社会イノベーション研究
第５巻第１号（1-16）
2010年１月

2009年10月14日掲載承認

The practices of university admissions and entrance examinations: Their impact on learning and educational programs

Dennis Riches

Introduction

Japanese universities are making sincere efforts to improve the quality of higher education, but the usual approach to this challenge may be flawed. Program evaluations focus heavily on the efforts of teachers and what happens in individual courses. Systemic problems and the stronger influences of the educational culture of Japan are usually overlooked, probably because this culture is taken for granted like the air we breathe. One of these systemic problems is the way students prepare to enter university and gain admission. This paper reviews the content of and practices surrounding English entrance examinations at Japanese universities. This review is done through the perspective of the field of language testing that has developed among British and North American scholars in applied linguistics. As opposed to a relativistic approach that sees entrance examinations as adaptations to their cultural context, this analysis raises serious concerns about the validity, reliability and ethics of English entrance examinations.

Because entrance examinations are high-stakes tests, they are produced under tight security and universities are motivated to keep their exams sheltered from critical review by outsiders. This in itself is the primary reason that valid and reliable tests cannot be developed.

It is common knowledge in Japan that the rituals of getting a place at a university are a dreary fact of life that one must simply accept. The general public knows intuitively about what testing specialists call negative washback effect on the way students learn for the six years before they enter university, and the way they learn afterwards. The Ministry of Education, Culture, Sports, Science and Technology (known by the acronym MEXT) suggested guidelines for reform several years ago, but there has been no broad effort to force all universities to abandon methods that are inappropriate. In this situation, each university is better off not taking the risk of reforming for the sake of good principles because doing so would threaten to upset the reliable flow of applicants. Experimentation would be highly risky.

In this paper I will discuss the reasons English entrance examinations should be reformed on moral grounds, and can be reformed even though each university feels constrained to change nothing until its competitors change in unison. The implementation of innovative admissions policies at foreign universities has yielded results that would make such reform appear much less risky than one would think.

Part 1　Basic considerations in test design

1.1　Validity

An essential step in test design is the establishment of the test’s validity. If a test has validity, it is a means of measuring what it claims to measure, and this involves relating test scores to achievement in other endeavors. Validity also means that the use and interpretation of test scores are appropriate. This can be established by longitudinal studies of test takers to see if their test scores correlate to later success in specific domains of ability. The makers of TOEFL, for example, make claims of predictive validity about future academic success in an English speaking academic environment. However, they make no claims about the present language proficiency of the test taker. Anyone who interprets TOEFL or TOEIC scores as a measure of English proficiency is making his own invalid assumptions about the meaning of such scores.

As soon as one attempts to establish validity, one understands that it is a subjective process in which value judgments and priorities come into focus. This is not a bad thing. Much worse is test design in which validity is not explicitly defined. Unfortunately, the situation at Japanese universities is that values and priorities are implicit and unexamined, buried in the traditional way of designing English tests.

When validity does not exist, test results are only self-referential. A high score on a multiple choice grammar test tells only that the test taker is talented at this particular multiple choice grammar test. There is no evidence of a relation to skills that need to be applied in 'real world' situations.

1.2 Reliability

Reliability is another feature of good test design, one which has to be developed alongside validity. A reliable test provides consistent and reliable results. An individual or a group should get similar results if they take, for example, a TOEIC test on one day, and another TOEIC test the next day. The test items are of course different, but the designers of the test are careful to assure that each edition of the test will produce similar scores with the same test takers. Reliability is established by doing statistical analysis and using the results in the training of evaluators and item writers.

The traditions of university entrance examinations make it difficult to establish reliability because this requires pre−testing the test. It is necessary to analyze test development and train test creators and evaluators; and this would require allowing outside experts to review the entire process of test creation and implementation, but universities have many strong motives for keeping their practices confidential. Unfortunately, the examination is a singular event that, because of security concerns, cannot be pre−tested on an experimental group of test takers. Furthermore, test scores have to be returned quickly, and there is no practical way to train the evaluators and ensure consistency among them. Likewise, there is no analysis of the results that is used to make improvements on the next test.

Moreover, test creation is supervised by senior faculty whose authority is difficult to challenge. In fact, the decision making authority is diffuse and unclear, which makes it unlikely that individuals will willingly engage in critical and open discussion of testing practices.

Part 2　Establishing Validity

Part 2 and 3 focus on problems in establishing validity and reliability with reference to the actual test item types and format of the English entrance examination of *X University*. Whether these item types are problematic depends entirely on judgments about what should be learned and how it should be learned. What skills do we want to foster in learners both as they prepare for such tests years in advance, and for years afterward as they continue to learn English? If such matters were explicitly discussed and decided beforehand, a test that is valid for a specific purpose could be developed. As it is now, we have English entrance exams that are valid for a rapidly industrializing 19th century nation that gave university education to only a small minority of its high school graduates. This situation persists because people have relied on tradition, and no one has taken the time to articulate objectives for English education and English tests in the contemporary setting.

2.1　Use of Japanese

A glance at a completed answer sheet reveals much about the flaws in the English entrance examination of *X University*. Almost everything that is written to answer the questions on this test of English is written in Japanese. A foreign observer might not even be able to guess that it was an English examination. It is weighted heavily toward asking the test taker to demonstrate *in Japanese* his understanding of English reading passages. There is no way for a foreign specialist to assess this test, or to compare it with English tests used in other countries. This is a symptom of a serious problem in English education in Japan; that is, it makes English an object of study within Japanese society, something to be picked apart, analyzed and left dead on the table. People who are forced to study a language in this way are unlikely to think of it as something to use for communication with the world outside Japan. A further problem with requiring answers written in Japanese is that the test becomes an unreliable measure of English proficiency. It tests both Japanese and English proficiency at the same time, and markers can easily become influenced by the test taker’s poor command of his native language, or even his poor penmanship.

2.2 Productive and Passive Skills

Of the four language skills - speaking, writing, listening and reading - the first two are productive and the last two are passive. In this entrance examination, almost all of the items demand only the passive reading skill, or passive recognition of correct answer choices. One short translation item requires the test taker to translate a compound sentence into English.

This type of examination might have been valid in the 19th century when Japan wanted to rapidly import technical knowledge from the West and disseminate it widely to the general population in Japanese. Communication and interaction with non-Japanese people was not a priority because only very few people had opportunities to travel, and it was assumed that knowledge was flowing in one direction, from the West into Japan. There was little notion, in Japan or elsewhere, that economic development depended on fostering self-expression and creativity, or that these things were valuable regardless of their economic utility.

2.3 Sampling

A different problem is that of defining the corpus of knowledge to be tested, and sampling it fairly. The MEXT has quite specific descriptions of the corpus of English that a high school graduate should know, but questions are chosen from this corpus in a random and unsystematic manner. To test knowledge of a few hundred idioms, the test includes two or three items about idioms. If a test taker managed to memorize 75% of the idioms in the corpus, there is still a high possibility that the three idioms appearing in the test are from the 25% of the idioms that he does not know. The sampling is too small to be a reliable measure of the learner’s knowledge of idioms. The same problem applies to testing grammar and writing ability. The single productive exercise in the test is the translation of a single sentence from Japanese to English. Again, the test taker may have an excellent knowledge of grammar and vocabulary, but not the grammar and vocabulary needed to do this short translation. The problem of sampling could be solved somewhat if the test did not require the test taker to spend so much time reading a long passage about which only a few comprehension questions are asked in Japanese.

Test validity would be greatly improved if the university could articulate the value it places on language use and communicative skills, and design a test that reflects this value and encourages applicants to develop these skills. There could be much less emphasis on passive reading skill, and the use of Japanese on the answer sheet could be greatly reduced. Finally, test writers could more thoughtfully select a sample of test items that fairly measures the knowledge of grammar, vocabulary and idioms in the prescribed corpus of high school English.

Part 3　Establishing Reliability

Reliability refers to the consistency of measurement. A single answer sheet should be scored the same way regardless of the rater who scores it. A person who takes one edition of the TOEIC test on Monday and another on Tuesday should get two similar scores. His results on TOEIC should correlate to his scores on another test that makes the same validity claims. Reliability will suffer if there is no analysis of test results. A test may have too many items that almost everyone gets wrong, or too many questions that almost everyone gets correct, thus weakening the power of the test to discriminate between test takers. In all other educational contexts, discrimination is a negative word; but in testing, it is the central objective.

Validity and reliability are not the same thing, but each supports the other. Validity is the interpretation and use of test scores; the statement of what the test measures, so of course this has to be supported by reliable measurement.

3.1　Inter-rater reliability

The reliability of English entrance exams suffers because the raters, or markers, are not adequately trained to rate the test items in a consistent way. Because of security concerns, they cannot see the test beforehand, and the test cannot be pre-tested before it is given on a single day of high-stakes testing. Ideally, an experimental group of test takers would provide answer sheets for the raters to rate ‘blind.’ The raters would rate the same papers not knowing who rated them previously or how they were scored. Afterwards, discrepancies could be discussed until all agreed on consistent standards for evaluating each item.

3.2 Inter-test reliability

Just as pre-testing and training could improve rating, it can also improve the test by allowing a comparison with other tests. If the test results did not correlate with the same test takers' scores on standardized tests, this would be cause for concern.

3.3 Human error

Several of the items on the entrance examination are multiple choice questions, but they are scored by humans. Although the questions are double-checked, such items can be more reliably and inexpensively scored by machines.

3.4 Item analysis

Each test item needs to be analyzed to find out how many test takers answered it correctly. If there is a small percentage of items that everyone got wrong, and a small percentage that everyone got right, this provides data on the lower and upper limits of the group's knowledge and ability. However, having too many such items renders the results less valuable because it becomes less possible to discriminate between the test takers in the group, and we must accept that discrimination is the whole point of doing testing in the first place.

Item analysis data can be easily obtained, but it is not used in creating the next year's examination. The committee for one year closes its files, and next year's committee sets about creating the next examination without first sitting down to discuss how the data indicates what needs to be revised.

One interesting finding that might come from item analysis is a difference between types of items. Multiple choice sections seem to have higher scores, while sections requiring productive use of language have lower scores. This situation is considered acceptable because the two types balance each other out and seem to be fair to the test taker. Such thinking is completely wrong; both sections of the test should be presenting a challenge to test takers. If one of them is not, this is merely wasting everyone’s time and should be revised or eliminated.

The negative effects of multiple choice tests have been widely discussed in education literature, and a recent Japanese Nobel laureate, quoted in the Asahi Evening News, referred to them as a sort of “educational pollution.” If language education is to be focused on language use by the learner, ways of measuring language ability ought to encourage the learning of this skill. In actual language use, speakers are not offered a choice of a, b, c or d as they try to quickly recall a word or form a grammatically correct sentence. The dilemma of multiple choice questions is that they increase a test’s reliability but decrease its validity; that is, they measure accurately a useless skill that has no proven connection to real world challenges. The information they provide about test takers’ proficiency is useless for making inferences about how an individual will apply such knowledge in the real world.

Nonetheless, if for practical reasons universities cannot abandon multiple choice items, much could be done to improve them. Distracters (the incorrect choices) need to be truly distracting; that is, a tempting choice that many test takers will consider choosing. If too many of the items are answered correctly by most of the test takers, the results become less useful.

These problems in establishing reliability exist mostly because of the essential high security needed for the single event of the annual entrance examination. The items cannot be pre-tested, and cannot be opened to external review and criticism beforehand. Without having seen the examination or having agreed on how to rate it until the day of the test, a group of some thirty amateur raters (amateur in the sense that testing is not their specialty and they receive no specific training) come together for four days to mark several thousand examination papers, then they go back to their main jobs. A review of these practices by an international expert in testing would likely invite withering criticism. The results from such a test can provide little useful information about the English proficiency of the test takers.

Part 4　Other Questionable Practices

There are several other problems with English entrance exams, and these are related to the cultural and institutional settings where they take place. University entrance has been decided for over a century by the existing testing practices and rituals, and indeed they seem to have their roots in Confucian philosophy and the imperial education system of ancient China. Throughout history, such competitive examinations have had their virtues as ways of advancing by merit rather than by connections and hereditary privilege. It is difficult for Japanese society to agree that there is a need for new solutions, to imagine innovative approaches to higher education, and then to implement innovations in a way that would still be perceived by the public as a fair way to select people for the most sought-after careers.

Nonetheless, there are so many problems around entrance examination practices that they have become a meaningless ritual that has a negative impact on the way people learn English and on their attitude about learning.

Firstly, English examinations do not follow the MEXT guideline that universities should move toward including a listening component in their entrance examinations. This is an essential step in moving the entire English education system toward the communicative model.

Secondly, universities need to consider what they learn from their examination results about the abilities of the test takers who actually accept a place at the university. Aside from the few highest ranking universities, most institutions at the next level below attract applicants who have average to high average scholastic aptitude. The results of English examinations at such universities usually yield an abnormal distribution curve, with very few test takers scoring above 80%. Those who do score high are likely to accept a place at a higher ranked university. Thus, the university ends up accepting a group that consistently scores 60-75% on an examination consisting of mostly multiple choice items, which has no validity or reliability. Only a vague, general conclusion can be made about such learners. They have learned some English, but they do not have a clear understanding of how to use it. They have not mastered the basics of the high school English curriculum. They will need to go over this material again as ‘false beginners.’ They are average.

The irony of this finding is that it could have been obtained more easily from other assessments such as high school performance and other standardized tests. If applicants could demonstrate their intelligence and scholastic ability in other ways, they could be admitted into the university to study English without having to sit for an entrance examination. Since at present most university English teachers have to go back over the basics of the high school English curriculum, it makes no sense to carry on with the pretense that the English entrance exam is a method of selecting only the best who have fully mastered this curriculum. In the present circumstances, the decisive cutoff line of entrance examinations is that which distinguishes between those who have learned a little, and those who have learned almost nothing.

Critics of this proposal will say that the present English high school education is too unreliable and inconsistent to be used for university entry, but this is only because extracurricular entrance examination preparation has undermined high school education. If more value were placed upon high school English education, its quality would improve greatly.

It may seem foolish to decrease the competitiveness of university entrance when it is the common assumption that a university should compete to get the very top students. However, in a society that sends such a high proportion of high school graduates to university, very few universities can hope to attract the very best. As mentioned above, if entrance examinations were ever a good idea, it was in the past when university education was available to only those in the top percentiles of academic ability. In the present context, entrance examinations are a sort of self−deception on the part of universities. These institutions create a difficult examination that few actually score highly on, and this creates the illusion, on paper only, that entry requirements are very rigorous. This is an attempt to uphold the university's reputation with a document for public display, and to avoid what Schwartz (2005) calls the "death spiral" – the vicious circle of declining reputation followed by declining quality of applicants.

In fact, accepting what are considered to be second rank students is not such a disadvantage to a second or third tier university. A study of reverse discrimination policies (known as affirmative action) at the University of Michigan law school led to some surprising results. Affirmative action programs were implemented to correct the historical injustices of racial inequality, but it was feared that admitting minorities who did not have the highest scores would be unfair, and that this would lead to lower standards and achievement in professional careers. However, a study by Adams et. al. (2000) that tracked students through their studies and their careers showed that people who scored very high on the LSAT and in undergraduate studies were no more successful than the people who entered thanks to affirmative action policies. Success was defined by measures of income, subjective satisfaction and contributions to the profession, or whether they even stayed in the profession[1]. One might say, yes, but their grade point average and their LSAT scores were not the highest, but this is to mistake a shadow for the real object. Grades and standardized test scores are not success. They are only supposed to *predict* real life success, and in this case, the results of the study only show that above a certain cutoff, which is lower than we would think, the scores do not predict much. The finding also supports a hypothesis that one might formulate based on anecdotal experience and the biographies of famous tortured geniuses. The sort of person who can get a perfect score on an academic proficiency examination may be a ‘failure’ in ways that are not measured by academic tests. Measures of ‘success’ that interest educators and policy planners are inherently measures of social success. What does the person receive from society? What does she contribute to society? No education system is deliberately oriented toward producing solitary geniuses who live on the fringes as non-conformists.

Schwartz has made similar observations about the elite institutions which actually get so many applicants with perfect SAT and high school records that they have no fair way to choose among them. He has suggested that Harvard could easily dispense with its expensive admissions screening and simply gather all the applicants who are “good enough” and select them by lottery. In these elite schools all the applicants have near perfect scores and applications, but over half of them need to be rejected. The situation is similar farther down the hierarchy. The applicants at a lesser ranked school all might fall within the same range of a few percentage points, so there is no fair way to decide who should be rejected.

Such theorizing has lent support to a growing movement in America to reduce or eliminate reliance on SAT scores and admit students based on their high school

[1] This study is, however, open to various interpretations. Job satisfaction is not to be confused with an observer’s measure of job performance, and “contributions to the profession” are difficult to quantify. Furthermore, affirmative action continues to be a factor in hiring and promotion decisions throughout these careers, so these continuing advantages will form a feedback loop with the continuing satisfaction felt by people who benefit from the policy.

record, or other skills and achievements. Pink (2004, pp. 57-59), for example, describes new methods of assessing “right brain” creative problem solving to be used in formal admissions screening. An organization in Boston, The National Center for Fair and Open Testing, has been advocating in favor of reform of university admissions since 2002, and it has already been influential in the few years of its existence. All of the schools that have abandoned reliance on standardized test scores report improved student satisfaction and performance, and improved reputation of the institution. A skeptic would notice that few elite universities are in this group, but this is beside the point. This is an innovation that is useful to the second and mid-ranking universities who want to give the best education possible to the students they actually get, not the ones they wish they could have.

This finding underscores what is easily forgotten in the preoccupation with competitive selection processes. High school graduates in Japan have already completed standardized national achievement tests and received grades and diplomas from a standardized national education system. Making them take entrance examinations is just overkill, or it is an admission that universities consider the public education system to be unreliable. Whether students succeed at university depends on the quality of their experience after entering university, and such quality is much more likely to be achieved if students have not experienced a phenomenon which their society refers to as “entrance exam hell.”

If universities still want to insist that prospective students take a difficult examination, they could rely on specialized test producers that have the resources to make reliable and valid tests that are open to public scrutiny. This would allow professors to devote more time to teaching and research in their specialties. Yet this would also require the individual professors and the universities to forego the financial incentives involved in holding entrance examinations.

Unfortunately, most universities are stuck on having their own branded examination as a way of signaling to the public that their standards are difficult to attain. The entrance examination is also a trick that any car salesman would understand. The customer must put down a deposit and commit to taking only the number of entrance examinations that he can afford. Instead of having a portable score from a standardized test that could be used to apply to any number of universities, the student is forced to apply through an entrance examination that is held only once a year. The student can afford the time and money for only a few examinations. It is curious that in a higher education system that is competing for ever fewer applicants, universities cannot make admissions procedures that offer more benefits and flexibility to applicants.

Conclusion

Universities worry a great deal about faculty development and improvement of their programs, but, curiously, there is little concern with the powerful influence of the selection process on the quality of secondary and post-secondary education. By the time students begin their first university courses, they have already been traumatized for several years by what in Japanese society is well known as "entrance exam hell." They begin their experience in higher learning with cynicism and a passive resentment about what they have been through, with an attitude that they have earned some time to relax. Learning is not enjoyable; it is just work. Changing these attitudes is an essential first step in fixing whatever else is wrong with university education, so no reform will be effective without such changes.

While this situation persists, faculty carry on with the annual ritual of student class evaluation in which respondents anonymously answer multiple choice questions about the quality of the teacher's voice and the classroom temperature, as well as some more serious questions. At faculty meetings we review the results, with the names of teachers and survey respondents all concealed, and wonder how to tweak the survey for the next year.

By this point, it should be clear that this paper is asserting that English entrance examination practices are not only a ritual of little practical use, but that they do actual harm to students and educational programs. This is the import of papers by Fulcher, Amrein and Berliner, Braun and many others who work in the field of testing. The ethical problems arising from testing and competition seem to be well understood, but difficult to remedy. The Japanese government and many universities have made efforts to admit a limited number of students by other methods, such as the national Center Examinations (national standardized tests of core subjects), ‘admissions office’ applications (AO), recommendations from high schools, etc... Unfortunately, the AO system has gained a reputation as an easy route in for those who are the least academically prepared in high school. The AO system could be used as a way for high achieving students to apply without having to sit for an entrance examination, but the established reputation of AO may lead to this possibility being overlooked. All in all, the majority of students still enter through entrance examinations, and reform in admissions procedures has been very gradual.

Each university can claim that it cannot change the established system when all other universities carry on with it. They will say that reform needs to be coordinated from above by the MEXT. This is true, but as a long time has passed and the government has not taken action. There is a moral imperative for universities to do the right thing and demand change from below. If entrance examination practices are having a harmful effect on attitudes toward learning, then there is no excuse for continuing with that which causes the harm.

If this argument is unconvincing so far, consider one additional point: American universities all struggle with similar dilemmas about how to make admissions fairer and how to diminish the negative effects of competition, but they do not turn to entrance examinations as a solution. If they really were such a great thing, elite universities in America would adopt this marvelous innovation from Japan and use it to solve their problems. Imagine the reaction if you told the faculty at Yale that they should form an entrance examination committee made up of the professors from the core subjects, and that they should devote less time to teaching and research and more to composing entrance examination questions. Yale too could have its own branded and unique entrance examination, but for some reason it chooses not to. Good ideas from Japan, such as hybrid cars, travel very quickly across cultural barriers, so there is something to be learned from the fact that entrance examinations have not been adopted elsewhere.

After all this criticism, the author is perhaps obliged to offer a solution or way out of the dilemma that universities find themselves in. For this, the suggestion is bottom up pressure on MEXT. If the government will not allow major reforms, it would be feasible, for example, for a dozen similarly ranked universities in the Tokyo area to band together to create a single high quality, reliable and valid test of

English proficiency which they could use jointly (some would worry that this would lead to illegal collusion in admissions, but this could be avoided). Such an entity could open the test to review by foreign experts in testing, and take the lead in establishing good practices; or they could just abandon entrance examinations altogether and focus more of their energies on teaching and research. If they engaged in a well planned public relations campaign appealing directly to the public, the government and other universities would have to follow the example of the leading group. This is just one way that universities could innovate and become more appealing to applicants, but it would first require acceptance of the assertion of this paper: English entrance examinations at Japanese universities have serious design flaws and harmful effects which oblige us to revise their design and use.

References

Adams, T. K., Chambers, D. L. & Lempert, R. O. (2000). Michigan's Minority Graduates in Practice: The River Runs Through Law School. *Law & Social Inquiry, 25* (2), 395-505.

Amrein, A. L. & Berliner, D. C. (2002, March). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives,* 10 (18). Retrieved March 5, 2009 from http://epaa.asu.edu/epaa/v10n18/

Au, W. (2007). High-Stakes Testing and Curricular Control: A qualitative metasynthesis. *Educational Researcher,* 36 (5). Retrieved March 13, 2009 from http://www.aera.net/uploadedFiles/Publications/Journals/Educational_Researcher/3605/07EDR07_258-267.pdf

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing.* New York: Oxford UP.

Braun, H. (2004, January 5). Reconsidering the impact of high-stakes testing. *Education Policy Analysis Archives* 12. Retrieved March 6, 2009 from http://epaa.asu.edu/epaa/v12n1/

Educating the Children. (2008, December 12) (editorial) *The Asahi Shinbun* [Tokyo]. Asahi News. Retrieved December 14, 2008 from <http://www.asahi.com/english/Herald-asahi/TKY200812110064.html>.

Fulcher, G. (1999) Ethics in Language testing. *TESOL Testing and Evaluation Special Interest Group Newsletter - Special Conference Issue,* Volume 1 (1). Retrieved March 20, 2009 from http://www.languagetesting.info/artlt.html

National Center for Fair and Open Testing. Boston. Retrieved March 10, 2009 from http://www.fairtest.org/

Pink, D. H. (2005) *A whole new mind: why right-brainers will rule the future.* New York: Penguin.

Schwartz, B. (2005, February 25) Top Colleges Should Select Randomly From a Pool of 'Good Enough.' *Chronicle of Higher Education - Special Supplement.*

TESOL (2009) Position Statement on English Entrance Examinations for Nonnative English Speakers at Schools and Universities. Retrieved April 4, 2009 from http://www.tesol.org/s_tesol/bin.asp?CID=32&DID=12303&DOC=FILE.PDF