A general overview of English tests administered in Japan[1]

Michihiro Hirai
mjhirai@beige.ocn.ne.jp
Kanagawa University

Abstract

As a means of communication, language reflects all aspects of human thoughts and activities; hence there are countless approaches to, and forms of, language tests as a means of assessing communication skills, depending on the purpose, domain, and other factors. I discussed various features and characteristics of the major English tests administered in Japan by genre such as general, academic, and purpose-specific. Just as it is important to understand what knowledge and skills each test is designed to evaluate and how, so is it also important first to realize the meaning of the test to the organization using it – be it a school or an enterprise – and then to ensure that the test serves that organization's objectives and priorities.

Keywords: testing and evaluation, correlation

Having spent more than 30 years as an engineer and a subsequent 10 years or so as a college English teacher, I have developed a grass-roots framework for discussing language education and testing, which combines the users', teachers' and learners' viewpoints. From this trilateral perspective, I have been observing the English test landscape in Japan by personally taking more than 50 English qualification tests open to the public during the past 40 or so years. In this three-part presentation, I first reviewed various parameters of language testing, then gave a comparative overview of the major English tests, and finally discussed some common misperceptions and misuses of these tests, primarily from the standpoint of industry. Considering the readership of this publication, here I am going to focus on the second and third parts, skipping the first part that discussed what is actually tested and how.

Major English tests

More than 50 different English tests are currently administered in Japan (ELT Services Japan, 2017), depending on how they are counted. ELT Services Japan's website titled 英語教育ニュース (soon to be closed down) lists 63 English tests as of the beginning of 2017, including composite tests (i.e., a combination of English and non-language knowledge/skills such as typing, accounting, and export/import trading) and tests for children, while missing many translation tests and brand-new tests such as the TEAP, not to mention minor ones such as the Pitman^{TM} ESOL. For the purposes of this paper, I classify them into three broad categories: general, academic, and purpose-specific.

General English tests are those designed to cover a general range of use of the language and include Eiken (英検), Kokuren Eiken (国連英検), the Cambridge Main Suite (KET, PET, FCE, CAE, and CPE), IELTS^{TM} (General Training), the Pitman ESOL, and the G-TELP^{TM}. Most of them, except the G-TELP, which does not have speaking or writing components, test all the four skills as well as grammar knowledge and test speaking skills in a face-to-face (real-person) interview mode. Most of them employ different test forms for different levels (i.e., each level or grade has its own set of questions, problems, and/or tasks), while IELTS uses a single test form to cover the entire range of levels called bands (i.e., one set of questions, problems, and tasks covers all the levels or grades) (Note: In IELTS's speaking section, the interviewer tailors test questions to the individual candidate).

[1] This report is a summary of an oral presentation at the JALT Hokkaido Conference on October 2, 2016.

Academic English tests are those designed to evaluate how capable the candidate (test-taker) is of keeping up with study at universities in English-speaking countries, while the exact acceptance criteria are usually left to individual universities. These tests include the TOEFL® iBT®, IELTS (Academic), and the GRE® General, the TEAP (Test of English for Academic Purposes), and its computer version the TEAP CBT. Historically, the TOEFL, developed by Educational Testing Service (ETS®), was practically the only major English test specifically designed for academic purposes, particularly for admission to American graduate schools. In the late 1990s, IELTS was split into General Training and Academic, and then the GRE General was introduced. Each of these tests employs a single form to cover the entire range of competence.

Recently, in response to the Japanese government's initiative to internationalize the Japanese educational system, Professor Kensaku Yoshida of Sophia University has led an ad-hoc team in cooperation with the Eiken Foundation to develop the TEAP. It was officially introduced in 2014 with only listening and reading components but now covers all the four skills and is expected to be widely used in Japan as an alternative to the English component of the National Center Test for University Admissions (センター試験).

Purpose-specific tests are very diverse in nature but can be divided into four subcategories: business or workplace English, English for tourism, technical English, and translation tests. Business or workplace English tests are those designed to evaluate the candidate's linguistic competence in business situations or at workplace, with varying degrees of business flavor. These include the TOEIC, BULATS (Business Language Testing Service) English, GTEC, BETA, Nissho Business Eiken, and TOBiS.

The TOEIC was developed by ETS in 1979 at the request of some representatives of Japanese companies, who were not quite satisfied with the Eiken test, which focused on school English (McCrostie, 2010). Their original intent was apparently to develop a more practical English test reflecting the kind of English actually used in general society. Since a number of major Japanese enterprises jumped on this bandwagon, the TOEIC has become widespread during the past three and a half decades. The popularity of the TOEIC, however, is essentially limited to Japan, South Korea, and Taiwan. The British presence is stronger in the rest of Asia as well as in Europe. In the meantime, ETS, which is a respectable organization with solid and professional research staff, has taken criticisms (Hirai 2002; Chapman, 2003) from outside seriously and has enhanced the TOEIC several times since 2006 by introducing speaking and writing components and adding a business flavor.

BULATS is available in four European languages: English, French, German, and Spanish. BULATS English was developed by Cambridge ESOL in the late 1990s, specifically as a business English test and is now one of the mainstream business English tests in Europe; however, it is not as well-known as the TOEIC in Japan. From my own experience of taking many English tests, I would rate it as the most business-oriented of all the tests.

Tests of English for tourism are, as the name suggests, mainly interpretation tests focusing on how to interact with foreign visitors and include the National Licensed Guide, the Travel English Test, and the Tourism English Proficiency Tests. In these tests, while the language content is not very challenging, a broad, often meticulous knowledge of Japanese culture and history is required, as well as some familiarity with the tourism industry.

Technical English tests are generally intended to evaluate the candidate's competence in technical communication in English and include the Waseda-Michigan Technical English Proficiency (TEP) and the Kougyou Eiken. These two tests employ distinct test forms for different levels. They both require a basic familiarity with general science and engineering, as well as a knowledge of specialized fields. In this regard, they seem similar to technical translation tests but there is a fundamental difference: The writing components of these tests, especially the TEP (at its highest level), place great emphasis on report structure and rhetoric and are therefore very challenging in their own ways.

The term “technical English” is a major misnomer. Traditionally, technical English used generally to deal with the kind of English used in the scientific and engineering community. For this reason, it has long been made light of in English education, and its tests have been unduly unpopular. In a sense, its name has served as a self-inflicted fetter. However, the principles, particularly the writing principles, taught in technical English have recently gained international recognition as methodologies and guidelines universally applicable to all sorts of professional English. As a result, the more appropriate term ‘professional English’ is gaining currency and drawing more general attention.

Finally, there are a variety of translation tests including the Software Translator Test, the Intellectual Property Test, the Translator Qualifying Examination (TQE), the Hon’yaku Kentei offered by the Japan Translation Federation (JTF), the Certified Professional Translator Test offered by the Japan Translation Association (JTA), and certification tests (in various language pairs) offered by the American Translators Association (ATA). In addition, if one stretches the definition of the term “translation,” the Test of Business Interpreting Skills (TOBiS) can also be grouped in this category.

Most translation tests take a single form consisting of several tasks, and the candidates are given grades. In general, these tests have very rigorous scoring criteria and focus on very discrete points (in other words, are very nitpicky). As a result, achieving the highest grade is a big challenge in any translation test, even for holders of the first grade in Eiken or the TEP, or those with a score of 900 or more in the TOEIC. It should also be noted that translation tests differ from technical writing tests in that the former demand strict adherence to the original text and proper choice of words.

Evaluating the English tests

In discussing linguistic competency from the viewpoints of knowledge and skills, I have been using a five-axis radar chart, shown in Figure 1, in which the axes *GV*, *R*, *L*, *S*, *and W* represent grammar and vocabulary, reading, listening, speaking, and writing, respectively. Using this radar chart, I would like to illustrate how the characteristics of a test affect the resulting perception of the candidate’s ability.

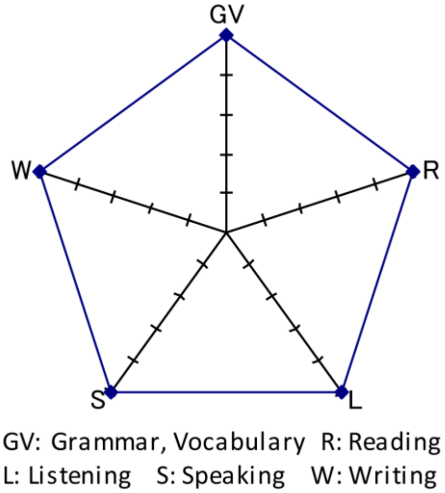

Figure 1. Radar chart of language skills.

Suppose the candidate has the ability profile depicted in Figure 2, which shows a relative strength in writing. If the test he/she takes has the evaluation profile (i.e., how thoroughly it evaluates each skill) shown in Figure 3, which tests *GV*, *R*, and *L,* but not *S* or *W*, as is the case with the conventional TOEIC, then his/her results will have the profile shown in Figure 4, since each axis will have a value equivalent to the product of its corresponding values in Figure 2 and in Figure 3. This means that his/her speaking and writing skills may not be reflected in the test results. In other words, the test itself serves as a filter to the real ability. Here lies one of the fundamental problems inherent in language tests. The great majority of people – most critically the stakeholders – tend to use the test results as the sole source of information for assessing the candidate's linguistic competence, yet the perceived ability does not reflect his/her real ability. If the test does not adequately test active (productive) skills such as speaking and writing, or even worse, does not test them at all, then those with good active skills but relatively weak passive (receptive) skills are significantly handicapped in the test and may eventually fail it. Conversely, those who excel in grammar and passive skills but have poor active skills have a better chance of scoring high in such tests and hence are more likely to be accepted into a renowned university or assigned to a well-rewarded position in a company. This misinterpretation of test results can have significant consequences especially in the case of high-stakes tests.

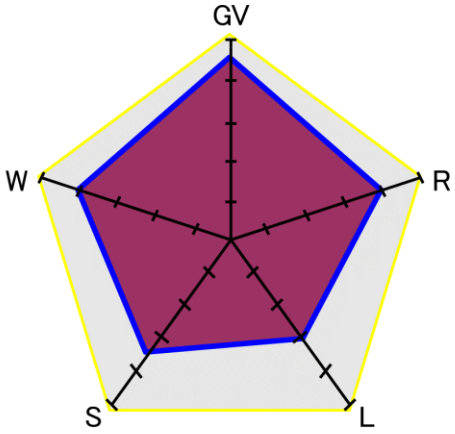

Figure 2. Radar chart of test candidate with strength in writing.

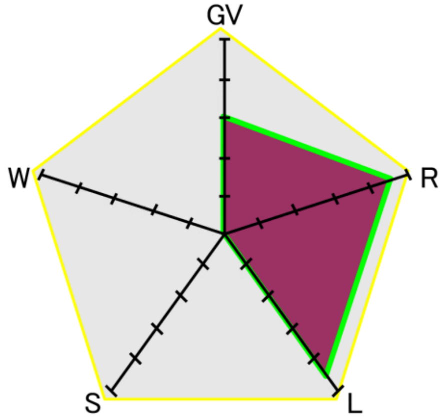

Figure 3. Evaluation profile of test lacking assessment of writing or speaking.

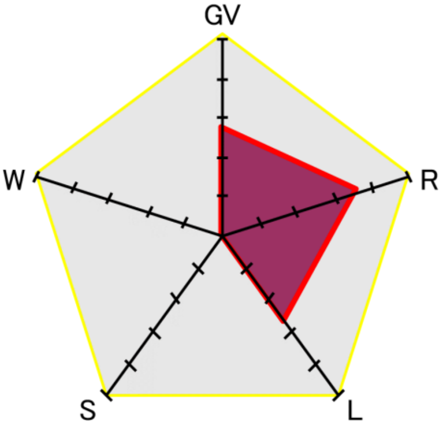

Figure 4. Radar chart of test candidate assessed by test lacking assessment of writing or speaking.

Whereas the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) (Council of Europe, 2001) provides a set of guidelines for evaluating language skills, it is no easy task to objectively determine how thoroughly each test evaluates different skills and also how difficult it is for most learners of English, primarily because most test organizations do not publish the actual test questions/problems, and also partly because test forms change from administration to administration. Based on a collection of personal memos on which I have jotted down my observations and impressions of the tests I have taken, I demonstrated in my presentation the profiles of several major tests as I perceive them. Table 1 summarizes my subjective comparison of difficulty level for some well-known English tests. In my view, among the dozens of general and academic English tests, the Cambridge Main Suite (especially the highest-grade CPE) is the most thorough and rigorous, whereas in the business and workplace domain, the BULATS English best reflects the actual needs of the business community.

Table 1
Level Comparison

CEFR	Eiken	Cambridge Main Suite*	IELTS Band	BULATS		TOEIC			TOEFL	
				Std	S, W	R, L	S	W	PBT	iBT
C2		CPE (A)	8.5 - 9.0	90 - 100	5					116-120
		CPE (B)	7.5 - 8.0			965-990	8	9	652-677	
	1	CPE (C)				940-960	7	8	640-651	111-115
C1		CAE	6.5 - 7.0	75 - 89	4	875-935	6	7	600-639	100-110
B2	pre- 1	FCE	5.0 - 6.0	60 - 74	3	700-870	5-6	6-7	540-599	76-99
B1	2	PET	3.5 - 4.5	40 - 59	2	500-695	4	5-6	470-539	52-75
						400-495	3-4	4-5	400-469	32-51
A2	pre-2	KET	2.0 - 3.0	20 - 39	1					
	3									
A1	4 - 5		0.0 - 1.5	0 - 19						

* For each test, the results are given in 4 grades. There may be overlaps between adjacent tests.

Note. This table summarizes the author's subjective evaluation of English tests.

Testing and learning: What they mean to business and industry

Let me stretch this pentagonal model further to illustrate how the choice of test affects the learner's progress. Figure 5 illustrates two cases, one with a passive-skills-only test (Test A) and the other with a four-skill test (Test B). It is in our nature that given any test, we tend to study only the subjects covered that it covers, especially when we are already overloaded with daily office work. Thus, with Test A, the learners are likely to stop studying or practicing active skills and, after a few years, to end up having improved only their passive skills, while letting their active skills deteriorate. On the other hand, with Test B, they will keep studying and practicing all the four skills and, after the same period of time, will have improved them all in a well-balanced manner. Accordingly, the choice of test critically molds the learners' skill profiles in the long run. This is what I call the "clothes make the man syndrome." (This phenomenon is also well known as the "washback" effects in the language testing community (e.g., Cheng and Watanabe, 2004)).

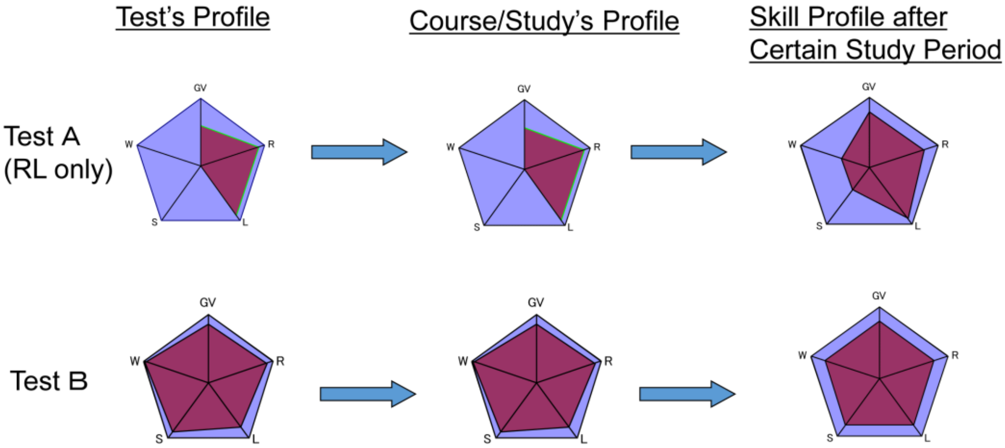

Figure 5. Washback effect of passive skills and four skills tests.

It is alarming that the prolonged use of, or dependence on, a passive-skills-only test such as the conventional TOEIC might eventually drive learners away from active skills. This is particularly true in Japanese companies, which are in great need of employees who can competently communicate with their international counterparts on the business front. A number of studies have revealed that active skills such as those required for presentations, negotiations, meetings, email writing, and report writing are high on the wanted list.

With that in mind, I have been proposing a T-square approach to corporate language education as shown in Figure 6. The idea is to build an elite pool of employees equipped with appropriate active skills (on the vertical axis) while gradually raising the average level of home-front staff on a long-term basis (on the horizontal axis). It is important to note that in language, quantity cannot substitute for quality, in other words, no matter how many 2nd grade speakers a company may have, it cannot beat a team comprising one top-grade person. In this respect, the human resources department should carefully direct the company's investment in language training, rather than spending its limited budget indiscriminately on all the employees. The same argument can apply to language education in schools and universities, *mutatis mutandis*.

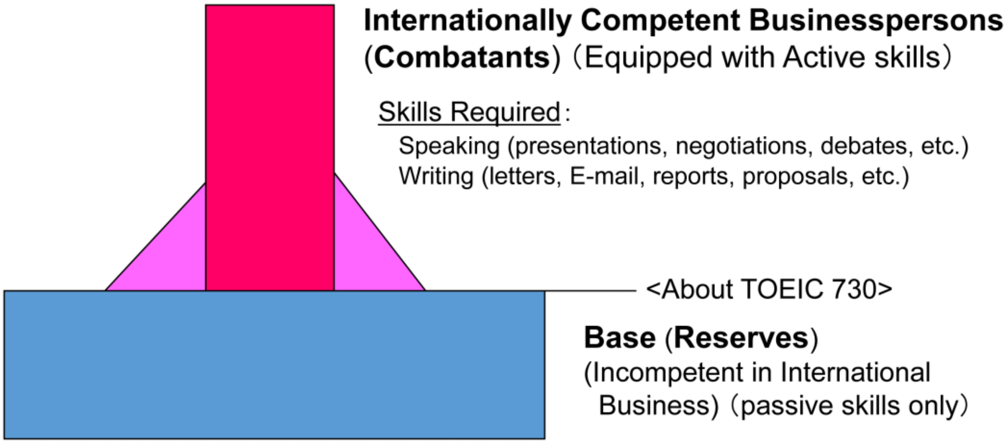
Internationally Competent Businesspersons (Combatants) (Equipped with Active skills)
Skills Required:
Speaking (presentations, negotiations, debates, etc.)
Writing (letters, E-mail, reports, proposals, etc.)
<About TOEIC 730>
Base (Reserves)
(Incompetent in International Business) (passive skills only)

Figure 6. Hierarchical education model of passive base skills followed by active advanced skills.

In this context, it would be worthwhile to review the requirements for language testing from a business perspective. First and foremost, companies should check how well the test they employ aligns with their business objectives and priorities. Without establishing their objectives and priorities and without looking critically at whether the test they employ corresponds with them, it is all too easy to fall victim to a lock-step mentality. No matter how good a test is as a general English test, it is pointless, and therefore a waste of money, to employ it if, for example, what the company needs most is business English competency. In addition to common selection criteria such as validity, reliability, and discriminability, I would list comparability (alignment with an international standard such as the CEFR), ease of use as a management tool, and convenience (ease of administering the test).

Misperceptions and misuses of English tests

In this section I would like to address some of the common misperceptions and the resulting misuses of English tests that are pervasive in Japan. The first misperception is that one test form can uniformly cover all levels – in other words, the notion that if the test can evaluate intermediate levels accurately, then it should also evaluate higher and lower levels equally well. The reality with most tests, however, is that the discriminability is rather poor at the very high and very low ends because of so-called ceiling or boundary effects. Suffice it to say that, in multiple-choice tests for example, one can easily obtain one third or one fourth of the full score by randomly or blindly selecting from the given lists of choices, even without knowing anything at all of the language. It is therefore not ideal to depend on a one-size-fits-all test covering all levels without regard to its intended use.

Another common misperception is to assume that a high score in one test automatically guarantees competence in the workplace. A typical example is the widespread notion that since an employee has scored over 800 in the TOEIC, he/she must be qualified to work as an international businessperson. In actuality, many of those with high TOEIC scores cannot write satisfactory business email or actively participate in business or technical meetings, since there is much more to doing business than merely using the language. Two major factors should be noted. First, the correlation between passive skills and active skills is not high enough to justify such assumptions, and second, language testing and real-life business are different domains. These two points are famously demonstrated in Figures 7 through 9 (Hirai, 2012a; Hirai, 2012b).

Figure 7 illustrates how BULATS Speaking Test scores correlate with TOEIC RL scores. While BULATS speaking scores do tend to increase as TOEIC scores increase, there is a great variance. BULATS Level

3 is the minimum required level for international businesspeople according to a nation-wide survey conducted by Koike, Takada, Matsui, and Terauchi (2010). The regression line shown in red intersects with the BULATS Level 3 line at TOEIC 930. It is also worth noting that 56% of holders of TOEIC 800 or more fail to reach BULATS Level 3. Figure 8 shows the correlation between BULATS Writing Test scores and TOEIC RL scores. The contrast between the two is more dramatic. The regression line does not intersect with the BULATS Level 3 line at all, which means that even with a perfect TOEIC score of 990, more than half of candidates would not reach BULATS Level 3 in business writing. Also, 71% of holders of TOEIC 800 or more fail to reach BULATS Level 3. Figure 9 compares the standard (RL) BULATS test scores of university students and non-student adults. The fairly substantial difference in average score (31.5 vs. 49.6) signifies that in the domain of business English, industry (work) experience plays a significant role.

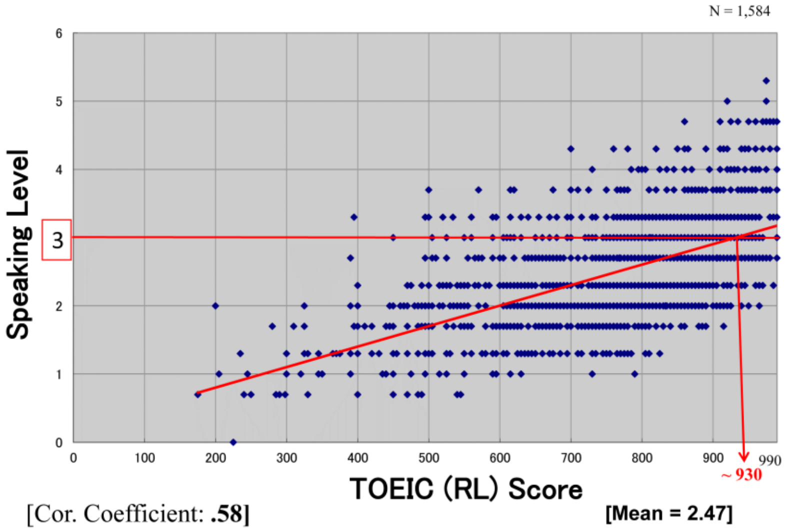

Figure 7. Comparison of BULATS Speaking Test scores with TOEIC Reading and Listening scores.

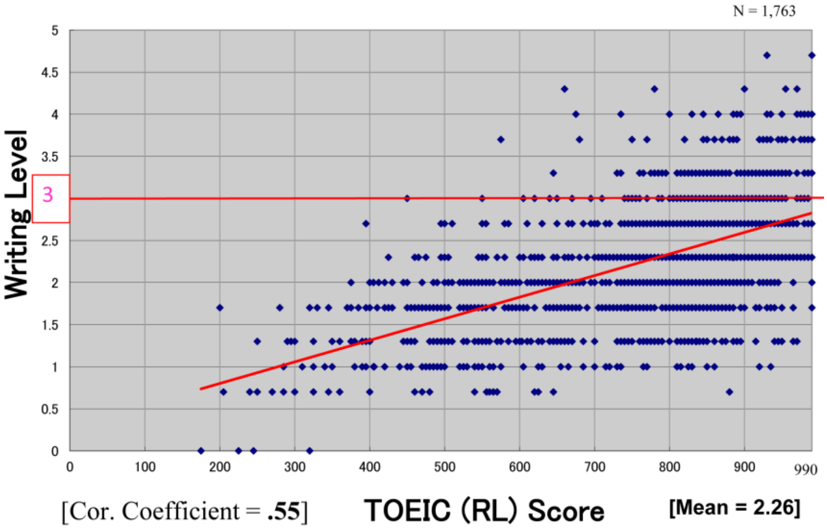

Figure 8. Comparison of BULATS Writing Test scores with TOEIC Reading and Listening scores.

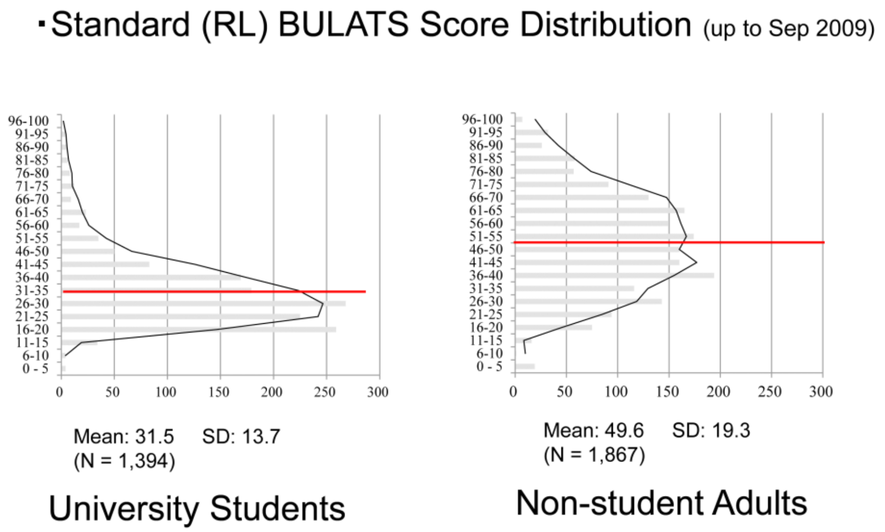

Figure 9. BULATS Reading and Listening test scores for university students and non-student adults.

These findings underscore the importance of choosing the right test for the right purpose. Whereas the TOEIC is very well designed as a general English test (more recently with additional workplace relevance), it would be a big mistake to mindlessly interpret the TOEIC RL score as a reliable indicator of business English competence. A test can only be "good" within the boundaries it is designed for and when its validity (i.e., its ability to test what it claims to test) is warranted. How to use a test is much more important than whether it is a "good" or "bad" test.

Looking ahead

In closing, it would be worthwhile to quickly review what is happening in the field of language testing and try to predict where it is heading. First of all, there is an inexorable shift towards online services. Second, the world is becoming increasingly aware of standardization initiatives such as CEFR, which allows us to compare and evaluate tests and materials with a common measure. Third, test developers continue to improve and enhance their tests by aligning them more closely to how English is used in real life. One example is the proliferation of ESP tests that purport to better serve the needs of industry. In academic English, there is a trend towards the integration of multiple skills. Finally, as with Go and Shogi, artificial intelligence seems likely to eventually make inroads into the world of testing, not to mention the sacred realm of scoring and grading. It may be time for mere humans to pack their bags and retire?

Conclusions

Language testing has many facets and should be viewed from various angles. From the viewpoint of education, it is important to realize how language tests affect the learners' study patterns and hence the formation of their skill profiles. In this regard, four-skills tests are much more desirable (if cost permits) than passive-skills-only tests, because active skills are what is needed most in the real world. From the viewpoint of language-test users, it is essential first to realize the organization's objectives and priorities and then to choose the test (or test battery) that most closely aligns with them. In Japanese industry today, it is crucial to use English tests as a means of fostering well-balanced skill profiles of employees in order to meet their international business requirements. In this respect, it is strongly recommended to use an

English test or test battery that is specifically designed to evaluate the four skills required in actual business situations.

References

Chapman, M. (2003). TOEIC®: Tried but undertested, *Shiken: JALT Testing & Evaluation SIG Newsletter*, *7*(3), 2-7. Retrieved on April 29, 2017 from http://jalt.org/test/cha_1.htm

Cheng, L., & Watanabe, Y. (Eds.). (2004). *Washback in language testing: Research contexts and methods*. Mahwah: Lawrence Erlbaum Associates.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press

ELT Services Japan. (2017). 英語の資格と検定試験, Retrieved on January 9, 2017 from http://www.eigokyoikunews.com/database/exam/

Hirai, M. (2002). Correlations between active skill and passive skill test scores, *Shiken: JALT Testing & Evaluation SIG Newsletter*, 6(3), 2-8. Retrieved on April 29, 2017 from http://jalt.org/test/hir_1.htm

Hirai, M. (2012a). Correlation between BULATS Speaking/Writing and TOEIC Scores. In Chartrand, R., Crofts, S., & Brooks, G. *The 2012 Pan-SIG Proceedings* (pp. 118-125). Tokyo: JALT. Retrieved on April 29, 2017 from http://pansig.org/archive

Hirai, M. (2012b). 受信型スキルテストで仕事における発信型能力を測れるか, *BULATS Journal 2012-2013*, 2-3. Retrieved on April 29, 2017 from http://www.hirai-language.com/wordpress/wp-content/uploads/2013/05/BULATS-Tsushin-2012-2013-Special-Issue-pp.2-3.pdf

Koike, I., Takada, T., Matsui, J., & Terauchi, H. (2010). *企業が求める英語力* (English Abilities Required by Corporations), Tokyo: Asahi Press.

McCrostie, J. (2010). The TOEIC® in Japan: A scandal made in heaven, *Shiken: JALT Testing & Evaluation SIG Newsletter, 14*(1), 2-10. Retrieved on April 29, 2017 from http://jalt.org/test/mcc_1.htm

Appendix

Homepages of English test websites in alphabetical order as of April 29, 2017

American Translators Association (ATA) Certification: http://atanet.org/certification/index.php

BETA (Businessmen's English Test & Appraisal): http://www.ilc-japan.com/tokyo/corporation/gogaku/beta2/bet

BULATS (Business Language Testing Service) English: http://www.cambridgeenglish.org/exams/bulats/

Cambridge English: Key (KET): http://www.cambridgeenglish.org/exams/key/

Cambridge English: Preliminary (PET): http://www.cambridgeenglish.org/exams/preliminary/

Cambridge English: First (FCE): http://www.cambridgeenglish.org/exams/first/

Cambridge English: Advanced (CAE): http://www.cambridgeenglish.org/exams/advanced/

Cambridge English: Proficiency (CPE): http://www.cambridgeenglish.org/exams/proficiency/

Certified Professional Translator Test (JTA 公認翻訳専門職資格試験): http://www.jta-net.or.jp/about_pro_exam.html

Eiken (Practical English) (英検): http://www.eiken.or.jp/eiken/

GRE (Graduate Record Examination): https://www.ets.org/gre

GTEC (Global Test of English Communication): http://www.benesse.co.jp/gtec/

G-TELP (General Tests of English Language Proficiency): http://www.g-telp.jp/english/

Hon'yaku Kentei (翻訳検定): http://www.jtf.jp/jp/license_exam/license.html

IELTS (International English Language Testing System): https://www.ielts.org/

Kokuren Eiken (English Proficiency Test in the Program of the Official Languages Test of the United Nations) (国連英検): http://www.kokureneiken.jp/

Kougyou Eiken (English Technical Writing Test) (工業英検): http://jstc.jp/koeiken/koeiken.html

Nissho Business Eiken (日商ビジネス英検): https://www.kentei.ne.jp/english

Pitman ESOL (English for Speakers of Other Languages): http://anshin-keiri.com/shikaku_02/01_24.html

TEAP (Test of English for Academic Purposes): https://www.eiken.or.jp/teap/

TEP (Waseda-Michigan Technical English Proficiency Test): http://www.teptest.com/outline.html

TOBiS (Test of Business Interpreting Skills): http://www.cais.or.jp/tobis/index.html

TOEFL (Test of English as a Foreign Language): https://www.ets.org/toefl

TOEIC (Test of English for International Communication): http://www.iibc-global.org/english/lr.html

TQE (Translator Qualifying Examination): http://tqe.jp/