult to discern what a high score on these tests is an indication of, other than of the test taker's performance on tests of a similar design. Test validity is easier to establish with criterion-based tests such as those based on a school curriculum, as the target language usage is usually smaller in scope (Hughes, 2003, p.55).

One final drawback of standardized testing is that it has been known to involve traits that are unrelated to language. The notion of "variable irrelevance" refers to how the possession of skills not associated with language ability can allow some test-takers to do better than others (Messick, 1989, p.7). Intelligence, memory, test-taking skills, confidence, familiarity with the test-design and cultural understanding are all qualities that can offer test takers an advantage (McNamara, 1990, p.19). Standardized test scores are unable to discern the degree to which these irrelevant skills are present in the test taker, limiting their ability to be inferential of language proficiency. Moreover, there is an issue of fairness to consider. A common concern among language teachers is for those students who can perform equally with other students in class activities, but for whatever reason, are not good at taking standardized tests. Over the course of their education from junior high school through to university, these students are severely disadvantaged in the standardized test-heavy language programs that they are required to take. Fairness in testing is thought to be an important contributor to test validity (McNamara and Roever, 2006, p.17-18).

Consequences beyond the Measure of Language Testing

Modern theories of validity require testers to look beyond the relationship between test items and language ability to examine the consequences that tests have on society, education and the test takers (Messick, 1989, p. 6). In general, consequences have been divided into two groups. Consequences pertaining to how tests exert an influence on studying and teaching in the classroom have been referred to as "washback", while consequences related to the effect that testing has on individuals and on society as a whole have been referred to as "impact" (McNamara and Roever, 2006, p. 235).

Washback

Washback refers to how tests influence a teaching context. If classroom study focuses on expected test content, there is the concern that less attention will be paid to areas that the tests do not cover. It is believed that high stakes measures of assessment are particularly affected by washback because teachers feel pressured by school administrations to teach test content, and students have to focus on what is covered

on these tests because the results can significantly affect their future (Cheng, 2008, p. 349).

Hughes discusses how tests can impact language education both positively and negatively (2003, p.1). Tests that reward a balance of all areas of language ability can exert a positive impact on language study because test takers must prepare comprehensively. However, standardized test content usually prioritizes the structural aspects of language ability, such as syntax and knowledge of vocabulary, over the functional aspects, such as communicative or strategic competence. It is for this reason that many have voiced concerns that the institutionalized use of standardized testing in Japan, in particular the entrance exam to university, has resulted in teachers disproportionately emphasizing grammar and reading comprehension over communicative language teaching in their classroom curriculums (Butler and Iino, 2005, p. 32; McNamara and Roever, 2006, p. 206; Okuno, 2007, p.148).

Impact

Although the use of standardized testing has been shown to produce some increases in classroom test scores, they quickly levelled off and were offset by several negative consequences (Hargreaves and Shirley, 2009, p.11). Scores from these tests tend to place the burden of responsibility squarely on the shoulders of teachers, pressuring them to increase the amount of test-coaching and "practice testing", training students by having them take mock exams (Saito, 2006, p.103). These practices lead to the narrowing of school curricula and have been shown to greatly decrease teacher satisfaction (Amrein and Berliner, 2002, p.33). Students are thought to be greatly impacted as well (McNamara and Roever, 2006, p. 206). The strain of preparing for high stakes exams has been shown to increase a negative association with test content and even increase dropout rates (Amrein and Berliner, 2002, p.33). In addition, families are burdened with the increased costs of supplemental exam preparation at cram schools, which also coach students through the use of mock exams (Saito, 2006, p.106). Relying on cram schools to help prepare students to take these tests impacts language education in another area as well. For financial reasons, "collusion" between university admissions departments and cram schools that offer preparatory classes for standardized tests is thought to have acted as an agent against the modernization of testing in Japan (Ross, 2009, p. 6).

External Testing Agencies

The use of external testing agencies to supply the means for benchmarking raises some noteworthy issues. It is believed that "the relationship of these externally

imposed standardized tests to teacher assessment within a curriculum is a matter of heated debate in virtually every setting in which such an arrangement has been established", and that the questionable validity of these high stakes tests is particularly "controversial" for L2 users (McNamara and Roever, 2006, p. 227). One reason is that the concerns of external agencies can differ from the needs of educators and from what is in the best interest of the test takers. Another is that the separation between assessment body and score user can result in a discrepancy between how scores were intended to be used and how they are actually used.

Standardized Tests as High Stakes Tests

The use of standardized language tests as high stakes means for social advancement is not unique to Japan. Researchers across Asia (Ross, 2009, p. 9-12) and in other countries have argued against their use.

Hargreaves and Shirley (2009) commented on their effect on the Canadian education system,

> … when high stakes events such as graduation depend on single or simple measures of performance that are linked to political targets, are cause for possible sanction, and are made public, the chance that they will distort the learning process are high. (p. 103)

Similarly, Linn (2000), an American educational psychologist, has stated,

> As someone who has spent his entire career doing research, writing, and thinking about educational testing and assessment issues, I would like to conclude by summarizing a compelling case showing that major uses of tests for student and school accountability during the past 50 years have improved education and student learning in dramatic ways. Unfortunately, that is not my conclusion…. Assessment systems that are useful monitors lose much of their dependability and credibility for that purpose when high stakes are attached to them. The unintended negative effects of the high stakes accountability uses often outweigh the intended positive effects. (p. 14)

In spite of the fact that experienced educators from around the world are speaking out against high stakes standardized language testing, Japan seems to be increasing its reliance on them.

To summarize briefly, stakeholders with power, such as policy planners, testing agencies and school administrations require measures that reduce language proficiency to a clearly definable score that can dependably be reproduced, resulting in language tests that prioritize consistency over accuracy, or, in other words, reliability over validity. The pervasive use of these tests has negatively impacted language education in Japan because it does not foster a balanced set of language

skills and because test scores are being used for purposes beyond those for which they were intended. How each test has specifically affected language education will be discussed in the next section.

External Agencies and their Standardized Language Tests

The Educational Testing Service and the Institute for International Business Communication (TOEIC). The Test of English for International Communication (TOEIC) was developed in 1979 by the Educational Testing Service (ETS) at the behest of the Japanese Ministry of Education. At first, TOEIC was rejected, as it was thought to threaten Eiken test favoured by the Ministry. However, Yaeji Watanabe, a former Ministry official, was able to secure support for its development, and TOEIC was eventually popularly received as a test of 'business' English. Mr Watanabe went on to become the Chairman of the Institute for International Business Communication (IIBC), the non-profit organization responsible for administering the test within Japan.

Both the creators and the distributors of TOEIC have been linked with practices that call into question their status as non-profit organizations. ETS has been accused of overcompensating its officers and for reaping excessive profits from the sale of their tests (Americans for Educational Testing Reform, 2007). Similarly, Mr Watanabe's long-standing tenure as Chairman of the IIBC has not been without controversy. He has been accused of hiring several former policy officials from the Ministry of International Trade and Industry, not coincidentally the government department that granted the IIBC permission to distribute TOEIC in Japan (McCrostie, 2010, p.3). Moreover, Mr Watanabe was criticised for appointing his girlfriend's son to an executive position within the IIBC, an act that was so strongly resisted that Watanabe was only able to do so after having fired half the board of directors (McCrostie, 2010, p.4). In 2009, Mr Watanabe was forced to resign after it was reported in the media that the IIBC had over 1.7 billion yen in unreported savings, prompting a warning from the Ministry of Trade and Industry (McCrostie, 2010, p.7). Before his departure, Mr Watanabe appointed his girlfriend's son, Murofushi Takayuki, as the current Chairman of the IIBC. As a test produced and distributed by non-profit organizations, it is not unreasonable to expect that all profits generated from the sale of TOEIC would be reinvested into test development and research. Many of the allegations of cronyism and the squandering of resources, however, suggest not only that these organizations are not closely scrutinized, but that profitability exerts a greater impact on test construction than they would have us believe.

There is strong evidence that TOEIC has become the benchmark for language assessment in Japanese society, evident in that many universities now use TOEIC to stream incoming students, and that MEXT has recommended that TOEIC scores be used in the hiring of new language teachers (MEXT, 2010). Furthermore, in the private sector TOEIC has become the "de facto" measure of the assessment of language ability used by companies across Japan (Chapman and Newfields, 2008, p.32).

Opinions of TOEIC are varied. While the ETS publishes validity research on its homepage (albeit not independent research), and there have been efforts made toward increasing the tests' overall validity, some criticisms remain (Chapman & Newfields, 2008, p.32). One concern is that TOEIC is being used as a test of general language proficiency even though it is primarily a test of reading comprehension and listening. In response to increasing pressure to more accurately assess spoken and written skills, in 2006 ETS started producing a separate test "TOEIC for Speaking and Writing." However, in 2009, while almost two million Japanese took the original TOEIC test, only 6,200 test takers sat for the TOEIC speaking and writing test (ETS, 2010), evidence that the TOEIC test based on reading comprehension is used as the benchmark for language proficiency. In this regard, TOEIC has been criticised for offering a separate test of writing and speaking instead of including a special section in the popularized version (Chapman and Newfields, 2008, p.35).

One other criticism has been levelled at how scores for TOEIC have been used. Although ETS maintains that "no single factor should be used as the sole criterion for any important educational decision" (ETS, 2010), there is a concern that companies and educational institutions are using TOEIC scores as the sole indicator of language proficiency with a test that does not adequately assess the language skills necessary for business (Childs, 1995, p.76; Hirai, 2009, p.8).

Society for Testing English Proficiency. The Society for Testing English Proficiency (STEP) was established in 1963 "for the purpose of popularizing and improving English in Japan" (STEP, 2010). Through consultation with the Social Education Council to the Ministry of Education, STEP developed and administered the "Jitsuyō Ginō Eigo Kentei", known as Eiken. The test was designed to offer a cost-effective and accurate test of English language proficiency.

Currently, Eiken is very highly regarded in Japan and is commonly utilized as a benchmark in educational institutions and in the private sector (Miura & Beglar, 2002, p.108). Every year, approximately 2.3 million people take Eiken, making it one of the most widely used standardized tests in Japan. The large majority of test takers are students in junior and senior high school, in order to measure their personal

progress, but test scores have also been used by institutions as a general measure of language proficiency. As an alternative to taking private entrance examinations, some high schools and universities now allow students to submit Eiken scores as proof of their language proficiency. In addition, one-third of prefectural boards of education require Eiken scores in the consideration of hiring new language teachers, and many companies now encourage their employees to take Eiken, relating level certification with employment and advancement (Miura and Beglar, 2002, p.108).

STEP makes a genuine attempt to provide a test that is relevant for test takers. It is a non-profit organization whose finances are overseen by the Ministry of Science Education and Technology, insisting that all profits are reinvested in strengthening the quality of its test. STEP's homepage contains an abundance of validity research as well as information that demonstrates a thorough understanding of current testing theory and a commitment to developing a test that creates beneficial washback for its test takers (STEP, 2010). However, while the research listed on the STEP homepage is convincing, more independent research would be helpful in support of its claims.

Although Eiken meets the strict reliability requirements of a high stakes standardized test, there are a few design qualities that set it apart from other similar tests. First, it offers levelled testing, and each test level is aimed at a different year of educational development, from junior high school through to university. The vocabulary for each level is linked to vocabulary for the associated year of school, based on interviews with teachers and students, which allows it to act both as a test of general linguistic proficiency, and to some degree as a measure of general progress within a school curriculum (STEP, 2010).

Eiken also offers subjectively assessed questions at higher levels in order to increase validity. According to STEP, at lower levels, when the stakes are also lower, a priority is placed on availability and affordability, and as the levels increase in difficulty and become more important for the future of the test taker, subjective test measures are thought to increase test validity (STEP, 2010). Once again, however, more independent research would be helpful to verify that these actions are beneficial to the test taker.

Although Eiken is likely the most accurate indicator of language competence of all the standardized tests discussed in this article, it has still not been established that Eiken alone is sufficiently comprehensive to merit its use as the sole indicator of language ability. While it may well serve its intended use as a measure of individual progress, other uses, such as a measure of communicative ability tied to advancement in a company, or as a general proficiency test for entrance in a

university, are more worrisome because they are not uses for which the test was designed.

National Center for University Entrance Examinations (Senta Shiken). The Senta Shiken, otherwise known as the Central University Examination, is a high stakes test constructed by The National Centre for University Entrance Examinations (NCUEE), an organization directly overseen by the Ministry of Science Education and Technology (MEXT). All students desiring to enter any public (and some private) universities in Japan are required to take this test. It is said to be "the only nationwide national standardized exam and the most heavily weighted" (Guest, 2008, p.88). As such, students spend considerable classroom time preparing for this test (Sakui, 2004, p.159), and many attend private cram schools for supplementary study (Saito, 2006, p.106). It has been reported that scores from this test alone are the sole criteria for the determination of admittance (Saito, 2006, p.102), although recent declining population rates have been causing universities to loosen their heretofore strict policies for gaining entrance (Guest, 2008, p.86).

While it may be true that the test has improved considerably in terms of its validity, and that considerations beyond test scores are gradually making their way into consideration for admission into university, the Senta Shiken still "stands as a bellwether of national policy regarding English pedagogical content" (Guest, 2008, p.88), and exerts a monumental impact on English education in Japan. Several criticisms of the Senta Shiken have been voiced.

First, MEXT has been widely criticised for making communicative language ability the focus of its policy initiatives while at the same time requiring all prospective university students to take a test that prioritizes reading comprehension (Crooks, 2001, p.36; Lamie and Lambert, 2004, p.92). Although it has undergone changes designed to make it more relevant to test takers, particularly in the addition of a listening component in 2006, even those who argue in favour of the Senta Shiken admit that it alone is not an adequate assessment of communicative ability (Guest, 2008, p.88). Many believe that language education in Japan will not sufficiently change until the severe impact of the Senta Shiken is addressed, either by eliminating it altogether or by altering it in such a way that all areas of language ability are equally evaluated (Butler & Iino, 2005, p.32; Okuno, 2007, p.148; Sakui, 2004, p.157; Samimi & Kobayashi, 2004, p.248). It is also felt that the overbearing burden of having to prepare for entrance examinations restricts innovative teaching approaches that would be required to introduce significant improvement in communicative English skills (Lokon, 2006, p. 9). Simply put, if the end goal of English education in Japan is performance on a test which does not include

questions that adequately assess writing, speaking or even listening skills, it is difficult to expect that students will devote their time to developing those skills, nor will teachers develop language curriculums that prioritize them.

Benesse (GTEC). The Global Test of English Proficiency for Students (GTEC) is a standardized test created by Benesse Corporation for the purpose of measuring the language proficiency of Japanese students in secondary education (Benesse, 2010). The ninety-minute exam contains sections on reading comprehension, listening and writing. While the writing section is subjectively assessed, no information is available regarding the standards by which evaluators score the results, which could affect reliability.

The most commonly used GTEC exams are the CORE test, aimed at junior high school students, and the BASIC test, given at high school. In 2009 more than 400,000 test takers in over 800 schools around Japan took the CORE or BASIC test. Although research into test validity for Benesse is scant, one published study documents the BASIC (high school) tests' consistency with the Common European Framework of Reference for Languages, a widely accepted scale for language proficiency (Negishi, 2006, p. 99). It should be noted, however, that this study does not offer support for the validity of the CORE test.

The catchphrase on the homepage of GTEC, "入試で必要な英語がもちろん、社会、留学先でも使える英語力を育むために", translates as "Not only for the purpose of entrance examinations, but to strengthen English necessary in society and for study abroad…" (Benesse, 2010), implying a fundamental misunderstanding of test use. First, as a norm-referenced test not based on any school curriculum, it is difficult to discern how such a test can positively impact English language study in Japan. Second, as standardized tests do not provide accurate assessments of performative aspects of language ability, the claim that they are able to prepare students for social language use is unsupported. Third, there is a strong implication that one of the main uses of GTEC is as an indicator of how well test takers will perform on future high stakes tests that they will have to face, such as high school entrance examinations, the Senta Shiken and TOEIC, reinforcing the perception that higher standardized test scores, rather than the development of real and practical language skills, should be the primary motivation for studying English.

A spokesperson for Benesse commented that GTEC was intended to provide a measure to help students reflect on their personal language development and for teachers to reflect on their teaching methods (Personal Communication GTEC, 2010). However, standardized tests are ill-equipped to comment on teacher performance or the efficacy of school curriculums (Popham, 2001: 27-28). These claims point to the

potential for GTEC to be used in ways for which it was not intended, such as a measure by which students are streamed into different English levels or to confirm or deny the merits of a school language program. Moreover, as Benesse does not make its own research regarding the establishment of validity, reliability, or impact available to the public, there is no way for outside parties to substantiate these claims.

One other area of concern is with the CORE Test used in junior high school. It has been documented that standardized tests are not only least accurate at lower levels but can prove particularly intimidating for students who are beginning their formal education in English (Amrein and Berliner, 2002: 55). This is particularly evident in the Benesse CORE test because as all years of junior high school students take the same test, the questions can prove frustrating and confusing for younger test takers.

Locally Made Standardized Tests

One area that is particularly concerning regarding standardized testing in Japan is the use of what can be called non-standard standardized tests. Private high schools, universities, and even some companies often rely on their own locally made standardized tests in order to make the determination of the language proficiency of prospective entrants. It is very likely that many of these organizations have not thoroughly researched their methods of test construction, nor do they undertake the rigorous procedures of statistical analysis that are currently being undertaken by the formal testing agencies mentioned in this article. The result is that these non-standard tests lose many of the benefits that regular standardized tests have while maintaining all of their limitations. In the same way that formal testing agencies should be required to, entrance examinations made by local boards of education, by language departments within schools or within private companies need to be available to outside parties for scrutiny.

Results

A summary table of pertinent practices of the external testing agencies that are being discussed is included below.

Table 1.1

Agency	Non-Profit?	Reports research?	Former tests published?	Negative Washback	Score Misuse
ETS (TOEIC)	?*	Yes	No	McCrostie, 2006: 32.	Butler & Iino, 2005: 31; Gottlieb, 2005: 69; Childs, 1995: 76.
STEP (Eiken)	Yes	Yes	Yes		Butler & Iino, 2005: 31.

High school entrance examinations	Yes	No	No	McNamara & Roever, 2006: 206.	
NCUEE (Senta Shiken)	Yes	No	No	Samimi & Kobayashi, 2004: 248; Lokon, 2006: 9.	Saito 2006; 102.
Benesse (GTEC)	No	No	No	Amrein & Berliner, 2002: 55.	Popham, 2001: 27-28.

*Both ETS, the organization which makes TOEIC, and IIBC, the organization which distributes TOEIC in Japan, have conducted activities which call into question their status as non-profit organizations.

ETS (TOEIC) and STEP (Eiken) make their validity and reliability research available to the public, but NCUEE (Senta Shiken) and Benesse (GTEC) do not. Moreover, of all the external testing agencies, only STEP publishes its former tests.

Another noteworthy finding is that, while STEP and the NCUEE are clearly NPOs, the status of ETS as an NPO is questionable, and Benesse is not an NPO. The implications of profitability and of how each agency publishes former tests and research will be examined in the discussion.

Although there is some research available regarding the effect that high school entrance examinations have on language education, there is very little available about how these tests are constructed and evaluated. The veil of secrecy with which boards of education and local schools construct these tests makes it extremely difficult to verify that they are reliable and valid tests of language proficiency.

Discussion

Standardized language tests have been shown to be widely used in Japan not because they are the most accurate measures of language proficiency, but rather because they serve the uses of more powerful stakeholders (McNamara and Roever, 2006, p. 209). They fulfil the requirements of a structuralist education system that promotes diligence and competence over performance (Samimi and Kobayashi, 2004, p. 250); they offer teachers clearly defined evidence that they are providing for the well-being of their students' futures; they provide data for schools to confirm or refute the merits of their English curriculums; and they offer concrete and quantifiable feedback for those in the government ministries to justify their policy initiatives (McNamara and Roever, 2006, p. 204; Solorzano, 2008, p. 314). However, for the test taker, standardized tests also suffer from serious issues of accuracy. First, they often reward skills that are unrelated to language ability, offering some an unfair advantage (Haladyna and Downing, 2004, p. 18). Second, the ways in which

test scores are used and the consequences that they produce in learning, although important factors in the determination of validity (Bachman and Palmer, 1996, p. 34), are often ignored because due to their subjective nature they are not easily incorporated into traditional validity research (Bachman, 2005, p. 6-7). Nevertheless, difficulty is not a justification for inaction. More awareness by stakeholders of the ethical and proper use of scores, and the creation of tests that can accurately assess language proficiency are clearly necessary.

Accountability

As the stakeholders that use tests are not in direct control of test construction, a means through which outside parties can have a clear understanding of how the tests are constructed is of obvious importance. There are two areas in which greater accountability should be expected of external agencies. A clear distinction between profit and non-profit status is relevant because for-profit agencies may feel greater responsibility to shareholders than they do to the other test stakeholders. Financial concerns may limit the use of more costly subjective testing procedures such as interviews and essay questions. In addition, research into validity and reliability may be foregone to increase profitability.

The other issue which concerns accountability is transparency. External agencies need to publicly, not privately, report their reliability and validity research so that the other stakeholders can verify that their research is sound. Moreover, agencies should publish former tests, not only so that test takers can use them in preparation for future tests, but so that other stakeholders have the opportunity to conduct independent research. It is noteworthy that only STEP (Eiken) satisfies the three stated concerns with regards to accountability (Table 1.1). On the other hand, Benesse satisfies none of them. The secrecy with which Benesse, a for-profit agency that is accountable to no one, constructs tests and conducts research makes their validity claims very difficult to verify. Furthermore, the makers and the distributers of TOEIC have been accused of using profits for purposes other than for improving test quality. Requiring all external test agencies to be transparent in how their test profits are spent and in how their tests are constructed would greatly enhance their claims of validity and provide them with built-in incentives to make improvements.

Test ethics

One way that external testing agencies can contend with some of the ethical issues surrounding testing is to ally themselves with organizations that exist on behalf of test takers. The Japan Language Testing Association (JLTA) offers information about testing research and theory and also authored the Code of Good Testing Practice,

which clearly outlines the responsibilities of test makers and score users, in short, saying that tests should be administered consistently, that tests makers must prove that their tests are accurate measures of the constructs that they were designed for, and that score users should recognize the limits of test results and should not misuse them (JLTA, 2010). Although organizations such as the JLTA have no enforcing authority in that participation is voluntary and there are no repercussions for not adhering to their codes of conduct, they are nonetheless important because they increase awareness of test issues and "raise the standard of professionalism" among test making bodies and scores users (McNamara and Roever, 2003, p. 139). Moreover, as ethical considerations are increasingly thought to be closely associated with the establishment of validity, belonging to these organizations and adhering to the Code is another way in which external agencies can make better tests. The JLTA currently has 190 individual members and 13 institutional members. Of the agencies discussed in this dissertation, only STEP and ETS are institutional members. Clearly, in light of the issues concerning the rampant use of standardized high stakes testing on-going in Japan, the JLTA needs to widen its membership, not only among external agencies but also among those who use the test scores.

Recommendations

Based on the points reviewed in the discussion, the following are suggestions that might be made toward the improvement of language testing in Japan.

1. The number of standardized tests that students in junior and senior high school take should be significantly reduced.
2. Greater awareness among teachers and administrators about the limitations of standardized tests is necessary to see that test scores are not misused.
3. The use of standardized test scores as the sole measure of language proficiency should be discouraged.
4. Greater accountability should be expected of external agencies that create tests used in the education system. In addition to complete financial transparency, all external agencies should be required to publish their research and former tests so that accuracy can be independently verified.
5. Private sector companies should be encouraged to stop linking standardized tests scores with promotion and advancement. For positions in which proficiency in English is anticipated, candidates should be required to take a criterion-based assessment centred on expected language use.

6. The creation of non-standard standardized entrance examinations at local boards of education, private schools and companies should be discouraged.

Conclusion

Evidence has been presented showing that standardized testing in Japan is being used in precisely the same circumstances as those that other countries have warned against. Many of these tests do not positively contribute to language learning because they do not adequately assess a balance of skills, not enough about the limitations and proper uses of such tests is known by the people who make interpretations from them, and external testing agencies are not adequately held accountable for the tests that they produce. Beyond this, there are other concerns, albeit ones for which empirical evidence is hard to come by. The tendency of standardized tests to advantage test takers with skills unrelated to language, and of these tests to be used not as measures of language *proficiency* but of language *potential*, is ethically questionable, especially when used within the formal education system. It is also becoming increasingly likely that as a result of the excessive use of standardized tests, high scores, rather than the desire to communicate with the outside world, has become the primary impetus for language study in Japan. If this is indeed the case, one wonders how long it will serve as a sufficient source of motivation after the desired test scores have been achieved.

Many of these practices suggest that stakeholders with greater power need to closely examine the rationale behind using standardized tests. Greater accountability by external testing agencies would do much to improve the validity of their tests. However, the secrecy with which many of these agencies have been allowed to operate has in itself acted against their own best interest.

Finally, while I have attempted to provide several answers regarding the impact of high stakes standardized language testing in Japan, it is also my sincere hope that the reader will be left, as I am, to wonder how tests known to have a high rate of variable irrelevance can be thought of as fair, how tests that do not concern themselves with the social aspect of language can be deemed valid, and how the use of tests with no regard for their social consequences can be considered ethical.

References

Americans for Educational Testing Reform. (2007). *AETR Report Card*. Retrieved from: http://www.aetr.org/ets.php

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Educational Policy Analysis Archives, 10*(18), 32-38.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly,* 2, 1-34.

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

Benesse. (2010). *GTEC for students*. Retrieved from http://gtec.for-students.jp/about/about.htm

Butler, Y., & Iino, M. (2005). Current Japanese reforms in English language education: The 2003 "ACTION PLAN". *Language Policy*, 4, 25-45.

Chapman, M., & Newfields, T. (2008). *JALT Testing & Evaluation SIG Newsletter, 12*(2), 32-37.

Cheng, L. (2008). Washback, impact, and consequences. *Language Testing and Assessment,* 7, 349-364.

Childs, M. (1995). Good and bad uses of TOEIC by Japanese companies. In J. Brown & S. Yamashita (Eds.), *Language testing in Japan* (pp. 66-75). Tokyo, Japan: JALT.

Crooks, A. (2001). Professional development and the JET Programme: Insights and solutions based on the Sendai City Programme. *JALT Journal, 23*(1), 31-46.

ETS. (2010). *About the TOEIC Test*. Retrieved from http://www.ets.org/toeic

Guest, M. (2008). *A comparative analysis of the Japanese University Entrance Senta Shiken based on a 25-year gap. JALT Journal, 30(1) 85-104.*

Haladyna, T., & Downing, M. (2004). Construct irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23,* 17-27.

Hargreaves, A., & Shirley, D. (2009). *The fourth way: The inspiring future for educational change*. Thousand Oaks, California: Corwin Press.

Hirai, M. (2009, October 5). Engineers must have English skills to succeed. *Japan Times,* p. 8.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

JLTA. (2010). *The JLTA code of good testing practice*. Retrieved from http://www.avis.ne.jp/~youichi/COP.html

Kobayashi, Y. (2000). *Japanese social influences on academic high school students' attitudes toward long-term English learning*. PhD thesis, University of Toronto, Canada.

Koike, I., & Tanaka, M. (1995). English in foreign language education policy in Japan: Toward the twenty-first century. *World Englishes, 14*(1), 13-25.

Lamie, J., & Lambert, S. (2004). Developing the communicative approach in Japan: An investigation into the Japan exchange and teaching programme. *Progress in Education, 13*(4), 83-100.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Littlewood, W. (2007). Communicative and task-based teaching in East Asian classrooms. *Language Teacher, 40*, 243-249.

Lokon, E. (2006). Will the new Center test make English language education more communicative in Japanese high schools? *The Language Teacher*,11(29), 7-12.

McCrostie, J. (2006). Why are universities abandoning English teaching for TOEIC training? *Oncue, 14*(2) 30-32.

McCrostie, J. (2010). The TOEIC in Japan: A scandal made in heaven. *JALT Testing and Evaluation Sig Newsletter, 14*(1), 2-10.

McNamara, T. (1990). *Language testing.* Oxford: Oxford University Press.

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

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

MEXT. (2010). Regarding the establishment of an action plan to cultivate "Japanese with English abilities." Retrieved from: http://www.mext.go.jp/english/

Miura, T., & Beglar, D. (2002). The Eiken vocabulary section: Analysis and recommendations for change. *JALT Journal, 24*(2), 107-129.

Negishi, M. (2006). How much do we have in common with a European framework of reference? The applicability of the CEFR to an IRT-based English proficiency test. In Yoshitomi, A., Umino, T., & Negishi, M. (Eds.), *Readings in second language pedagogy and second language acquisition*. (pp. 83-100). Tokyo: John Benjamins Publishing Co.

Popham, J. (2001). Uses and misuses of standardized tests. *NASSP Bulletin, 85,* 24-31.

Okuno, H. (2007). A critical discussion on the action plan to cultivate "Japanese with English abilities". *The Journal of Asia TEFL*, 4(4) 133-158.

Reesor, M. (2002). The bear and the honeycomb: A history of Japanese English language policy. *NUCB Journal of Language, Culture, and Communication, 4*(1), 41-52.

Reesor, M. (2003). Japanese attitudes to English: Towards and explanation of poor performance. *NUCB Journal of language, Culture, and Communication, 4*(1), 57-65.

Ross, S. (2009). Language planning policy in Asia. *Language Testing, 25*(1), 5-13.

Saito, Y. (2006). Consequences of high stakes testing on the family and schools in Japan. *Journal of Educational Policy, 3*(1), 101-112.

Sakui, K. (2004). Wearing two pairs of shoes: Language teaching in Japan. *ELT Journal, 58*(2), 155-163.

Samimi, K., & Kobayashi, C. (2004). Toward the development of intercultural communicative competence: Theoretical and pedagogical implications for Japanese English teachers. *JALT Journal, 26*(2), 245-261.

Sasaki, M. (2008). The 150-year history of teaching English language assessment in Japanese education. *Language Testing, 25*(1), 63-83.

Solorzano, R. (2008). High stakes testing: Issues, implications, and remedies for English language learners. *Review of Educational Research, 78*(2), 260-329.

Spolsky, B. (1985). What does it mean to know how to use a language? An essay on the theoretical basis of language testing. *Language Testing, 2*(2), 180-91.

STEP (2010). *Eiken Test in Practical English Proficiency*. Retrieved from http://stepeiken.org/

Weir, C. J. (1990). *Communicative language testing*. New York: Prentice Hall.