

Title	The University English Entrance Exam and Its Implied Effect on EFL Pedagogy in Japan
Sub Title	日本の大学入試と高校英語教育とその影響
Author	Shea, David P.
Publisher	慶應義塾大学日吉紀要刊行委員会
Publication year	2009
Jtitle	慶應義塾大学日吉紀要. 英語英米文学 (The Hiyoshi review of English studies). No.55 (2009.) ,p.97- 132
JaLC DOI	
Abstract	
Notes	
Genre	Departmental Bulletin Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=AN10030060-20091218-0097

慶應義塾大学学術情報リポジトリ(KOARA)に掲載されているコンテンツの著作権は、それぞれの著作者、学会または出版社/発行者に帰属し、その権利は著作権法によって保護されています。引用にあたっては、著作権法を遵守してご利用ください。

The copyrights of content available on the KeiO Associated Repository of Academic resources (KOARA) belong to the respective authors, academic societies, or publishers/issuers, and these rights are protected by the Japanese Copyright Act. When quoting the content, please follow the Japanese copyright act.

The University English Entrance Exam and Its Implied Effect on EFL Pedagogy in Japan

David P. Shea

The university entrance examination plays a pivotal and often polarizing role in discussions of English language education in Japanese high schools. While many critics see the exams as exerting a negative influence on EFL pedagogy, some view them in a more positive light. Rohlen (1983), for example, called entrance exams the “dark engine” driving EFL pedagogy, creating heated competition and instruction tailored to tests, but he acknowledged the overall motivational effect of the exams and their function in maintaining high standards and “habits of hard work, persistence, and mastery of detail” (p. 319). Along similar lines, Zeng (1999) argued that the exam system exerts an overall positive influence on high school education, working to both reinforce core cultural values of fairness and meritocracy, and equalize access to higher education. Exams, Zeng asserted, prevent the affluent from taking unfair advantage of the system and grant lower income families a better chance of gaining admission to selective universities than other procedures.

Nonetheless, English entrance exams have been widely criticized for their poor construction and negative influence. Brown (2002; see also Leonard, 1998) has spoken prominently against the system, charging that the exams are detrimental “nonsense” that bright students can avoid by

leaving to study at North American universities (p. 96). Brown charged that the exams are unfair, unreliable, and invalid because they are “developed by amateurs who know nothing about test design” (p. 100) and fall far short of psychometric standards. He pointed out that few if any exams are piloted or analyzed statistically, which means that results are actually invalid indications of actual language proficiency. Brown (1996) also raised other issues of ethics, pointing out that given the great deal of money made from constructing exams in-house, universities have a “vested interest” in maintaining the traditional system. Speaking from an insider’s perspective, Murphey (2004) similarly criticized the prevalent practice of constructing exams without carrying out fundamental item facility or differential statistical tests prior to implementation.

In terms of pedagogy, entrance exams are generally seen as exerting strong influence on EFL instructional practices, promoting a structural linguistic model that rewards passive knowledge of grammar and vocabulary and encourages a traditional focus on intensive word-for-word reading and translation (Hiramatsu, 2005, p. 116). Similarly, Shohamy (2005) stated that the exams “require grammatical knowledge and literacy skills ... in direct contrast to a recent national reform effort requiring oral communication classes in the high school English curriculum” (p. 106). Poole (2003) also charged that exams do not follow communicative curriculum guidelines established by the Ministry of Education’s course of study, suggesting that university test makers are either (1) unsympathetic to the aims of state planners, (2) ignorant of proper testing methodology, (3) unaware of the negative effect of grammar-translation pedagogy, or (4) all of the above (§3.3, ¶3). Reporting on a broad survey of teacher perceptions and belief, Gorsuch (2001) suggested that many English teachers “feel they are expected to prepare students for university entrance exams by having

students translate English passages into Japanese, taking vocabulary quizzes, and focusing their instruction on developing students' linguistic knowledge at the expense of linguistic skills" (§3, ¶4). McNamara and Roever (2004) asserted that the structure of the exams "hamper the teaching of oral skills" (p. 206) and promote entrenched values and ideologies (p. 209).

There are indications that the influence of entrance exams may be changing, if not waning. The recent decline in the number of Japanese young people of college age is seen as exerting pressure on universities to relax admissions standards (e.g., Mulvey, 2001). The number of 18 year old applicants nationwide has decreased to equal the total number of seats available at all universities, and dramatic demographic changes have pushed some colleges and universities to reduce the role of the exam in admissions decisions, especially at private institutions. Many universities have adopted a broad range of "assessment tools" including recommendations, interviews, and recognition for cultural activities and community involvement (Amano & Poole, 2005, p. 694). Further, not a few universities have special entrance procedures (including separate exams) for returnees with flexible and arguably less rigorous admissions requirements (Goodman, 2003). Many private institutions also guarantee admission to graduates of attached high schools. While demographic changes are undeniably significant, however, competition for entrance to first-tier universities continues to be strong (Hooghart, 2006) and the credentialist ideology prevalent in Japanese society that drives the traditional selection process "remains fundamentally unchanged" (Amano & Poole, 2005, p. 694). It is fair to assume, then, that the entrance exam has not weakened its influence at those high schools where students hope to enter a competitive university near the top of the pyramid of prominence and rank, according to which Japanese society is still largely structured.

Characteristic Features of Entrance Exams

Analysis of entrance exams themselves suggests a diverse and unsystematic complexity (Zeng, 1999). Typically each university and, in many cases, each department within each university constructs its own exams. The non-affiliated Center (*sentaa*) exam, produced under the aegis of the Ministry of Education, has recently exerted a standardizing force on the overall process, because it is used by public (and increasingly private) universities as a screening test. However, most universities, including major public institutions such as the University of Tokyo, follow-up the Center exam with independently constructed second stage (*niji*) exams (Guest, 2005), which are developed in house (Murphey & Sato, 2003) and used to make further admission selections after consideration of Center exam scores.

Brown and Yamashita (1995) carried out a seminal analysis of entrance exams at 21 of the largest, “most prestigious” universities, ten public and ten private, including the Center exam. They calculated differences in item type and readability levels, in addition to comparing the exams at private and public universities, looking at what skills were measured and whether test items were receptive or productive, discrete-point or integrative, or translation-based. Brown and Yamashita found “tremendous variation” in question type, length, difficulty, and kind of response across the 21 exams. They concluded that, given the intricate construction of the tests and the complexity of item types, exams assessed “testwiseness” (i.e., knowledge of exam procedures), which they concluded was likely “more important than actual abilities in English” in determining test scores (p. 27). Brown and Yamashita also found that nearly all the exams were receptive in orientation, composed of discrete-point multiple choice items with little written or productive use of the language. There was no spoken English

required on any of the exams, and few included a listening component. Many exams called for extensive translation of English into Japanese, and most had a level of reading difficulty far beyond the range of high school EFL textbooks. In particular Brown and Yamashita criticized reliance on translation as a specialized skill woefully out of date with contemporary approaches to FL education, asserting that “translation was abandoned years ago in ESL instruction” (p. 28).

Many critics feel that “little has changed” in the decade following Brown and Yamashita’s study (Cornwell, Simon-Maeda, & Churchill, 2007, p. 122). Kikuchi (2006) carried out a replication of Brown and Yamashita, and argued that receptive, discrete point items still dominate most exams. Kikuchi found that item type “varied tremendously” (p. 85) across exams (with combinations of multiple choice, short answer and translation), and that most exams were extremely difficult in terms of readability. Although exams at a few smaller schools were “fairly easy,” most larger universities used difficult passages constructed at or above college-level readability levels, which would prove difficult even for native speakers. However, Kikuchi reported major changes in the character of recent exams. Some universities, he pointed out, included productive items, calling for students to generate a written response, especially involving summation.

Perhaps the most significant change in the entrance exam system was the introduction of a listening component on the 2006 national Center test (Murphey, 2006). Students had to listen to short oral passages and respond to written questions, selecting answers in multiple choice format. While many (but not all) universities have recently introduced listening components on the second stage exams, Sage and Tanaka (2006) carried out psychometric measures of the 2006 Center listening test and found low construct validity as well as item differentiation and item facility indices. They suggested that

the exam was unable to reliably discriminate oral proficiency, especially when compared with commercial tests such as TOEFL. The Center test has nonetheless been widely heralded for its potentially positive effect on the high school curriculum to promote more practical, conversationally oriented speaking and listening instruction.

Mulvey (1999) and Guest (2000) presented alternative views of recent entrance exams and their effect on pedagogy. Both argued that despite widespread perceptions, the exams are no longer receptive instruments focused on grammar and translation. Rather, they suggested that exams function largely as a measure of reading comprehension. Mulvey asserted that classroom teachers who rely on traditional methods of instruction generally ignore the communicative thrust of exams, but Guest contended that the Center exam actually encourages progressive teaching in high school EFL classrooms. Guest (2005) extended his argument, carrying out a analysis of the 2004 Center exam as a “focal point” of the exam system. He held that the exam required an “integrative knowledge of English” and “holistic and comprehensive understanding of a text’s flow” (p. 1188). Although the format of the Center exam was primarily multiple-choice, there were no questions requiring translation or knowledge of decontextualized grammar. Guest’s analysis pointed to an overall concern for communicative proficiency, but it raised critical questions about the preliminary role of the Center test and larger structure of the entrance exam system, since it is the more difficult and specialized follow-up exam of each university to which HS teachers and students orient. Guest’s analysis also pointed to the complex ways in which the entrance exam works to shape classroom instruction.

Washback on the Curriculum

Although Shohamy (2001) cautioned that using tests to drive instruction

may be undemocratic and ethically questionable, there is widespread feeling that improving pedagogy in the Japanese EFL classroom is contingent upon changing the entrance exam. Many proponents of communicative language teaching advocate examinations as a form of “positive washback” (Brown, 2000) to improve EFL pedagogy in high school classrooms. For example, Akiyama (2003; see also McNamara & Roever, 2006) asserted that HS teachers ignore speaking skills because they are “not included in the highly competitive university admissions tests that formed a primary goal of teaching at the senior high school” (p. 205). Empirical investigations of the effect of washback in a variety of settings suggest, however, that the effect of tests on shaping curricular innovation is not always clear (Wall, 2000). High-stakes tests do not “automatically lead to innovation in teaching and learning” (Watanabe, cited in Newfields, 2005).

Chapman and Synder (2000) reviewed a wide range of international research on the effect of high-stakes tests on instructional practices and found that changes in key tests rarely produced change in instructional practice. Even when teachers “understood the examination requirements at a cognitive level, they were often unable to make necessary changes in the classroom to improve student performance. Moreover, experimenting with new pedagogical approaches was threatening to teachers, who felt under enormous pressure to look good and show good local test results” (p. 462). In Hong Kong, Cheng (1998) found that it was easier to change the content of textbooks and classroom materials than teacher behavior, and it was even more difficult to influence teacher attitudes and values about what *should* be taught, using tests as incentives. The influence of Ministry mandated curricular reform was strongest on textbooks and the “what” teachers addressed in class. There was less impact on “how” teachers actually taught material and the methodologies they used.

In Guangdong, China, Qi (2005) looked at attempts to use the National Matriculation English Test, a standardized norm-referenced measure, to change EFL classrooms focused on grammar and vocabulary. Qi found that in spite of the intentions of test developers, teachers turned supposedly communicative classrooms into “permanent testing halls and students into frequent test-takers” (p. 162). In some cases, teachers were reluctant to introduce communicative activities in the classroom because of their own lack of proficiency, although most teachers simply believed they should emphasize memorized knowledge of vocabulary and linguistic structure even though items that tested linguistic knowledge constituted “only a small portion” of the test (p. 156).

In Japan, Watanabe’s (2004) ethnographic study of high school teachers also contradicted popular assumptions that entrance exams shape classroom pedagogy. Observation of three EFL classrooms made clear that, while one teacher used past tests as written practice exercises, another focused on listening activities and pairwork to develop skills “usable in real-life situations” (p. 139). Watanabe argued that “teacher factors, including personal beliefs, past education and academic background, seemed to be more important in determining the methodology a teacher employs” (p. 130). In other words, a focus on test preparation did not bring about either automatic or uniform changes in instructional practice across instructional contexts. Watanabe’s analysis suggests that teachers make pedagogic decisions based on how they perceive what best improves student proficiencies, but these assumptions are often biased and inaccurate. Teachers nevertheless attribute negative results of pedagogy to the exam and justify instructional decisions in terms of its influence.

Looking at the Todai Exam

In the remaining sections of this paper, I propose to look specifically at the 2007 and 2008 English entrance exams of the University of Tokyo (Todai), which as the most selective university in the country, exerts an enormous influence on the high school EFL curriculum. I carried out a close analysis of the exams in an attempt to see the kind of proficiencies that were measured, as well as the assumptions embedded in the tests, with a particular concern to delineate what students are expected to know and do. In particular, I wanted to consider the need for grammar-translation in the high school EFL classroom, and to what extent *yakudoku* instruction would be useful to prepare for the exam. To this end, I approached the analysis with an interpretive lens (Hubbard & Power, 1993; Patton, 1989), seeking first to generate a analysis of core categories. I carried out estimates of readability levels and vocabulary frequency, calculating Flesch-Kincaid Grade and Flesch Reading Ease measurements. I also looked at the kind of questions that appeared on the tests. Whereas Brown and Yamashita (1995) and Kikuchi (2006) asked whether the exams included translation, I asked what kind of translation was required and to what effect. Further, I looked at test items in terms of the quality of distractors and considered implicit assumptions about how English should be studied and the kind of students who would likely do well on the test.

Because of space limitations, I have not appended the exams to this paper, but they are available online along with correct answers and commentary provided by major *yobiko* preparation academies. In the following section, I present a general description of the exams, followed by analysis of key features and discussion of pedagogic implications.

Characteristics of the Exam

The exams were divided into five distinct sections (1–5). The first contained two reading passages (short and long), focusing on reading comprehension. Section 1–A asked to provide a written summary in Japanese (2–3 sentences), and Section 1–B contained four multiple-choice questions asking students to demonstrate global understanding of the reading. The second section focused on written production, with 2–A giving a 9–turn conversation in Japanese between a professor and his student, which students had to summarize in English. The 2008 test, however, used a prompt in English of an email exchange involving a homework assignment that students were to complete. In Section 2–B, there was a cartoon on the 2007 exam that students had to describe in English, and on the 2008 exam, a prompt in Japanese asked students to describe how transportation might change in the next 50 years. Section 3 involved a 30 minute listening component, and Section 4 dealt with grammar and translation: 4–A asked students to identify the incorrect word in five sentences embedded in one-paragraph reading passage, and 4–B required translation of three sentences, also embedded in a longer paragraph. Section 5 also focused on reading, with a short story about which students were asked comprehension questions.

With six extended reading passages, short translations and summaries, and a 30 minute listening component, the exam was difficult. The tests ran over 25 pages in length (excluding the listening script) with a total of 41 questions, 26 on the written component, 15 on the listening. Emphasis was placed on both global understanding and detailed comprehension of oral and written texts, coupled with a significant productive element involving summary, explanation, translation, and dictation. Topics covered a range of academic disciplines (e.g., anthropology, science, psychology) and questions

incorporated a variety of item types. For example, Section 1 on the 2007 exam involved a three-paragraph discussion of reader-response theory (1-A), and a ten-paragraph essay describing waste management systems in India (1-B). The corresponding section on the 2008 exam was composed of a short essay on the correlation between appearance and personality (1-A) and a discussion of the discovery of microbes in an underground diamond mine in South Africa (1-B). Other reading passages included a one-paragraph description of super volcanoes, a two-paragraph discussion of the human dimensions of medical treatment, a two paragraph reflection on multilingual identity, and a one paragraph discussion of the significance of email as opposed cell phones. The short stories in Section five revolved around themes of adolescent alienation and the responsibilities of growing up.

The listening component of both exams contained three oral passages, two of which were related. The 2007 test consisted of two essay-style lectures, the first on walking versus car culture, the second on anthropology in Africa. The third passage was a seminar-style discussion that followed the second lecture, involving a professor and three students. The 2008 test contained one lecture, a discussion of recent elections in Bhutan, followed by two segments of a planning commission meeting between architect, builder and city government representatives, discussing the design of a proposed apartment building. In both cases, questions for the oral passages were written in the exam booklet and distributed prior to being played, with passages played twice. There was no speaking component on either test.

The exam contained a striking assortment of question types, which ranged from fill-in the blank to reading comprehension, choice of appropriate words, and arrangement of sentences. Questions involved translation, dictation, and short-paragraph summary, in both English and

Japanese. The distribution of item types was the same on both tests. A clear majority of questions (25 of 41) were multiple-choice, but almost half of these (11 of 25) required high-level skills involving text manipulation, fill-in the blank, and word order, as well as deletion-insertion of text, and choice of appropriate title. Nearly all (13 of 14) of the traditional multiple choice items that required selecting the right answer from four alternatives were in the listening section. There were six questions that asked students to produce written text: two 2–3 sentence-length summaries of long reading passages (one English, one Japanese), a short paragraph description (in English) of the cartoon illustration, and a 1–2 sentence interpretation (in Japanese) of a short phrase (e.g., *go back home*). There were four items requiring translation: three sentence-length excerpts and a short phrase (*down the hall*). Productive items totaled nearly a quarter of the reading questions (6 of 26), and likely carried a good deal of weight, given the prominence and degree of attention required.

A second core characteristic of exams was the level of difficulty. The test was distinguished by a combination of time constraints, abstract discourse, range of topics, and variety of item types. It would pose an overwhelming challenge for all but the most talented and experienced high school students. Successful candidates would have an advanced language proficiency, with an ability to read across disciplines with speed and fluency, write with facility, and listen with a high degree of concentration and comprehension. At the same time, it is not clear what constituted a passing grade, since the exam was norm referenced for selection, and scores were likely aggregated with other subject tests. That is, a student who “passed” the Todai exam (i.e., was admitted to the university) might not have scored as high on English as other students who were not accepted. What is clear, however, is that the exam is very difficult to prepare for, in terms of specific language skills. Rather, the

exam presumed broad range of reading experience and all-around academic proficiency. A strong degree of intellectual curiosity would also prove a great help.

In terms of readability, the exam was roughly equivalent to an introductory level U.S. college textbook. Flesch-Kincaid readability scores, a general indication of the number of years of schooling required to comprehend a text, averaged 11.8 for the two reading passages in the first section of both 2007 and 2008 exams. The Flesch Reading Ease index, which provides a gauge of text difficulty on a scale of 1 to 100, with lower scores more difficult (professional academic journals score in the mid 30s to 40s, news magazines in the mid 50s), generated scores in the upper 40s. The indices confirm that the first two reading passages in Section 1 were compactly worded scholarly texts. Readability indices for the short stories in Section 5, however, were noticeably higher, indicating the descriptive style of an extended narrative and a greater degree of accessibility.

Other measures of vocabulary present a complementary picture. Analysis of word frequency demonstrated that both reading and listening passages were densely structured, with a significant percentage of academic and infrequent “off-list” vocabulary terms, defined as technical, field-specific, and/or unusual words that appear primarily in intermediate and advanced texts, “off” the basic list of the 2000 most frequently occurring vocabulary. While the percentage of items on the academic word list (AWL) attests to the academic tone of a passage, the percentage of off-list vocabulary reflects breadth and scope. For example, the 2007 passage on reader-response theory (Section 1A) had a high percentage of terms on the AWL (8.2%), and only a few off list tokens (1%). In contrast, the short story registered a relatively low readability score with few academic terms (1.5%), but a high percentage (10.8%) of unusual and likely unfamiliar

words, with proper names (e.g., *Pizza Delight, Michigan*), technical terms (e.g., *reflux lines, circus equipment*), and culturally specific colloquialisms (e.g., *early shift, broke up, memorial service*), comprising nearly 11% of total vocabulary.

Unusual words occurred throughout the test. The paragraph on multilingual identity (2008, 4–A) included the terms *integrated lifestyle, monolinguals, and intimate, intensely*. The reading selection (2007, 2–B) on waste-management systems, was similar, with a combined AWL and off-list total of 13.7%, a particularly high figure due to unfamiliar, unique terms related to the economics of rag pickers (e.g., *pheriwallahs*) and other aspects of recycling and waste disposal. The highest percentage of academic and unfamiliar vocabulary terms, however, occurred in the listening section. The lecture on Evans-Pritchard’s anthropological study of the Azande in central Africa (2007, 3–B), while not as difficult as the reading texts (1–A and 1–B) in terms of grade level and readability indices, contained a higher percentage of off-list technical words and specialized vocabulary. Phrases like *Azande belief systems, constructive role of inherited spiritual power and magic, and full confirmation of an oracle’s message*, illustrate the level of difficulty.

Cognitive Complexity

The test required little knowledge of abstract grammatical rules. Most questions, including multiple choice, required integrative thinking and high level cognitive skills of analysis, synthesis, and interpretation. Students were asked to approach reading passages with bottom-up and top-down strategies to grasp overall themes as well as specific information and ways in which particular linguistic elements conveyed broader meanings. Students needed to understand what fit into a passage and what was inappropriate, and they had to read strategically, skimming and searching for supporting details, as

well as look for causality and consequence.

To illustrate, I focus on the second reading passage (Section 1–B) that described waste management in India (2007) and democratic elections in Bhutan (2008). The four questions insured that students read the entire passage and were able to identify the basic theme. Further, the questions asked students to determine how pieces of the text fit together, on both paragraph and sentence levels. The first question asked students to insert an extracted sentence into the appropriate position. To do this, students had to grasp the discursive flow of the paragraph, deciding (on the 2007 test) whether the sentence established the introductory theme, illustrated basic categories, or summarized the paragraph. In the second question, students were asked to do the opposite: to delete (2007) or designate the least relevant sentence (2008) from one of the central paragraphs. This task also required a synthetic grasp of cohesion, recognizing discursive linkages between the overall theme and illustrative details. The third question involved arranging the order of four concluding paragraphs set apart from the passage. In this task, students had to discern the central argument and thematic connections among the paragraphs. Students also had to recognize the summative tone and reference to the overall thesis. The fourth question asked students to choose the most appropriate title, and the distracters employed vocabulary that would appeal to someone who had not read the passage with understanding and sensitivity.

Although the section's four questions were multiple choice, none demanded abstract knowledge of linguistic rules. Focus was placed uniformly on the content and broader social and economic issues raised by the text. The questions demanded integrated comprehension, and an ability to recognize how a text held together with explanation and illustration. Students also needed to generate a top-down conceptual frame to evaluate

the different interpretive possibilities presented in the questions. In this respect, student were asked to engage in authentic reading practices, skimming for details, making guesses about new and unknown vocabulary based on textual clues, and drawing on worldly knowledge and past reading in other subjects.

Written Production

The exam measured written proficiency in a variety of ways. Students had to write summaries, explanations, and interpretations, and productive questions called for analysis, synthesis and interpretation. Almost a third of the test (3 of 10 sections) required generating some form of written response. In particular, Section 2 was focused on composition. The 2007 exam (2–A) asked students to summarize (in English) a transcript of a Japanese conversation between a professor and a student having problems with English. On the 2008 exam, 2–A asked students to complete an email exchange involving the division of a homework assignment. The message was to be composed in the voice of a Japanese student, proposing an alternative to a potentially unfair division of labor. In Section 2–B (2007), there was a cartoon to describe in English. On the 2008 exam, 2–B gave students a prompt in Japanese, asking to explain (in English) how transportation will likely change in the next 50 years.

In other sections of the exam as well, students had to compose text, going beyond proposed alternatives on multiple choice items. In the reading section (1–A & 1–B), students summarized (in Japanese) the central thesis of the passage about reader response theory (2007) and the correlation between appearance and personality (2008). In response to the short story, students had to explain (in Japanese) what key phrases meant within the context of the narrative. In Section four, students had to translate three sentences, while

in the listening section, there were dictation questions that required an exact rendering of two expressions (2007) and a nine-word phrase (2008).

In each case, students had to articulate ideas in their own words, synthesizing and pulling together different points of view. The tasks required condensing a broader narrative (or in the case of the cartoon, constructing a narrative), distinguishing critical details, and utilizing key concepts from the readings. Students had to make connections, recognize contrasts, and in the case of the illustration, note the irony of the situation. When writing an English summary based on a Japanese prompt (and vice versa), students were not able to rely on vocabulary from the original text, but they had to generate their own phrasing. Importantly, the writing was not simply expression of opinion, but was based on text reference, with the goal to demonstrate familiarity with academic writing conventions.

Use of the L1

The exam presumed a high proficiency in Japanese. In fact, there was so much L1 that the test would be more accurately described as a measure of bilingual proficiency than English skill per se. All directions were in Japanese, and students had to read and interpret extended Japanese prompts. Many of the questions were also phrased in Japanese. On the 2007 exam, 10 of 11 questions that asked about the short story (Section 5) were in Japanese, two requiring translation. Students chose from Japanese alternatives and had to compose Japanese answers, involving summaries, explanation, and translation. Use of the L1 thus involved a range of discourse skills; without a strong productive and receptive command of L1 literacy, students would certainly not do well on the exam.

The first section of the test required a 2–3 sentence summary in Japanese of the English passage. Section 4 called for students to translate

three sentences embedded in a longer paragraph. In addition, three questions elicited the L1 in response to the short story on the 2007 exam: translating the phrase *down the hall*, and explaining the phrase *go back home*. Only one section (4-B) required a straightforward translation, a word-for-word rendering of three complex sentences taken from a paragraph on patient care in contemporary medical systems (2007) and recent effects of cell phones and email communication (2008). For example, the first sentence of 4-B (2007) read:

A phone without wires, so small that it fits in a pocket, containing such miracles of technology that one can call home from the back seat of a London taxi without thinking twice, is still just a phone.

The excerpt contained colloquial phrasing, such as *human touch* and *without thinking twice*, and students had to grasp not simply lexical and grammatical features but also overall meaning. Although students did not have to demonstrate a holistic comprehension of the overall paragraph, as they did in reading sections 1-A & 1-B, the excerpts for translation were located in the broader context of the paragraph, and in one case, there was specific instruction to clarify the external reference for the pronoun *it*. The translation tasks were thus situated within a broader bilingual recognition of the L1, and Japanese was assumed to be a natural and effective way to check comprehension, clarify ambiguity, and explain patterns of L2 meaning. At the same time, it is important to note that translation was a small part of the larger use of Japanese, which included reading as well as written summary, explanation, and interpretation.

Grammar in Context

Only one section of the test (4–A) required explicit grammaticality judgments, although this was limited to the 2007 exam. What is more, students were not asked to explain grammatical rules or the reasoning for the judgment. Rather, the task was to delete one syntactically inappropriate word from each of five sentences which were, in turn, embedded in a six-sentence paragraph. Thus the question read:

- 1) were they to explode, super volcanoes would [have] set off terrible earthquakes
- 2) they have done so [for] at least three times
- 3) looking for [an] information
- 4) pointed to some [of] answers
- 5) it will [ever] be possible

The words in brackets designated terms to be deleted, consistent with conditional, durative, non-count, partitive, and adverbial constructions, respectively. While it would take a high degree of grammatical sophistication to articulate the rules of syntax explaining why the particular words were inappropriate, students were not asked to give a rationale for the choice. The question located grammaticality judgments in the larger issue of comprehension, and students were able to make decisions based on intuition about the semantic content, or simply on how the words seemed to fit within the flow of the sentence. For students who did not have an intuitive feel for the language, explicit grammatical knowledge of the language may have helped, though not necessarily. In this respect, the exam involved attention to structural features of lexis and syntax, situated in the context of

reading. What is more, attention to grammaticality decreased in the 2008 exam, as the Section 4–A question dealt with semantic rather than syntactic appropriateness of words (a point that I discuss in more detail below).

The overall focus on context was limited, however, and it would be inaccurate to say that the exam was primarily concerned with reading comprehension. On the 2007 exam, for example, most questions in Section 5, the multi-layered short story, actually worked to preclude reading the text as narrative and considering its content and thematic elements. Instead, students were asked to read intensively, with a narrow focus on sentence-level discourse and vocabulary meaning. Most of the questions (6 of 11) required choosing the appropriate word or phrase that would complete a sentence, as in the following example (with choices in parentheses):

Years of living in Brooklyn ... had given Rebecca a new (perspective, sight, transformation, way).

Beyond choosing the correct answer, there was no need to consider the broader *content* of the protagonist's new perspective or the connection (or contrast) with the old. The task was simply to demonstrate understanding of word meaning. However, not all of the questions about the short story required an explicitly a narrow focus. Three of the eleven questions did demand a holistic grasp of narrative content. One question asked students to complete an interactional exchange between the story's protagonist and her childhood friend, choosing four of six possible phrases. The question was similar to the narrowly focused fill-in the blank questions, except that students had to deal with a discourse unit larger than the sentence, synthesizing conversation across multiple turns. The question, embedded within the larger narrative of the story, was the only conversational item on the exam.

The last question concerning the short story asked students to choose the protagonist's age, and the answer was located, not at the end but in the middle of the story, to which students would have to skim. Once students found the passage, they would have to calculate the protagonist's age based on the description of her stepfather, who

had moved in when Rebecca and Tracy were eleven. Their real rather had left three years earlier. Rebecca hadn't seen him for twenty years.

In this way, students could calculate that Rebecca was twenty eight.

Testwiseness

In spite of the overall focus on literacy and cognitively complex academic skills, both exams were seriously flawed by awkward construction and ambiguity. The diverse complexity of item types would certainly be confusing for untutored students, thus conflating language proficiency with test-taking preparation. More importantly, not a few questions demonstrated flaws in phrasing, with ambiguous distracters and more than one potentially right answer. Other items contained unfamiliar, culturally specific background knowledge and confusing double negative phrasing. Given the high-stakes nature of the exam and the potential for external variables such as familiarity with the format, it is clear that the test was not a reliable measure of EFL proficiency.

The listening component of the 2007 exam was particularly problematic. There were three subsections (a, b, c) with five questions each (two short dictation and thirteen multiple choice), and more than half (7 of 13) of the multiple choice items contained what seemed to be some form of compromising feature. Question A3, for example was phrased using

multiple negatives, asking for a reason *not* given for “why people *avoid* walking,” further compromised by a third negative distracter (*inefficient*). The student was thus presented with a negative sequence of three counterfactuals: *inefficiency was one reason not given for people to avoid walking*.¹ Question B4 also used a double negative (that an oracle might *fail* because someone might *not* have been aware of the problem). Question A5 offered two distracters, fifty yards and a quarter of a mile, which conflated cultural background with listening comprehension, since yards and miles are unfamiliar measurement terms in metric Japan.

Most questions in the listening section involved synonymic equivalents, but in some, ambiguous phrasing generated semantic mismatches. Question A4, for example, asked about the suburbanization of the mind, even though the lecture concerned San Francisco, which is an *urban* area. Question C5 asked what point of the professor’s lecture did one of the students in the seminar “accept” when he said “I see.” Not only did the phrase *I see* not necessarily imply acceptance, two of the incorrect distracters accurately captured the point made in the lecture immediately prior to the right answer. In sum, the phrasing of the question introduced both semantic and pragmatic ambiguity. Other questions related to the lecture also involved semantic uncertainty. The correct answer for question B2 stated that witchcraft was considered a “natural” ability, but “magical power” is in an important sense not “natural” at all. In question B5, the right answer stated that the unreliability of magic oracles helped Azande society run smoothly. The awkward phrasing (that unreliability helped stability) was compounded by the possibility of another right answer, that royalty protected witches, since

1) Question B4 on the 2008 exam also used double negatives in the distracters, asking why architects are *forbidden* to use a color *not approved* by the residents, or use *less than* 60% of a site’s area.

the lecture stated that royal animals were always used in the magic rituals.

Questions dealing with the short story (2007) also created problems with testwiseness. At four-pages, the section was the longest reading passage, with the largest number of questions (11). Because only two questions required reading the passage for understanding of its content, students who approached the text with the intention to guess the story would be making a potentially crucial error, given time limitations, thus giving an advantage to those who had prepared particular test taking strategies. Another serious problem involved the diversity of questions. There were seven item types among the eleven questions, which would prove difficult to anyone not already familiar with the format. Students had to switch from questions related to possible alternatives, filling-in a missing word, translating an underlined phrase, selecting an appropriate translation, choosing appropriate phrases to complete a conversation exchange, writing an explanation of a crucial phrase, and calculating the age of the protagonist. Another measure of testwiseness involved one particularly labyrinthine question type that required arranging a jumbled series of seven words into appropriate order to fill-in the blank, deleting one word from the seven (there was an extraneous word), then presenting the second and fifth words as the right answer, as in the following:

Even (_____) a teenager, Rebecca was embarrassed
(as been had more she than would)

The correct answer was *than* and *been*.² While the task required an

2) The correct answer was *than* and *been*: Even more **than** she had **been** as a teenager, Rebecca was embarrassed.

understanding of grammaticality, it also demanded familiarity with the format to decipher the instructions quickly and accurately, determining not only word order but also word number. Students unacquainted with the task would face a significant disadvantage. In the case of the listening section, however, no amount of preparation would unravel the awkward and ambiguous phrasing of the item distracters.

Linguistic Ideology

While there were serious problems with its construction, the Todai exam nevertheless worked to endorse particular social and linguistic values related to how English should be studied and used, both in the high school FL classroom and in Japanese society at large. Within this ideological frame, the test represents an implicit attempt to exert positive washback on HS EFL instruction. First, the test adopted a primarily communicative view of language use. Grammar was presented as a contextualized tool for making meaning, essential for close reading but secondary to the larger discursive engagement with the text. Use of both the L1 and translation was affirmed, but not as discreet rules and isolated expression, or as a substitute for a focus on the text. Instead, the L1 served a bilingual capacity, as a means to generate analysis, interpretation, and other cognitively demanding academic skills. The test also endorsed an international perspective that extended beyond the cultural sphere of the US or UK, incorporating distinctly Asian topics and contexts. Further, the test situated language study within other disciplines, including anthropology, social economics, and literary analysis. Additionally, students were positioned vis-à-vis English as active readers who pay careful attention to detail, but who also use practical strategies to guess new and unknown words (such as *maal*) based on the context and wider knowledge of the world, including knowledge developed in the L1.

Second, the test presented an explicit model of how to use English in the language classroom. For example, in the second reading selection (2–A) on the 2007 exam, one of the students directly asked the professor how to study English, and the professor’s advice constitutes a meta-statement about culturally appropriate attitudes. While the student articulates a popular view, that exposure to authentic English improves listening comprehension, the professor, speaking with the authority of the nation’s top-ranked university, rejects this passive approach. He asserts that contact with English, including native speakers and widely available electronic media, will not in and of itself develop language proficiency. Only concerted effort and dedicated study will allow students to master the foreign language. It is the individual effort to use books and dictionaries and other resources the teacher recommends, that the student – under the teacher’s guidance and tutelage – will succeed in being admitted to Tokyo University. This narrative reproduces core cultural values of Japanese education, involving effort, dedication, and respect for established authority and experience.

Although straightforward teacher-fronted lectures and written text materials are endorsed as the primary means of knowledge acquisition, the seminar discussion of anthropological theory (3–C, 2007 exam) also fills an ideological role, demonstrating a model for participation in group discussion, often presented as the prototypic activity of authentic English use. The seminar, composed of the professor and three students, Joe, Don, and Rumiko, demonstrates a key narrative about intercultural communication, locating Japanese nonnative speakers within the larger English speaking world. The seminar portrays students who are attentive and who ask questions that are relevant and based on information presented in the lecture. What is more, the students are neither shy nor reluctant to speak out and express personal opinion in exchanges with the professor

and classmates. Students ask for clarification, make connections, and use illustrative examples based on their experience. Through their participation, they also recognize the position of the group, another core cultural value, incorporating everyone's contribution and point of view. For example, one of the students, Don, states in the discussion that

surely we can study the same thing for very different reasons. Joe thinks we should study something to improve our own way of doing things. Rumiko wants to understand the way they think. Professor Shelby wants to explain what holds their society together and makes it work. Why can't we say that these are all good reasons?

The professor approves of Don's epistemic synthesis and his inclusion of the disparate points of view expressed by other members of the group.

In particular, Rumiko's participation in the seminar symbolizes the ideal way to use English on an international stage. As the only Japanese student, Rumiko is outnumbered by the native speakers who surround her, but she is neither shy nor intimidated. She articulates her ideas in a reasoned but assertive way and does not back down when challenged. She is a Japanese who can indeed say no, clearly and unequivocally. When her classmate Joe rejects the example she presented, Rumiko replies "I don't agree" and backs up her assertion with clear explanation. When Joe continues to insist on his interpretation, Rumiko does not avoid confrontation. She responds with an equally sharp question, implying both negative assessment and broad mindedness: "So you think we have nothing to learn from them?" When Joe employs a similarly framed negative tag question, Rumiko explicitly dismisses his point:

- Joe: they'd be more interesting if they were right, wouldn't they?
- Rumiko: Not at all. It's more interesting to see how their beliefs can make sense to them, even if they are wrong.

Going tit for tat with use of emphatic intonation, Rumiko not only wins the exchange, she has the last word, as at this point, the professor shifts the discussion to Don's synthesis, thus mediating the disagreement and affirming the core value of group cohesion, leaving no one left out. In this respect, Rumiko embodies the ideal Japanese English speaker in her engaged participation, reasoned argumentation, and confident speakership.

Changes in the 2008 Test

The 2008 exam bears a strong resemblance in terms of overall length and structure to the previous year's test. Each consisted of five sections of broadly comparable length and readability, with only minor variation. On both tests, for example, Section 1A asked for a Japanese summary of a two paragraph reading passage that was highly academic in tone. While the 2008 passage was slightly shorter, with a fewer vocabulary terms from the academic word list, it had more off list words of unfamiliar vocabulary. The reading passages in the second section (1-B) measured roughly the same in terms of readability level. Questions were similarly structured and phrased, with clear focus on discourse organization and flow. There was also a strong focus on content. While the 2008 passage on 1-B was considerably longer (901 words to 708), grade level and word familiarity was largely equivalent. The short stories that end both tests were likewise similar in terms of tone and difficulty level. The 2007 version was noticeably longer (1249 vs. 975 words), but the 2008 version involved contrastive points of view and was thus perhaps moderately more difficult to understand.

Within the fundamentally similar structure and orientation, however, there were several important differences. Perhaps most noticeably, there was more English in the 2008 questions. The prompt on Section 2A on the 2007 version, a 9–turn exchange in Japanese, became an email exchange in English that students were to complete by adding explanation. While both questions required examinees to compose text in English, the 2008 prompt was itself in English.

There was also a subtle shift to a more academic tone on the 2008 exam. While Section 2–B of the 2007 exam presented the humorous cartoon of a UFO, the 2008 version involved a verbal prompt, where students were asked to imagine changes in transportation. Both prompts required a short essay written in English involving creative imagination and expressive skill, but the 2008 version allowed students to touch upon critical issues related to environmental protection and social organization.

There were other slight differences. The 2008 listening section was shorter in length, and the excerpts had a more interactional quality. There was one academic lecture instead of two. Two of the listening excerpts were, like the 2007 seminar, group discussion among multiple speakers, but the participants spoke more conversationally, with less information packed into each turn. The second excerpt was noticeably less complex, with easier readability indices. While there was less of a scholastic tone, the speakers adopted different perspectives on the shared topic. In other words, in terms of overall readability, the 2007 exam was slightly more academic, the 2008 exam somewhat more interactional.

Discussion

In conclusion, it is fair to say that the Todai English exam does not fit the traditional image of Japanese university entrance exams focused on discreet

skills of grammar rules and decontextualized vocabulary. But neither does it focus exclusively on reading comprehension, and it would be incorrect to describe the exam as a measure of reading proficiency. The test would be better understood in terms of bilingual literacy. The wide variety of questions demanded an advanced academic ability integrating reading skills with both written production and listening comprehension. Students had to interpret a range of academic discourse with fluency and attention to detail, demonstrating synthetic understanding of complex texts, within a largely (but not exclusively) multiple choice format. Reading passages required a holistic grasp of textual ideas, as well as detailed understanding of sentence-level discourse features, word meaning, and (to a lesser extent) grammatical accuracy. Writing tasks called for summarizing the content of reading passages, while the extended listening component presented students with substantive and detailed academic lectures. The test included a significant amount of Japanese, although word-for-word translation comprised only a small part of the overall exam. The test also illustrated implicit meta-narratives about the study of English, suggesting that the ideal student of the foreign language is willing to articulate ideas, engage in discussion, and stand by one's opinions.

In spite of the overall orientation to language as discourse in context, however, the exam was seriously flawed as an objective measurement of FL proficiency. The exaggerated complexity of item types was confusing and rewarded familiarity with the format. Students who have not undergone specialized training are not likely to do well on the exam. More importantly, a number of questions, particularly on the listening component, contained serious ambiguities and unfamiliar, contradictory distracters, which rendered the test unreliable and ultimately invalid. As a result, the analysis I have presented in this paper suggests that those who criticize Japanese entrance

exams for lack of psychometric rigor and quality control are correct to challenge the fairness of university entrance exams and even the legitimacy of a system where tests are constructed in-house without pilot study or external review.

The problem, however, is not simply about building a better test. As Brown (2002) and others have pointed out, it is fundamentally unfair to make high stakes admission decisions based on the results of a single measurement. Nevertheless, the notion of the test as an objective evaluation uncorrupted by preferential treatment or partisan influence, taps into deeply held cultural values (Zeng, 1998), and the idea of impersonal exams administered without favor to all applicants retains strong institutional and public support. What is more, universities find it difficult if not impossible to fundamentally change the way the exams are constructed, given economic and administrative constraints imposed by the Ministry of Education. Effectively, the hands of universities are tied, and without more imaginative guidance and coordination, individual universities have neither the academic nor financial means to change the system.

It is certainly true, then, that from a macro perspective, the exam imposes a unfair model of assessment that has deleterious social effects. From the standpoint of teachers who work with students in the HS classroom, however, there is little room to ignore the test or wait for the Ministry to introduce a more informed policy. The test retains its influence over the HS curriculum in spite of flaws in its construction. Schools are still inevitably viewed in terms of the exams, and parents as well as students still rate success on exams as paramount, while teachers are still judged by how well they prepare – or perhaps how they are *perceived* to prepare students for the exams. The question is, however, what kind of preparation is actually effective.

On the one hand, analysis of the Todai test does not substantiate the view that the preparing for the entrance exam precludes communicative pedagogy. To the extent that high school EFL teachers implement traditional *yakudoku* instruction, based on explanations about grammar rules and intensive, word-for-word translation, it is clear that students will not be preparing for the exam as it is constructed, but rather engaging in pedagogic practice with which the teacher may be more familiar (Watanabe, 2004) or established school culture may consider more effective (Sato, 2002). On the other hand, Oral Communication classes where students practice personal greetings, engaging in tasks like asking directions and exchanging opinions with native speaker ALTs will likely be of little help preparing for the exam or giving students the linguistic and interpretive skills to address academic texts, either on the test or in “authentic” contexts. Consequently, the exam affirms *neither* the orthodox model of intensive grammar-centered *yakudoku* instruction *nor* the conversational syllabus of oral communication. Instead, analysis of the exam suggests that what is paramount is an academic facility with literacy, the ability to read serious texts critically with attention to accuracy and detail, coupled with productive proficiency in the L1 as well as the L2. On a practical level, the exam suggests that, in preparation, students engage in both intensive *and* extensive reading, involving attention to form and critical interpretation, in oral discussion and writing. In sum, the exam calls for a text-centered classroom, where the teacher uses Japanese *judiciously* as a pragmatic tool to explain grammatical structure and check comprehension, but also uses English to read as well as talk and write about texts, with the goal to deepen understanding of content, ideas, and interpretations.

The Todai exam is admittedly difficult, and as the most selective institution in the nation, the test sets the bar higher than most if not all

Japanese students can jump. At the same time, the exam reflects recent shifts in the Japanese high school EFL curriculum, particularly since introduction of the Ministry's *Action Plan*, establishing "super" language programs that provide an expanded range of English courses and opportunities for accredited study abroad (Butler & Iino, 2005). The exam also illustrates the increasing globalization of Japanese society, with expanded access to international travel and residence in English-speaking societies, at least for some segments of the population. As a result, more and more HS students are developing the proficiencies that are tested by the Todai exam. Perhaps more importantly, the test challenges assumptions about effective pedagogy because it goes beyond the interactionist paradigm that reduces the term "communicative" to spoken conversation (van Patten, 1998) privileging face to face interaction with native speakers (Cook, 1999). Instead, the exam endorses what Kramsch (2006) calls "symbolic" competence. As such, the Todai test is consistent with emerging conceptions of FL education as not simply about communication but rather, in Kramsch's terms, understanding the conditions that make communication possible, not only how to achieve a communicative task, but also "the nature and the purpose of the task itself" (p. 250). Doing this, Kramsch points out, requires critical engagement with text-based literacy and a "literary imagination" to locate meaning in social, cultural and historical understandings.

Shohamy (2001) has stated that the "strength of tests lies not only [in] their technical quality but in their use in social and political dimensions" (p. 162). In this paper, my analysis suggests that while the Todai exam is without doubt a seriously flawed instrument in terms of construction, its influence on EFL pedagogy is not nearly as negative as often portrayed, especially by those who assume that the role of entrance exams (and other high stakes tests like TOEIC and TOEFL) is to make the FL classroom more

communicative. What the Todai exam demonstrates is not a dated model of grammar instruction but an ideal of academic study and challenging engagement with an increasingly complex and multicultural global world in English. The Todai exam (and one might argue the same for other universities, including Keio) is located within this academic vision of FL education, suggesting grounded practice in the classroom and an integrated view of what it means to study a foreign language.

References

- Akiyama, T. (2003). Assessing speaking in Japanese junior high schools: Issues for the senior high school entrance examinations. *SHIKEN: JLT Testing & Evaluation SIG Newsletter*, 7 (2), 2–8.
- Amano, I. & G.S. Poole. (2005). The Japanese university in crisis. *Higher Education*, 50, 685–711.
- Brown, J.D. (2002). English language entrance examinations: A progress report. In A.S. Mackenzie & T. Newfields (Eds.), *Curriculum Innovation, testing and evaluation: Proceedings of the 1st annual JALT pan-SIG conference* (pp. 95–106). Tokyo: JALT. Online at <http://jalt.org/pansig/2002/HTML/Brown.htm>
- Brown, J.D. (2000). University entrance examinations: Strategies for creating positive washback on English language teaching in Japan. *Shiken: JALT Testing & Evaluation SIG Newsletter*, 3 (2), 2–7. Online at http://jalt.org/test/bro_5.htm
- Brown, J. D., & Yamashita, S.O. (1995). English language entrance examinations at Japanese universities: What do we know about them? *JALT Journal*, 17 (1), 7–30.
- Butler, Y.G. & Iino, M. (2005). Current Japanese reforms in English language education: The 2003 “Action Plan.” *Language Policy*, 4, 25–45.
- Chapman, D.W. & Synder, C.W. Jr. (2000). Can high stakes national testing improve instruction: Reexamining conventional wisdom. *International Journal of Educational Development*, 20, 457–474.
- Cheng, L. (1998). Impact of a public English examination change on students’ perceptions and attitudes toward their English learning. *Studies in Educational Evaluation*, 24 (3), 279–301.

- Cook, V. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly*, 33 (2), 185–209.
- Cornwell, S., Simon-Maeda, A. & Churchill, E. (2007). Selected research on second-language teaching and acquisition published in Japan in the years 2000–2006. *Language Teaching*, 40, 119–134.
- Goodman, R. (2003). The changing perception and status of Japan's returnee children (*kikokushijo*). In R. Goodman (Ed.), *Global Japan: The experience of Japan's new immigrant and overseas communities* (pp. 175–194). London: Routledge Curzon.
- Gorsuch, G. (2001) Japanese EFL teachers' perceptions of communicative, audiolingual and yakudoku activities: The plan versus the reality. *Education Policy Analysis Archives*, 9 (10). Online at <http://epaa.asu.edu/epaa/v9n10.html>
- Guest M. (2005). Teaching progressively ... for the *senka shiken*! In K. Bradford-Watts, C. Ikeguchi, & M. Swanson (Eds.), *JALT 2005 conference proceedings* (pp. 1182–1191). Tokyo: JALT.
- Guest, M. (2000). "But I have to teach grammar!": An analysis of the role "grammar" plays in Japanese university English entrance examinations. *The Language Teacher*, 24 (11), 23–28.
- Hiramatsu, S. (2005). Contexts and policy reform: A case study of EFL teaching in a high school in Japan. In D. Tedick (Ed.), *Second language teacher education: International perspectives* (pp. 113–134). Mahwah, NJ: Lawrence Erlbaum.
- Hooghart, A.M. (2006). Educational reform in Japan and its influence on teacher's work. *International Journal of Educational Research*, 45, 290–301.
- Hubbard, R.S. & Power, B.M. (1993). *The art of classroom inquiry: A handbook for teacher-researchers*. Portsmouth, N.H.: Heinemann.
- Kikuchi, K. (2006) Revisiting English entrance examinations at Japanese universities after a decade. *Jalt Journal*, 28 (1), 77–96.
- Kramsch, C. (2006). From communicative competence to symbolic competence. *The Modern Language Journal*, 90 (2), 249–252.
- Leonard, T.J. (1998). Japanese university entrance examinations: An interview with Dr. J.D. Brown. *The Language Teacher*, 83 (3), 303–318. Online at <http://www.jalt-publications.org/tlt/files/98/mar/leonard.html>
- McNamara, T. & Roever, C. (2006). *Language testing: The social dimension*. Malden,

MA: Blackwell.

- Mulvey, B. (1999). A myth of influence: Japanese university entrance exams and their effect on junior and senior high school reading pedagogy. *JALT Journal*, 21(1), 125–142.
- Mulvey, B. (2001). The role and influence of Japan's university entrance exams: A reassessment. *The Language Teacher*, 25(7), 11–17.
- Murphey, T. (2006). Practical reasons for praising entrance exams. *The Japan Times*, Tuesday, Feb. 7, 2006. Online at <http://search.japantimes.co.jp/cgi-bin/ek20060207a1.html>
- Murphey, T. (2004). Participation, (dis-)identification, and Japanese university entrance exams. *TESOL Quarterly*, 38 (4), 700–710.
- Murphey, T. & Sato, K. (2003). Deconstructing the making of Japanese university entrance exams. In C. A. Coombe & N. J. Hubley (Eds.), *Assessment practices* (p. 41–48). Alexandria, VA: TESOL.
- Newfields, T. (2005). Voices in the field: An interview with Yoshinori Watanabe. *JALT Testing & Evaluation SIG Newsletter*, 9 (1), 5–7. Online at http://jalt.org/test/wat_new.htm
- Patton, M. (1990). *Qualitative evaluation and research methods*. Newbury Park: Sage.
- Poole, G. (2003). Assessing Japan's institutional entrance requirements. *Asian EFL Journal*, 5 (1). Online at <http://www.asian-efl-journal.com/march03.sub5a.php>
- Qi, L. (2005) Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing*, 22, 142–173.
- Rohlen, T.P. (1983). *Japan's high schools*. Berkeley: University of California Press.
- Sage, K. & Tanaka, N. (2006). So what are we listening for? A comparison of the English listening constructs in the Japanese National Centre Test and TOEFL. In T. Newfields, I. Gledall, M. Kawate-Mierzejewska, Y. Ishida, M. Chapman, & P. Ross. (Eds.) *Authentic communication: Proceedings of the 5th annual JALT Pan-SIG conference* (pp. 74–98). Online at <http://www.jalt.org/pansig/2006/HTML/SageTanaka.htm>
- Sato, K. (2002). Practical understandings of communicative language teaching and teacher development. In S. Savignon (Ed.), *Interpreting communicative language teaching* (pp. 41–81). New Haven: Yale University Press.

- Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. New York : Longman.
- Shohamy, E. (2005). The power of tests over teachers: The power of teachers over tests. In D.J. Tedick (Ed.), *Second language teacher education: International perspectives* (pp. 101–111). Mahwah, NJ: Erlbaum.
- Shohamy, E. (2007). Language tests as language policy tools. *Assessment in Education: Principles, Policy & Practice*, 14 (1), 117–130.
- VanPatton, B. (1998). Perceptions of and perspectives on the term “communicative.” *Hispania*, 81 (4), 925–932.
- Wall, D. (2000). The impact of high-stakes testing on teaching and learning: Can this be predicted or controlled? *System*, 28, 499–509.
- Watanabe, Y. (2004). Teacher factors mediating washback. In L. Cheng & Y. Watanabe, Y. with A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 129–146). Mahwah, NJ: Lawrence Erlbaum.
- Zeng, K. (1998). *Dragon gate: Competitive examinations and their consequences*. London: Cassell.