University Entrance Exams Renewal
—From Translation to Communication—

By
Jon ROWBERRY*, Yoko KINOSHITA *and* Craig YAMAMOTO*****

Abstract

本論文は日本の私立大学における大学入試問題改正プロジェクトの経過報告である。入試問題の更新は、コミュニカティブ英語教育と実際の言語使用状況に基づいた教材活用の促進を推薦する文部科学省の新学習指導要領の実施に合わせている。本プロジェクトの第一段階として開発チームは試験背景にある多数の実用的制約を配慮しながら、試験目的に最も適合させるべく実行可能な一連の試験仕様を作成した。 それに続き、実地試験済みの二種類の試験形式見本に対し試験課題と項目を開発した。試験結果の記述統計量では試験項目の作成は大部分において成功したことを示したが、試験の信頼性、妥当性をさらに高めるにはまだ多くの取り組みが必要であることも示した。

This paper is an interim report on a project to revise the English entrance examinations for a private university in Japan. The exam renewal coincides with the implementation of MEXT's revised high school curriculum guidelines which endorse communicative language teaching and promote the use of materials based on actual language use situations. During the first stage of the project the development team attempted to establish a workable set of test specifications in order to best meet the objectives of the examinations within the considerable constraints of the testing context. Subsequently, the team developed tasks and items for two prototype exam forms which were piloted and then field-tested. Descriptive statistics of the trial results suggested that the team had been largely successful in generating items but that considerable work still needs to be done in order to further enhance the reliability and validity of the test.

Key Words: language testing, MEXT, communicative language teaching, washback, reliability, validity

1. University English Entrance Exams in Japan

Traditionally in Japan the vast majority of university admissions are determined on the basis

*Professor, Sojo International Learning Center
**Assistant Professor, Sojo International Learning Center
***Senior Assistant Professor, Sojo International Learning Center

of results in examinations in the nationally administered Center Test as well as in locally administered examinations set by each institution. Since they are used as the basis for decisions about which students to accept or reject, these exams should be considered high stakes tests. Therefore, it is incumbent upon the committees responsible for creating them to do the best job that they possibly can, despite the numerous constraints and challenges of the testing context.

Although the low birth-rate and consequent surfeit of places has brought some flexibility to entrance procedures at many Japanese universities, the entrance exam remains the linchpin of the admissions system and this is unlikely to change for the foreseeable future. Moreover, the status of English, in particular, has been much enhanced by recent initiatives such as the Project for Promotion of Global Human Resource Development aimed at "improving Japan's global competitiveness and enhancing the ties between nations", the Global 30 project to promote the internationalization of Japan's universities, and the Re-Inventing Japan Project which aims "to foster human resources capable of being globally active" (MEXT, 2014). Consequently, English ability is increasingly viewed as an important attribute not only for admission to language or linguistics departments but also to business, science, technology, and other programs.

Meanwhile, the high school English curriculum has undergone significant reform in order to replace traditional grammar translation approaches with more communicative teaching methods, with course of study guidelines now requiring classes to be taught principally in English (MEXT 2011). More recently, the announcement in January 2014, subsequent to the selection of Tokyo as the host city for the 2020 Olympics, of the English Education Reform Plan Corresponding to Globalisation (MEXT 2014), has served to strengthen and accelerate this shift towards a communicative curriculum.

Currently, however, the majority of university English entrance exams seem poorly suited to measuring candidates' English proficiency in the communicative contexts apparently desired by MEXT. If specifications exist at all for these tests they are rarely shared beyond the committees responsible for creating the items, and even within these committees there is considerable confusion over the true purpose of the test. Cook (2013a) found little consensus amongst those involved in test creation about whether their tests aimed to assess candidates' English competencies, school achievement, general suitability for university study, or even whether the tests served purposes entirely unrelated to pedagogy, such as enhancing the reputation of the institution or generating additional revenue. Meanwhile, the quality of the tests themselves has frequently been called into question. Test creation committees typically have no formal training in assessment (Aspinall, 2005) and test items do not undergo piloting, often because of fears over test security (Leonard, 1998). Moreover, items tend to be overly reliant on archaic language and highly complex reading passages (Brown and Yamashita, 1995; Kikuchi 2006), and target receptive skills at the expense of productive skills (Guest, 2008). Unsurprisingly, therefore, significant concern has been expressed about the negative washback effects of English entrance examinations on the school curriculum and the materials and methodologies employed in schools (see, for example, Brown 1995).

However, it should be noted that many of the criticisms levelled at university entrance tests come from expatriate teachers who may have limited understanding of the Japanese perspective or working practices (Stapleton, 1996). In fact, Mulvey (1999) and Stout (2003) claim that the negative influence of entrance examinations on high school pedagogy has been overstated, while Guest (2008) suggests that some of the open-ended task-types utilized in many of the tests compare favourably to the discrete point multiple-choice items commonly used in high stakes standardized

testing. Moreover, despite their limited formal training, members of test development committees are often able to draw on many years of experience in creating items and grading tests within their own specific testing contexts, and tend to be well attuned to the characteristics of test-takers. Nonetheless, it seems clear that in many institutions changes to the English entrance exams could impact positively on test-takers and test-makers alike.

2. Reform of the University English Examinations

This paper will describe a project currently taking place at a small, private university in Japan to reform the English entrance examinations. The aims of the project are to draw up formal specifications for the tests before going on to create and trial assessment tasks which best meet those specifications within the considerable constraints of the testing context. By the end of the three year project it is hoped that prototype test papers will be available which can serve as a model for the creation of subsequent tests. These prototype tests can also serve, along with accompanying documentation, to inform the various stakeholders about the rationale for the various task types utilised thereby initiating an ongoing process of test validation.

The current test

Six iterations of the English examination are drawn up each year taken by a total of 1000 to 1500 applicants in one of two versions. One form of the examination (Test A) is developed for the Faculty of Pharmaceutical Science. There are two iterations and, since entrance to this faculty is highly competitive and it is a compulsory paper, it targets a relatively high level. The other test form (Test B) is for the use of other faculties within the university, the majority of which are in science or engineering disciplines. Four iterations are required for this test but since it is just one of a number of optional papers, it is taken by relatively few applicants and targets a somewhat lower ability range than Test A.

The test has traditionally been created by English teaching faculty members, all but one of whom was Japanese. In recent years, however, the number of expatriate (native-English speaker) teachers on the test development committee has increased such that the committee is now chaired by an expatriate professor and comprising a more or less even balance of local (Japanese) and expatriate teachers. Although the expatriate members of the group have some training and experience in language testing none of the committee members could be described as assessment specialists. However, the desire to reform the test is evident at all levels of university management and all members of the group are enthusiastic about (or at least open to) revising the test in order to enhance its validity and reliability.

The constraints of the testing context

A number of factors, largely beyond the control of the test development committee, serve as severe constraints in developing the English examinations. Most of these are likely to be shared with comparable institutions. Such constraints are encountered at all stages of the development process. For example, the need to administer the tests in multiple locations over a number of iterations means that it is not practical to deliver the test electronically, so it must be realised in a traditional paper-based format. Moreover, the large number of test takers and lack of appropriately skilled administrators rule out a speaking section, while the lack of resources rules out a listening section. Consequently, it was impossible to realise the test creation committee's first preference of designing a four skills test.

In relation to test evaluation, security concerns mean that it is impossible to trial test items and therefore to obtain reliability estimates in advance of a test. Then even after the test it has not been

possible to run statistical analyses because responses have not been digitized. Furthermore, Japanese institutional culture seems somewhat resistant to reform and innovation with the fear of making a trivial mistake, such as a minor spelling or punctuation error, appearing to trump the fear of making a bad test. Moreover, in the absence of formal published test specifications, past test papers, which pass quickly into the public domain in order to help prospective applicants prepare for future test iterations, become the de facto specifications. Since these papers are scrutinised thoroughly by high schools and cram schools, test development committees hesitate to innovate for fear of being accused of being unfair or unreasonable (Cook 2013b) preferring to stick with item types and content areas that have been used in the past.

Research questions

Acknowledging these constraints and the consequent challenges to reforming the test, the following research questions were devised for the project:

1. What test specifications are appropriate for the university's English entrance examinations for use from 2016 onwards?
2. What task types are most appropriate for meeting these test specifications?
3. Is it possible to actualise the revised test in multiple-choice format?

Guiding principles

After thorough analysis of the current tests and discussions amongst the test creation committee and other stakeholders, it was agreed that as far as possible the reformed examinations should:

- avoid reliance on discrete decontextualized items targeting specific lexis or grammar points but instead present items within clear communicative contexts;
- draw on a wide range of texts as source material including, but not limited to, letters, emails, reports, advertisements, instructions, descriptions, narratives and conversations, as well as more traditional information texts such as articles;
- require students to demonstrate an ability to apply their learning in a variety of sub-skills such as inferring meaning from context, scanning texts for specific information, synthesising information from multiple sources, summarising, etc.;
- measure productive as well as receptive skills;
- use multiple-choice items where possible in order to facilitate fair and efficient scoring as well as to facilitate post-test statistical analysis of tasks and items.

Given the constraints outlined above, these goals represent aspirations rather than concrete specifications; nonetheless they still served as a set of guiding principles and even if only partially realised would represent a significant enhancement to the quality of the test.

It was also recognised that in order for the reform to be successful it would be vital to communicate effectively to stakeholders within and beyond the university a clear rationale and explanation of the changes. By communicating goals and methods clearly to high school students, as well as their teachers and parents, it may be possible to influence the ways in which they prepare for the test, or even their whole approach to English language study, thereby creating positive washback (Messick 1996).

3. Methodology

A team of four was convened in order to develop two prototype tests. Having established the guiding principles the team was able to turn its attention to the challenging task of developing appropriate items. The team met regularly in order to discuss potential task types and then worked individually or in pairs to create items. Some existing tasks were redesigned to conform better to the guiding

principles, and some entirely new tasks were developed. The resultant items were collated into the two test forms, which were substantially revised following discussion and evaluation until the final trial versions of the tests were drawn up. A preliminary pilot was conducted on the higher-level test form by administering it to a group of 25 students who were estimated to have a level of English proficiency similar to that of the actual test takers. These test results were analysed and following discussions with the test development team necessary revisions were made to both tests. Full piloting took place in December 2014. 103 students took Test A and 91 took Test B.

Test Format

Test A consisted of a 30-item multiple-choice section worth 70% of the total marks and a writing section worth 30%. The multiple-choice section comprised of 10 discrete items targeting students' knowledge of lexis and grammar within short samples of spoken text and an additional 20 items based on four extended texts. The texts used were:

- a conversation in which two people discuss an incident that occurred recently in their neighbourhood;
- a short news article;
- a letter promoting a new medication; and
- a longer article about the health impacts of space travel.

In the writing section test-takers were asked to explain whether or not they would like to travel to outer space, providing reasons for their opinion. They were advised to spend at least 15 of the 60 minutes available on this section and to write at least 100 words.

Test B comprised 36 multiple-choice items divided into 4 main sections. The first section focuses on discrete grammar and vocabulary items largely modelled on items from the current version of the test. The second section presents a series of short conversations accompanied by items in Japanese designed to probe test-taker's pragmatic competence. The third section featured questions based on four short narrative texts, in which test-takers were required to compare and contrast the information given in the texts. The final section consisted of a 500-word article with accompanying comprehension questions.

In both tests, the items based on readings targeted a range of skills. These included the ability to find specific information, understand particular words and phrases, infer the meaning of unknown words and phrases from context, recognise the purpose or genre of a text, select pragmatically appropriate language for a given context, and synthesise information from a variety of sources.

Reliability and validity

Following the trials, the scores for the multiple-choice sections were analysed to generate the mean, standard deviation, standard error of measurement, and reliability coefficient (using Cronbach's alpha). Reliability refers to the internal consistency of test scores or the extent to which the same test-taker might be expected to get the same score if he or she were to take the same test in two different (hypothetical) instances. In addition, individual items were analysed by calculating the item facility (IF) to investigate the level of difficulty of each item, and the point biserial correlation coefficient (r pbi) to determine the extent to which each item could discriminate between the higher and lower ability test-takers. A distractor analysis was also conducted in order to evaluate how successful the development team had been in coming up with convincing distractors.

It is also important to examine the test in terms of validity. However, whilst reliability can be determined statistically immediately following the test trial, judgements regarding validity are more descriptive in nature and need to account not only for the test itself (internal validity) but also wider factors relating to the context in which the test is conducted. Consequently, test validation should be seen as an ongoing process (Messick, 1989) rather

than a one-time snapshot. The efforts of the group described above to draw up test specifications at least provide a starting point from which a validity argument might begin to be constructed.

Finally, it should be noted that practicality, defined by Green as "the extent to which the commitment of resources to a system of assessment is justified by the benefits it brings" (2014, p. 60), is an essential quality of all assessments. Therefore, the group also needed to carefully consider not only how well each task-type functioned but also whether it would be feasible to create items of sufficient quality and in sufficient quantities given the limited availability of time and the other constraints outlined above.

4. Results and discussion

Overall test performance

The goal for high stakes standardized tests is to achieve a reliability coefficient of over 0.90. However, given the constraints of the testing context it would be unreasonable to expect a Japanese university entrance examination to achieve a reliability score this high so the team set itself the more modest goal of a score of, or close to, 0.80. For both the trial tests, the reliability coefficient falls somewhat short of this but within the 0.70 to 0.80 regarded as good for classroom tests (see tables 1 and 2). In the case of Test A, the lower score is probably due to the smaller number of items (30 as opposed to 36). However, in both cases, the test reliability seems at least satisfactory for a first attempt and the goal of achieving a coefficient of at least .80 would seem to be attainable following revision and the inclusion of additional items[1].

Table 1. Mean, standard deviation, standard error of measurement and reliability scores for Test A

	Raw	%
Mean	14.81	49.37
SD	4.64	15.47
SEM	2.47	8.23
Reliability（Cronbach's Alpha） 0.72		

Table 2. Mean, standard deviation, standard error of measurement and reliability scores for Test B

	Raw	%
Mean	19.98	55.49
SD	5.44	15.12
SEM	2.59	7.18
Reliability（Cronbach's Alpha） 0.77		

In the case of Test A, the multiple-choice section comprised 70% of the total score with the remaining 30% based on a writing task. This task was seen as a vital part of the test because, as well as mitigating the reliability concerns outlined above, it also forces the test-taker to demonstrate an ability to produce language rather than to merely understand it. Of course, the inclusion of a writing section raises problems of its own, particularly in respect of scoring. In order to make the scoring system as fair as possible, a writing rubric was drawn up consisting of three bands: task achievement, organisation and language. Following the pre-trial test iteration, all four members of the group, along with a colleague experienced in the assessment of English writing, were asked to score a number of responses using the writing rubric. The group then met to compare and discuss their ratings and to make any necessary revisions to the rubric

[1] *A note of caution should be sounded here in relation to the interpretation of test scores. Even highly reliable standardized tests typically report a standard error of measurement of around 5%. Following the general rule of thumb that to predict the amount of change which can be expected in individual test scores we should multiply the standard error of measurement by 1.5, a score of, say, 60% actually represents any score within the range of 52.5% to 67.5%. In the case of the trial test scores reported above the standard error of measurement is even greater this is also likely to be the case for the majority of English entrance tests currently employed at universities across Japan. It is extremely important, therefore, that those responsible for admissions decisions are made aware of these limitations and do not use a single test score as the sole means of assessing a candidate's suitability.*

itself. Then for the full pilot, responses were copied and distributed to all five raters with each script being graded by at least two different people. The ratings given were then analysed to determine rater leniency and the scores generated were adjusted as appropriate. It was found that four of the five raters, rated consistently and comparably but one rater was significantly more lenient than the other four. However, since the scores of this rater were consistent it was simple to adjust the scores. The fact that all five raters were able to consistently apply the rubric was a positive finding and the inclusion of the writing section greatly enhances the validity of the test as a measure of communicative competence rather than simply of linguistic knowledge.

Item Analysis

Test A

In terms if item difficulty, it can be seen that the items represent a wide spread of difficulties and, therefore targeted the whole test population effectively (see table 3). However, only 8 of the 30 items generated item difficulty scores of over 60%, meaning that for the remaining 22 items fewer than 60% of the test-takers were able to answer correctly. The inclusion of additional items towards the easier end of the difficulty scale would provide better balance and would bring the mean scores closer to the desired score of 60%, as well as making the test more accessible for lower proficiency test-takers.

Table 3. Item facility (IF) and point biserial correlation coefficient (r pbi) for each item in Test A

Item #	IF	r pbi
1	0.49	0.25
2	0.60	0.22
3	0.48	0.36
4	0.68	0.36
5	0.56	0.35
6	0.13	0.29
7	0.43	0.38
8	0.41	0.28
9	0.81	0.30
10	0.57	0.34
11	0.70	0.34
12	0.88	0.39
13	0.34	0.25
14	0.14	0.06
15	0.42	0.34
16	0.58	0.36
17	0.40	0.38
18	0.52	0.14
19	0.47	0.21
20	0.55	0.40
21	0.83	0.35
22	0.62	0.39
23	0.49	0.30
24	0.57	0.25
25	0.76	0.51
26	0.55	0.42
27	0.62	0.25
28	0.29	0.18
29	0.44	0.34
30	0.44	0.11

Item 6 was the most difficult but since it discriminates reasonably well could be retained in order to target the high level test takers. Item 14 on the other hand is not only extremely difficult but also failed to distinguish between the stronger and weaker students and would therefore need to be revised or removed. Items 18, 28 and 30 also have a low discrimination index and require further evaluation. However, all the remaining items performed at or above the generally accepted level of .20 and can be regarded as satisfactory.

Test B

For the other test, there was a balance of easy and difficult items, although four of the items were answered correctly by fewer than the 25% we would predict from random guessing and only 7 of the 36 items were answered correctly by over three-quarters of the test-takers (see table 4). Therefore, again, a higher proportion of easier items would improve the test. In terms of discrimination, one item (number 32), produced a negative score meaning that the weaker test-takers were likely to perform as well or better than the higher proficiency test-takers. This is clearly problematic and, although it was not immediately clear to the test development team why this particular item should behave in this way, it was removed from the item bank. An additional four items were found to have point biserial correlation scores below the acceptable cut off of .20. In the cases of numbers 5 and 13 the items seem to be failing to discriminate well simply because they are very easy and a case

can be made for retaining a small number of such easy items in order to ensure that the test is accessible to even the lowest proficiency test takers. In the case of item 31, ability to answer correctly depended on the students' knowledge of the idiomatic verb 'to run out of time'. It is clear that even the higher-level test takers struggled with phrasal verbs and other forms of idioms, even those which are relatively common in spoken English. The team was divided on whether to retain such items. On the one hand, as examples of commonly-occurring natural language, it seems that they should be worthy of inclusion; on the other hand, test-takers seem to have had very limited exposure to such language and may not reasonably be expected to answer correctly.

Table 4. Item facility (IF) and point biserial correlation coefficient (r pbi) for each item in Test B

Item #	IF	r pbi	Item #	IF	r pbi
1	0.81	0.47	19	0.51	0.60
2	0.74	0.23	20	0.84	0.38
3	0.20	0.30	21	0.50	0.38
4	0.40	0.41	22	0.59	0.43
5	0.92	0.16	23	0.53	0.46
6	0.36	0.40	24	0.75	0.43
7	0.30	0.40	25	0.49	0.32
8	0.56	0.32	26	0.83	0.51
9	0.75	0.33	27	0.33	0.42
10	0.81	0.26	28	0.45	0.16
11	0.76	0.21	29	0.24	0.21
12	0.68	0.39	30	0.42	0.46
13	0.89	0.17	31	0.22	0.04
14	0.68	0.24	32	0.24	-0.13
15	0.57	0.28	33	0.57	0.39
16	0.47	0.38	34	0.62	0.27
17	0.64	0.55	35	0.29	0.38
18	0.77	0.41	36	0.33	0.26

One very positive finding from the research was that the items from the fourth section of the test (numbers 16-26) seemed to function very well. Although none of the 11 items from this section could be said to be very difficult (the lowest difficulty index reported was .47), all the items discriminated well, generating point biserial correlation scores from .32 to as high as .60. This section was an entirely new task-type devised by a member of the development team in which test takers are presented with four short written texts on a similar topic (in this case holidays abroad), each by a different narrator. The test taker is then presented with a series of statements in Japanese and asked to select which of the four narrators the statement best fits. There are a number of advantages to this task type in addition to the ability to discriminate effectively. Firstly, the texts are relatively easy to generate and although care needs to be taken creating and translating the distractors it is a highly efficient task type insofar as one text can generate 10 or more items. Secondly, it is relatively easy to adjust the item difficulty simply by modifying the length or complexity of the texts. Finally, having the distractors in Japanese ensures that the task focuses on the test-takers' ability to comprehend the text rather than rely on test-taking strategies such as matching the language of the item to the language of the text.

However, the other new task type (items 11-15) was considerably less successful. Although four of the five items were satisfactory in terms of their discrimination scores, the item facility covered a fairly narrow range bunched towards the easy end (from 0.57 to 0.80) suggesting that a significant proportion of test-takers were insufficiently challenged by this section. Perhaps even more importantly, the item writers felt that these items were extremely difficult to create. Each text required careful scripting and it proved extremely challenging to come up with convincing distractors, as a result of which these were the only items on either of the tests that had only three, instead of four, answer choices. Also, since the distinctions between the distractors were quite nuanced, translating them into Japanese proved to be highly

challenging. Given the considerable time and effort required to come up with just four functioning items, the team agreed that it would not be feasible to retain this item type in the actual test.

5. Interim Conclusions

It is anticipated that it will take up two more years before the team can be confident that the reformed test is fit for purpose. Therefore, any conclusions drawn at this stage are somewhat tentative. Nonetheless, it is clear that the development team have made some important discoveries that should help to inform the test development process. These include the following.

1. Since the tests scores make a significant contribution to admission decisions, the overall reliability needs to be improved. The inclusion of additional items to ensure that each test contains at least 40 items should help to achieve this.
2. Members of the test team are well skilled in creating items with suitable distractors. The majority of items were able to discriminate well between the more and less able test-takers and each answer option attracted at least some responses suggesting that the test developers are able to create convincing distractors. However, for better balance and to ensure that scores are distributed around the desired mean of 60%, future iterations of the test should include a slightly higher proportion of easier items.
3. Some new task types, particularly the items based on the four comparable texts described above, performed well and could be considered for inclusion in the final forms.
4. The original aim of redesigning the tests into a purely multiple-choice format may have been misguided. Although the multiple-choice format has many benefits and can lead to increased reliability, it also has serious limitations which can impact negatively on test validity. For example, the multiple-choice format may advantage test-takers who have developed effective test-taking strategies and disadvantage test-takers who have a high degree of communicative competence but less reliance on learnt forms. Perhaps even more pertinently, it is extremely challenging and time-consuming to produce good multiple-choice items so the time saved by scoring the items electronically is lost to item creation. Moreover, without being able to trial items in advance of the test, it is far from clear that this effort is an investment worth making. Instead, an optimal solution might be to convert as much of the test as possible to multiple-choice but to retain some open response element, such as the writing section in the case of Test A or cloze, sentence completion, or word reordering elements for Test B. Such tasks will be developed and evaluated further over the coming year.
5. Even when there is agreement about the desirability of change, there may be less agreement about the pace, degree and approach so the first stage of reform should be to build consensus between the various stakeholders both within and beyond the university community. This may include the university faculties, the university admissions and academic affairs sections, high school teachers, cram school teachers, parents and, of course, the test-takers themselves. Effective communication with stakeholders should ensure that students and their high school teachers are receptive to changes and may serve to facilitate positive washback.

6. Next Steps

Although the group has had some success in developing tasks, not all the task types were successful so we will need to continue to develop and trial alternatives. Moreover, even for the successful task types, the number of items produced remains small so we need to keep adding

to this bank of items. The current aim is to have a fully trialled prototype of Test B completed by the end of the 2015/6 academic year and a prototype of Test A ready by the end of 2016/7. These prototype tests can be shared with stakeholders and serve as sample papers for high school students preparing to sit the first round of the reformed exams in 2016/7 or 2017/8.

Once the format of the test has stabilised the group can devote more attention to building a validity argument supporting the test's use. As well as continuing to statistically analyse iterations of the test in order to measure reliability, item facility and discrimination, it will also be necessary to conduct analyses using Rasch modelling in order to evaluate item targeting and dimensionality. The former will show us how well the test matches the abilities of the test-takers, while the latter will help us to determine the extent to which variance exists within the test scores which cannot be accounted by the English ability construct. As far as possible the results of these statistical analyses should be made available to stakeholders in order to enhance the tests' credibility.

In the meantime, of course, the committee must continue to create tests conforming to the old specification to be administered while the new version is being developed. This time-consuming process severely limits the time available to work on the new tests but provides an opportunity to introduce some of the less controversial changes incrementally over the course of several years. These include a gradual replacement of the Japanese test rubric with instructions in English, the conversion of selected items to multiple-choice format and the inclusion of more items based on inference.

Acknowledgement

The authors would like to thank ex-colleague Hana Craig for the design and development of the new task-type described in the item analysis section for Test B, as well as for her assistance in conducting the test trials.

References

Aspinall, R. (2005). University entrance in Japan. In J. S. Eades, R. Goodman & Y. Hada (Eds.), *The 'big bang' in Japanese higher education*, (pp. 199-218). Melbourne: Trans Pacific Press.

Bachman, L. and Palmer, A. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Brown, J. D. (2000). University entrance examinations: Strategies for creating positive washback on English language teaching in Japan. *Shiken: JALT Testing and Evaluation SIG Newsletter,* 3(2), 2-7.

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.

Brown. J. D. (1995) English language entrance examinations in Japan: Problems and Solutions. *JALT 95 Conference Proceedings*, 273-283.

Cook, M. (2013a). The multipurpose entrance examination: Beliefs of expatriate ELT faculty. *The Language Teacher, 37*(1), 9-14.

Cook, M. (2013b). You say you want a revolution? Changing Japanese University Entrance Examinations. *The Kyoto JALT Review*, (1), 17-44.

Fulcher, G. (2010). *Practical Language Testing*. London: Hodder Education.

Green, A. (2014). *Exploring Language Testing and Assessment*. London and New York: Routledge.

Guest, M. (2008). Japanese university entrance examinations: What teachers should know. *The Language Teacher*, 32(2), 15-19.

Kikuchi, K. (2006). Revisiting English entrance examinations at Japanese universities after a decade. *JALT Journal, 28*(1), 77-96.

Leonard, T. J. (1998) Japanese University Entrance Examinations: An Interview with Dr. J. D. Brown. *The Language Teacher, 22*(3), 25-27.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: Macmillan.

Messick, S. (1996). Validity and washback in language

testing. *Language Testing, 13*(3), 241-256.

MEXT. (2011). Five Proposals and Specific Measures for Developing Proficiency in English for International Communication Retrieved on October 16, 2015 from http://www.mext.go.jp/component/english/_icsFiles/afieldfile/2012/07/09/1319707_1.pdf

MEXT. (2014). English Education Reform Plan corresponding to Globalization. Retrieved on October 16, 2015 from http://www.mext.go.jp/english/topics/_icsFiles/afieldfile/2014/01/23/1343591_1.pdf

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.

Stapleton, P. (1996, March). A reaction to J. D. Brown's recent inquiry on the English entrance exam. *The Language Teacher, 20* (3), 29-32.

Stout, M. (2003). Not guilty as charged: Do the university entrance exams in Japan affect what is taught? *The ELJ Journal 4*(1), 1-5.