Developing an Online Placement Test for the Japanese Program at UC Berkeley

Yasuko Konno Baker and Noriko Komatsu Wallace
Lecturers in Japanese, Department of East Asian Languages and Cultures

The purpose of our BLC project was to develop an online adaptive test for placing students in the appropriate course in the Japanese language program. The creation of an online test was recommended as part of an external review in Fall 2011; moreover, an online placement test would more efficiently place the more than 100 students taking the test each year, and an online test would provide a more objective means of placement, since currently tests are graded by different instructors with varying standards for placement.

Designing the Test

We began planning for this large project in the spring of 2012, and it is still in progress. Table 1 shows what we have done as of Spring 2014, and what we have scheduled for the future to complete the process. Please note that we are dealing with six levels, each level corresponding to one semester of the first three years of Japanese language study.

Table 1: Project Schedule

Level of test	Spring 2014 fellowship period					Spring 2015 Fellowship period				
	Start of Spring 2014	Spring 2014	End of Spring 2014	Summer 2014	End of Fall 2014	Start of Spring 2015	Spring 2015	End of Spring 2015	Summer 2015	Start of Fall 2015
Level 1 (J1A)	Gave NT J1B	Analyzed NT		Analyze NT	Give NT		Analyze NT			Give PT
Level 2 (J1B)		Wrote NT	Gave NT J1B	Analyze NT				Give NT	Analyze NT	Give PT
Level 3 (J10A)	Gave NT J10B	Analyzed NT		Analyze NT	Give NT		Analyze NT			Give PT
Level 4 (J10B)		Wrote NT	Gave NT J10B	Analyze NT				Give NT	Analyze NT	Give PT
Level 5 (J100A)	Not started due to textbook change Fall 2013				Give NT (level 4 & 5)		Analyze NT			Give PT
Level 6 (J100B)					Write NT		Write level 6 Q's	Give NT J100B J100X J102 J104	Analyze NT	Give PT

NT = norming test PT = placement test

As can be seen from Table 1, the process involves 4 steps:

1) Create questions and categorize them (see below).
2) Administer a norming test to students currently enrolled in the program. The norming test verifies that students currently in the program can handle the

levels we assign to questions based on the current curriculum. Thus, material we've assigned to level 3 should be passed at a high percent by students who have completed level 3 and above, whereas students at lower levels should do poorly on the same question.

3) Analyze the results of the norming test, modifying or deleting questions that seem too hard (students who have completed a the level still get a question at that level wrong at a high percentage) or too easy at the assigned level (students at levels lower than the assigned level get a question correct at a high percentage).
4) Administer the placement test.

It should be noted that the online test is one of three components of the process of placing students in the appropriate level. After administering the placement test and looking at the results, we will conduct a short interview with students and, if placement at higher levels seems warranted, students will be asked to write a short essay. The reason we use these three different components is because applicants to our program have widely different abilities in terms of speaking proficiency and writing proficiency. To place them into the course level that will be best for them, we need to measure each of these skills on a case-by-case basis. The oral interview and short essay help us do this.

Questions and Categories. The questions are in a standard multiple-choice format. In accordance with best practices for reliable multiple-choice questions, there are four choices for each question.

We wanted to create a variety of questions, but realized that if the software was pulling questions from a database randomly, students might see a disproportionate number of questions of one particular type. Therefore we created categories and subcategories for questions, and the program allows us to specify that students receive a minimum number of questions from each category and subcategory.

We chose three categories of questions: grammar, *kanji*, and reading comprehension. Under the *kanji* category, some examples of subcategories are *kanji* meaning, *kanji* reading (the pronunciation in a particular context), and discrimination of similar *kanji*. The grammar category is further divided into subcategories for particles, pragmatics, functional expressions, etc. The reading comprehension category has no subcategories. Each reading passage consists of two multiple choice comprehension questions.

Question Construction. Wherever possible, we tried to create questions that are natural and meaningful and also reflect Japanese culture and attitudes. For example, originally the kanji part contained the question、

「日本では（　　）になるとさくらがさきます。」
"When spring comes in Japan, cherry trees bloom. "

a. 夏　　b. 秋　　c. 冬　　d. 春

The correct kanji for spring is found in (d.).

As a simple kanji question, this sentence is fine, but it does not reflect Japanese culture, since flowers, including cherries, bloom everywhere in the world in the spring. If we want to reflect a unique characteristic of Japanese culture, we would say, 「日本では（春）になってさくらがさくと、お花見をします。」 "When spring comes in Japan and cherry trees bloom, people go to see the blossoms to celebrate." Celebrating the blooming of the cherries reflects Japanese culture because it is a special tradition in Japan. However, while this is good information to introduce to the student, it is not necessary to know that about Japanese culture in order to correctly answer the question. The assumption of cultural knowledge is a valid component in the placement process.

Also, we try to make questions that require an understanding of pragmatics, because when we teach language, we cannot ignore the context in which the language is used. For example,

Choose the most appropriate expression for the situation. Yamada-san and Smith-san are classmates.

山田：スミスさん、

a. 先生が聞いた質問が分かりましたか。
b. 先生が聞いた質問がお分かりになりましたか。
c. 先生がお聞きになった質問がお分かりになりましたか。
d. 先生がお聞きになった質問が分かりましたか。

スミス：いいえ、ちょっと…。

Yamada: Smith-san, did you understand (分かりました・お分かりになりました) the question the teacher asked (聞いた・お聞きになった)?
Smith: No, I didn't.

This question tests whether the applicant can understand proper use of honorific expressions. Japanese has a complex system of honorific expressions that indicate honor and/or respect toward the hearer or the person spoken about.

In the above example, the boxed phrases are honorific descriptions of the teacher's action (asking) and underlined phrases are honorific expressions of Smith's state (understanding). The verb "ask" describes the teacher's action, so Yamada-san *should* use an honorific expression. On that one point alone, answer (c.) or (d.) would both be possible answers—both have the honorific form of the verb, as the boxes indicate. However, "understand" refers to his classmate's state, and Yamada-san *should not* use an honorific

expression for his classmate Smith-san. That then eliminates answer (c.) that has the honorific form of Smith's state, as the underlining indicates. Thus, the correct answer is (d.).

This is an example of the type of pragmatics questions that we are making. We want to test whether applicants understand the appropriate use of honorific expression in the context of Japanese culture.

Sometimes we include in questions cultural concepts or facts that we want to teach via the question itself but which we do not expect applicants to already know. The correct answer should *not* require prior knowledge of that concept or fact.

For example, the following grammar questions ask the student to select an appropriate particle. The student can answer this type of question even if she or he has never heard the specific details about Japanese culture. We want to integrate cultural concepts and facts into questions but we do not want to test the extent of the applicants' "cultural" fluency. For example,

> *Insert an appropriate particle for each blank to make a correct sentence.* [The answer has been inserted below, for clarity.]
>
> 日本には商品__を__温めてくれる自動販売機がある。
>
> In Japan there are some vending machines that warm up food automatically and serve it.
>
> 日本のまんがは左から右__に__ページを開く。
>
> The pages of Japanese manga magazines open from left to right.

These questions contain Japanese cultural information that we want to include, but they do not directly *test* those points. The student is entirely able to answer correctly without prior knowledge of the cultural point.

Data collection and analysis

In these norming tests, we gave two different levels of questions to two different levels of students, Level 1 questions to the students who had just finished the first semester of Japanese, and Level 3 questions to the students who had just finished the third semester of Japanese. The results we obtained indicated for each question what percent of students who saw that question got it right (the Item Facility value[1]) (Brown 2005, p.67). Within a particular level, we were also able to determine how well better students did on a question

[1] Brown (2005) expresses these values as proportions between 0.00 and 1.00, rather than percentages.

compared to weaker students at the same level (the Item Discrimination score). However, we lacked the data necessary to discriminate between students at different levels that is essential for making placement tests. So, for example, for a particular question that we had assigned to Level 3, we know how students who had completed that level did on the question, but we lack the data for how second-semester students performed on that question (since we didn't administer Level 2 questions in the norming test given in January 2014). Therefore, we cannot determine whether this question discriminates between students at different levels. Ideally, for a Level 3 question, third-semester students get the question correct, and second-semester students get it wrong. We will obtain this data at the next norming test in December 2014.

Thus, a question that has been designated as a Level 3 question, for example, should be correctly answered by students who have completed third semester Japanese. Questions with low percentages need to be examined carefully to determine why the scores were low: Were the directions clear? Was a particular distractor too attractive? If all questions in a particular category or subcategory had low Item Facility (IF) scores, then that might indicate problem areas in the curriculum. Questions with low IF scores need to be modified or discarded from the database.

The first norming tests covering questions at Levels 1 and 3 were given to the students between January 29 and 31, 2014, just after the beginning of the spring semester at UCB. The students were told to log into the BLC placement test site by choosing the class they were enrolled in at that time and entering their student ID number.

Table 2: Testing completed in spring 2014

Course #	Level	Date given
J1A	Level 1	Feb. 2014
J1B	Level 2	April 2014
J10A	Level 3	Feb. 2014
J10B	Level 4	April 2014

Participation. We thought about several options to encourage student participation: a) make the placement test mandatory without any extra credit, b) make it voluntary with extra credit, or c) make it voluntary without any extra credit. We worried about how many of the students would participate in the norming test if we made it totally voluntary, but did not want to give them an extra burden in addition to their regular classroom assignments, which are already quite heavy, so we decided to make it voluntary with extra credit. We made an announcement in class asking for participation and also posted the announcement on bSpace (the UC Berkeley course management system). We told the students that participation was optional, but that the participants could receive extra credit of 1% of their course grade if they completed all 50 questions. Students were able to take the test from anywhere and at any time using their own computer or the ones in the computer lab at school.

The results of this first NT (norming test) were sent to us on a spreadsheet which contained the questions, the four choices of answers, the number of participants who received each question, the number of students who chose each answer, and the percentage of correct answers for each question.

Table 3: Student Participation in Norming Tests, January 2014

Level	Took the Norming test	Incomplete	Wong level	Completed the right level	Completion ratio
Level 1	94 (75%)	5	10	79	63%
Level 3	50 (83%)	6	10	34	57%

Table 3 shows that 94 out of 125 Japanese second-semester students took the Level 1 test, which was 75% of the class, and 50 out of 60 fourth-semester students took the Level 3 test, which was 83% of the class. However, we discovered that five second-semester students and six fourth-semester students began the Level 1 / Level 3 tests, respectively, but did not complete them. Furthermore, we realized that ten students at each level had taken the test for the wrong level. We had to discard these students' data, in addition to the data from students who didn't take the entire test, which was a significant loss of data. Therefore, we had a 63% completion ratio for Level 1, and 57% for Level 3. After considering these low rates of participation, we think that in the future we should make the Norming Tests mandatory, so that we can acquire more data.

Question database. This norming test drew on a database of 117 questions for Level 1, and 141 questions for Level 3, which includes *kanji*, grammar, pragmatics, and communicative functions (such as asking a favor and offering to help someone). Each participant received 50 questions randomly selected by the software. We anticipated that it would take about 30 minutes on average to take the test. We also predicted that each Level 1 question would receive about 34 answers and each Level 3 question about 12 answers, which we believed would be sufficient data for our analysis.

As seen in Table 4, the average percentage of correct answers for Level 1 was 82% and for Level 3 was 69%; we were surprised to find that the average score for the Level 3 test was much lower than the one for Level 1. We expected that the students would score between 75 and 80% average correct. We wondered if the way we made questions was appropriate for the purpose of this test, and how we can find the answer to this question.

Table 4: Counts of questions and answers on norming tests

	Level 1 test	Level 3 test
Total number of questions	117	141
Number of students	79	34
Number of questions given to each student	50	50
Estimated avg. number of answers/question	50*79/117=34	50*34/141=12
Avg. percent correct	82%	69%

In order to find out why the average percent of correct answers for Level 3 was very low, first we had to examine questions with low average percentages of correct answers and determine whether there was a particular distractor causing the low average percentages, or what other factors might have caused a low percentage. If a particular distractor was frequently chosen, we will consider ways to make the distractor less attractive in order to raise the average score and make them more appropriate to include in the final placement test. We should be able to raise the scores of some of the questions with low percentages correct, and make them more appropriate to include in the final placement test.

Item analyses. In order to evaluate the effectiveness of the individual items on the placement test systematically (a process known as Item Analysis), Brown (2005: Chap. 3-4) suggests conducting three different analyses: A. Item Format Analysis, B. Item Facility Analysis, and C. Item Discrimination Analysis. However, since we haven't carried out the norming test across the different levels, at this point it may only be appropriate for us to look at the questions with a low rate of correct answers (IF analysis) and carry out Item Format Analysis.

The purpose of Item Format Analysis is to see "the degree to which each item is properly written so that it measures all and only the desired content." (Brown 2005, p.42) This analysis usually depends on tester's judgment rather than scientific analysis. For example, there is a checklist in Brown's General guidelines (Brown 2005, p.43) for most item formats which includes questions such as "Is the item format correctly matched to the purpose and content of the item?", "Is there only on correct answer?", etc.

Issues

Inadequate instructions. After examining the questions with a low rate of correct answers, based on these guidelines, we found some problems in our questions. For example, questions that test a student's ability to offer to do a favor for someone in Japanese at Level 3 had a low rate of correct answers. We believe that learning Japanese pragmatics is an important part of learning the language, which third-semester students should master. We presented these questions involving pragmatics in conversation style, because it is crucial that the grammatical patterns of the sentences be learned in a specific context. However, particularly for "how to offer a favor" questions, it was very difficult for

us to make three plausible incorrect answers, so in this subcategory, we made three correct answers and one incorrect one for some of the questions. The instructions for these questions say, "Choose the INCORRECT answer…". This type of instruction might have confused the students if they hadn't read them carefully, so we asked the programmer to present the instructions and questions physically closer to each other on the screen and to highlight the instructions so that the students would not miss them. We will try to test these questions again next time.

Question length and vocabulary. Some questions in the *kanji* or grammar categories on the Level 3 test contained either difficult or extensive vocabulary, which we believe contributed to low scores on those particular questions. As a result, we re-examined the Level 2 and 4 questions in preparation for conducting the next norming test, and made the sentences shorter and simpler, so that only relevant information is presented. In this way, the students should be able to focus on the intended target of each question.

Here is an example of a grammar question that we simplified by eliminating difficult vocabulary. Students have to choose the correct form of the verb.

The old version said:

学生１：EU の失業率は大変だね。特に若者達に仕事がないそうだね。
学生２：うん。私もそのニュースを見て
a.考えられた
b.考えなかった
c.考えさせられた
d.考えさせた

Student 1: The unemployment rate in the EU is really bad, isn't it? It seems like they just don't have jobs, especially for young people.
Student 2: Yeah. I saw that on the news and ( ).
a. I was thought or I was able to think (passive and potential form).
b. I didn't think (negative past).
c. It made me think (causative passive form, lit. 'I was made to think').
d. I made (somebody) think (causative form).

The correct answer is (c.). As you see, the topic is pretty advanced, and the vocabulary in the underlined part is difficult for the third-semester students (even though they have studied these words in class), so we changed the question to make it simpler. The new version says:

学生１：いじめの問題は大変だね。
学生２：うん。私もそのニュースを見て（　　）。
a. 考えられた
b. 考えなかった

c. 考えさせられた
d. 考えさせた

The revised version has the same options for the answer, but in this version, Student 1 simply says: “Bullying is a big problem, isn’t it?” Since students are more familiar with the topic of bulling than with the unemployment rate in the EU, and there isn’t any difficult vocabulary, they can focus on the content of the conversation and choose the appropriate form of the verb 考える ‘to think’.

Here is another example of a question that we simplified dramatically. Students read the paragraph and choose an appropriate connective in the parentheses. This is the old version of the paragraph:

> （スミスさんの日記）今日は、日本語のクラスメートといっしょにピクニックをした。11 時にさくら公園で会って、一時間ぐらいバレーボールをした後、みんなで昼ご飯を食べた。天気もよかったし、みんな食べ物をたくさん作って来たので、おいしい物を食べながらゆっくり色々な　話をした。楽しかった。（a.けれども　b.それに　c.その上　d.それが）来ると言っていた木村さんは来なかった。どうしたのだろう。
>
> (Smith’s Diary) Today, I went on a picnic with my classmates from Japanese class. We met in Sakura Park at 11 o’clock, played volleyball for an hour, and then had lunch. The weather was nice, and everybody had made and brought a lot of food, so we had a nice conversation about all kinds of things while eating delicious food. It was a lot of fun. (But) Kimura-san didn’t come even though she had said she would come. I wonder what happened to her?

As you see in the original version, we included a lot of information about Smith-san’s activities in the park in order to explain his regret that Kimura-san didn’t show up although she had said she would come. However, this question involves a lot of reading, and actually the students would have enough information to choose the correct connective without the underlined part. So, we decided to omit that part to make it shorter and simpler, so that the students can understand the situation without having to read too much. The new version says:

> （スミスさんの日記）今日は、日本語のクラスメートといっしょにピクニックをした。（a.けれども　b.それに　c.その上　d.それが）来ると言っていた木村さんは来なかった。どうしたのだろう。
>
> (Smith’s Diary) Today, I went on a picnic with my classmates from Japanese class. But Kimura-san didn’t come even though she had said she would come. I wonder what happened to her?

We are going to use the questions which we have revised by replacing difficult terms with more familiar vocabulary and eliminating irrelevant information in the next Level 1 and Level 3 norming tests, and compare the results to see whether these changes actually improve the effectiveness of questions.

Assigned question difficulty. In some cases we need to reassess the difficulty level of the assigned questions. For example, for the level 1 exam there are six questions (most students would see some but not all) dealing with the *tokini* construction, which is a subordinating conjunction that tells when something happened and can be translated as *when, after, or while*.

In the norming test, the questions on clauses with *tokini* had a correct answer ratio of only 40-50%, and the reading comprehension question with a *tokini* clause was only 53% correct, which we consider quite low.

This is an example of a question containing a *tokini* clause:

> The students choose the correct sentence for a reply to the question.
>
> 学生Ａ：かぶきを見たことがありますか。
> 学生Ｂ：a. いいえ、でも、日本へ行った時に見るつもりです。
> b. いいえ、でも、日本へ行く時に見るつもりです。
> c. いいえ、日本へ行く時に見ました。
> d. いいえ、日本へ行った時に見ました。
>
> Student A: Have you ever seen a *kabuki* performance?
> Student B:
> a. No, but I am going to see one when I have arrived in Japan (after going to Japan).
> b. No, but I am going to see one when I go to Japan (in the process of going to Japan).
> c. No, I saw one when I went to Japan (in the process of going to Japan).
> d. No, I saw one when I went to Japan (after going to Japan)

Only 36% of the Level 1 students answered this question correctly. In Japanese, the tense of the sentence appears in the verb of the main clause, which comes at the end of the sentence (underlined). The verb before *tokini,* which means "when" in this case, doesn't carry the tense of the action; instead it is marked to indicate whether the action in the clause is completed or in process, which is very different from English, although the forms in Japanese are past (perfective) or non-past (in process). The correct answer is (a.) because "I am going to see *kabuki* performance when I have arrived in Japan." However, 64% of the students chose (b.) for their answer. We can see that the students chose the answer where both verbs have the non-past form and we can assume that this mistake is caused by interference from English, where the present tense is typically found.

However, we couldn't find a clear reason for the mistakes in the other five questions with *tokini*. The structure with *tokini* is introduced near the end of the first-semester course, so it is possible that the students hadn't had enough practice using this structure. But are the low IF scores really because first-semester students haven't mastered how to make a subordinate clause using *tokini*? Are the questions with *tokini* not appropriate for the first-semester students' proficiency? In order to find out more about this, we decided to create additional questions with subordinate clauses for the next Level 1 norming test. We also will examine how second-semester students perform on *tokini* questions.

Uneven distribution of answers. First, we found that some questions with very high or very low IF scores received only one or two answers, indicating that the distribution of questions across students was very uneven. The computer program randomly assigns questions to each student, but we found it necessary to put some restrictions on the randomness so that each question is answered more or less equally. The algorithm for generating questions was modified to choose first a category or subcategory requiring a student response, and then choose the question from that subcategory with the fewest responses. In this way, all the questions in each category or subcategory will receive approximately the same number of student responses.

Uneven distribution of types of questions. We also found that the distribution of the types of questions across students was uneven. Because the program was set to choose questions randomly, reading comprehension questions were assigned unevenly. As a result eight Level 1 students had to answer all five reading questions, 26 students answered 4 questions, 32 students answered 3 questions, 12 students answered 2 questions, and one student was given only one reading comprehension question. For level 3, one student had to answer all five questions, eight students answered 4 questions, 13 students answered 3 questions, 9 students answered 2 questions, and 5 students were given only one reading comprehension question. Because each reading comprehension question consists of 3-5 short paragraphs with two multiple choice comprehension questions, this type of question takes longer to answer, and is more difficult than many of the *kanji* and grammar questions. We are very sure that this uneven distribution of reading comprehension questions made this NT very unfair. In order to solve this problem, the program was modified to allow instructors to indicate both a minimum and a maximum number of questions in each (sub)category. For the next norming test, each student will be given 1 reading question to answer, and only 1.

Future Directions

We still need to complete an analysis of the data for the purpose of creating an appropriate placement test for our program. As mentioned above, the purpose of the placement test is to see which level is most appropriate for a student who has studied Japanese outside of our program, so that they can study together with the current students without problems.

Our initial analysis, reflected in this paper, was based on the results of the first norming tests for Level 1 and Level 3, which were given to first-semester and the third-semester students, respectively. In May 2014, we administered the first norming tests for Levels 2 and 4 to second-semester students and fourth-semester students, deliberately assigning them some questions at lower and higher levels, in addition to the bulk of questions at the level they had just completed. In addition to the Item Facility Analysis at all levels of questions, this additional data will allow us to carry out an Item Discrimination Analysis on questions, i.e., compare how students at different levels perform on a particular question. This will enable us to establish whether a particular question effectively discriminates between student levels, an essential for a placement test. Our next step will be to create questions for Levels 5 and 6, and then go through the norming process two more times in order to winnow out ineffective questions. In this way, we should be able to create a good placement test for our program and have it ready to use by the beginning of the fall semester in 2015.

Conclusions

We have spent almost two years on planning and creating questions for Levels 1, 2, 3, and 4, and finally we were able to conduct the first norming test for these levels this semester. However, we encountered many unexpected problems, such as the small number of participants, and unusable data due to students who didn't complete all the questions or who took the wrong level of test. We also made some mistakes in the process of creating questions, such as having an unbalanced number of questions in each category, and creating too many easy questions for both levels, and too many difficult questions for Level 3. Many of these mistakes could have been avoided with more careful planning and consideration, but having gone through the process for the first four levels, we will use our experience as we construct questions for Levels 5 and 6.

References and Suggestions for Further Reading

Reference

Brown, J. D. (2005). *Testing in Language Programs: A Comprehensive Guide to English Language Assessment. .* New York: McGraw-Hill.

Suggestions for Further Reading

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

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

Hughes, A. (1989). *Testing for Language Teachers.* Cambridge: Cambridge University Press.

Ito, S. (2008). *Nihongokyoshi no tame no test sakusei manual*（日本語教師の為のテスト作成マニュアル）. Tokyo: ALC.

Kondo-Brown, K. (2012). *Introduction to Assessment for Japanese Language Teachers*（日本語教師のための評価入門）. Tokyo: Kurosio Publishers.

MacNamara, T. (2000). *Language Testing*. Oxford: Oxford University Press.

Onozuka W. & M. Shimada. (2008). *Nihongokyoshi no tame no Excel de dekiru test bunseki nyumon*（日本語教師のための Excel でできるテスト分析入門）. Tokyo: 3A Corporation.

The Society for Teaching Japanese as a Foreign Language. (1991). *Nihongo test Handbook*（日本語テストハンドブック）. Tokyo: Taishukan Shoten.