Analysis of a new Japanese language placement test battery using G-theory and Rasch model programs

Etsuko Toyoda
The University of Melbourne
and
Yoji Hashimoto
The University of Tasmania

Abstract

This paper reports the analyses of a newly introduced placement test battery in the Japanese Program at the University of Melbourne. Focus is placed on the reliability of two innovative tests employed, SPOT (Simple Performance Oriented Test) and SKAT (Simple *Kanji* Awareness Test). SPOT and SKAT aim at obtaining maximum information on test-takers' receptive proficiency in spoken and written forms of the language with minimum time and human resources. The test results for 82 students were analysed using two test analysis programs, Generalizability theory based program, GENOVA, and Rasch measurement model based program, QUEST. In the course of the analyses, the two test analysis packages successfully complemented each other, and showed that they were powerful tools for supplying relevant data for test improvement. The analyses showed that both tests were highly reliable, and that they could be further improved through revisions.

1. Introduction

One dilemma faced by anyone involved in the development and/or management of language testing is the conflicting demands to attain test quality and time efficiency. It is particularly the case with designing language placement test devices: placing a large number of

Melbourne Papers in Language Testing 2001 Volume 10.1 pp. 49-66. The Language Testing Research Centre, The University of Melbourne.

students efficiently in a very tight time-frame with minimum resources is an issue that requirescritical attention.

In the Japanese Program at the University of Melbourne, two innovative time-saving tests, SPOT and SKAT, have been employed, replacing the conventional time-consuming placement test procedure. In order to use these tests as self-standing placement devices, their reliability and validity need to be established. This paper reports the findings from a study investigating the reliability of both tests using two test analysis application programs. The paper further discusses prospects of possible test improvement.

2. Background

The dilemma mentioned above has been experienced by Japanese programs at many Australian universities in recent years. Since the late 1980s, efficient placement procedure for Japanese language courses has become increasingly demanding as the number of studentshasincreased and as the proficiency gaps among them have widened.

Under these circumstances, a more efficient and accurate placement procedure has been sought. In the Japanese Program at the University of Melbourne, new attempts have been put in place since 1998. The innovation involved the gradual introduction of two newly developed tests: SPOT (Simple Performance Oriented Test) and SKAT (Simple *Kanji* Awareness Test). The ultimate aims of the new placement test procedure are: to minimise the resources needed to conduct placement tests, and at the same time, to produce highly accurate placement outcomes.

The current placement procedure is composed of five components as shown in Table 1 below.

Table 1. The new placement procedure

Component	Time
1) SPOT (Simple Performance Oriented Test)	10 minutes
2) SKAT (Simple *Kanji* Awareness Test)	10 minutes
3) Written composition	30 minutes
4) Oral interview	up to 5 minutes per student
5) Personal language background questionnaire	3 minutes

The advantages of this arrangement are two-fold: 1) there is no need to develop different tests for different levels; 2) thereby the test can be administered simultaneously for all students at varying stages of language learning.

As will be explained later, the skills that SPOT and SKAT aim to measure overlap roughly with those covered by the components 3) written composition and 4) oral interview, which are more conventional forms of placement measures. At this stage of the development, due to the fact that both SPOT and SKAT are yet to establish themselves as self-standing devices, they need to be used concurrently with other measures. Further evaluation is essential for them to be used independently of other testing tools. Nevertheless, compared to the pre-1998 placement procedures, in which five different level-by-level tests had to be devised and administered separately, the introduction of SPOT and SKAT has created a drastic reduction in the time and resources required for decision making (Hashimoto, 1999; Hashimoto, 2000)[1].

[1] Hatasa and Tohsaku, 1997 reports on a successful placement procedure trialled in a university level setting, in which SPOT was the sole measurement tool.

3. SPOT and SKAT

3.1 SPOT

SPOT, or Simple Performance Oriented Test, was initially developed in the early 1990s at the University of Tsukuba in Japan (Ford-Niwa, 1997). This dictation type test is known to have high empirical validity through practices at various universities both in Japan and overseas (Kobayashi et al., 1996; Hatasa and Tohsaku, 1997; Murakami, 1998; Hashimoto, 2000). It is capable of indicating test-takers' language proficiency levels in an objective format in a short period of time (around 10 minutes). Trial reports from those institutions supply evidence that SPOT scores have a high correlation with students' grammatical knowledge and oral/aural proficiency at all levels from beginners' to advanced (Ford-Niwa et al. 1995; Ford-Niwa, 1998).

As shown in Figure 1, an item in SPOT is a short sentence written in Japanese, but with one *hiragana*[2] letter that has a grammatically meaningful function missing. The current version of SPOT[3] has 60 items similar to this. During a SPOT session, each of those unrelated sentences are read one after another. As test-takers look at each of them, they will simultaneously hear the sentences played by audio-tape at normal speed and must fill in one and only one correct missing *hiragana*.

[2] *Hiragana* is one of the two types of syllabic script *kana* used in Japanese writing system. Words that carry grammatical functions are normally written in *hiragana*.

[3] There were several versions of SPOT available at the time. The easiest version (ver. 3B) was used here.

Figure 1. SPOT item sample

なに
そこ (　　) 何をしているんですか。

SPOT is a type of indirect integrated test. It aims to capture test-takers' language proficiency by measuring their grammatical knowledge as well as the ability to instantly process language heard, both of which are necessary in real-world communication (Ford-Niwa et al., 1995; Ford-Niwa, 1998).

3.2 SKAT

While SPOT measures students' aural processing skills, SKAT aims to measure students' recognition ability of written forms of Japanese. The Japanese writing system employs two types of script, *kanji* (morphographic characters) and *kana* (syllabic letters). *Kanji* has an important role in Japanese written materials as most content words are written in *kanji*. Abundant evidence suggests that adequate recognition of *kanji* characters is crucial for Japanese literacy (Dobson, 1997; Okita, 1995; Hatano, 1986). Research suggests that, as learners advance in learning *kanji*, they gradually become more aware of information embedded within *kanji*, which leads to more efficient *kanji* recognition (Koda, 1999). It was therefore necessary to devise a test to measure students' *kanji* awareness as part of the placement test. For this purpose, the "Simple *Kanji* Awareness Test (SKAT)" has been developed.

SKAT is a direct discrete-point test. The test has 6 tasks, each of which measures a different aspect of ability required for *kanji* recognition. There are 5 multiple-choice questions in each task. A supplementary answer "n" was provided for "no idea". The structure of the test is illustrated as follows.

Figure 2. SKAT examples

1. Visual identification ability

The first task is to find *kanji* that is identical to the given *kanji* from the list of graphically similar characters.

eg. 木 a) 水 b) 木 c) 米 d) 大 n) no idea

The answer to this example is b).

2. Segmentation ability

The students are asked to find a *kanji* that does NOT have the given component. In other words, their task is to spot the given component in a character.

eg. 目 a) 歩 b) 相 c) 県 d) 眠 n) no idea

The answer is a), the first character. Other characters all share the component, 目, although some look thinner or smaller.

3. Decomposition ability

The third one is to measure students' ability to break down the *kanji* into parts. There are four sets (one correct set and three false sets) of constituent elements or components of the given *kanji*, and the task is to find the right set of components.

eg. 明 a) 日+月 b) 目+月 c) 月+月 d) 日+丹 n) no idea

The answer is a). The target character is made up of 日 and 月.

4. Ability for pattern recognition

Kanji can be categorised into several patterns. In this section, the students are asked to tell the correct pattern of the given *kanji.*

eg. 休 a) ⊟ b) ◫ c) ⧈ d) ▣ n) no idea

The answer is b), as the character has two parts, right and left, that can be separated from each other.

5. and 6. Ability to infer meaning and pronunciation symbols

Tasks five and six are to measure students' ability to utilise the information of meaning and pronunciation symbols. The tasks are to infer the meaning or pronunciation of a blurred *kanji* that had been shaded off leaving the symbols aside.

eg. a) speak b) listen c) drink d) come n) no idea

eg. a) kou b) kan c) shou d) sei n) no idea

N.B: Easy characters in terms of complexity were deliberately chosen for the examples in order to make the instructions clearer.

The inferring ability for solving tasks five and six is slightly different from the other abilities, as it requires awareness that part of *kanji* can be utilised for inferring the lexical information of *kanji*. It is important to elaborate on the above-mentioned symbols. *Kanji* consists of one or more components. Some components can show the meaning category of *kanji* and are called meaning symbols. For example, the meaning of *kanji* 晴 is "sunny", and the meaning symbol of the *kanji* 日 shows that this *kanji* is in the meaning category of "the sun". The information that meaning symbols convey is often very useful for inferring the meaning of an unfamiliar *kanji*. There are also components called pronunciation symbols that show one of the pronunciations of *kanji*. Knowing the function of pronunciation symbols is also useful as they often tell you how to read an unfamiliar *kanji* or *kanji* word. For example, the same *kanji* 晴 is pronounced /sei/, and it has a pronunciation symbol 青 or /sei/ in it.

In short, task five was to infer the meaning of the given *kanji* and task six was to infer the pronunciation of the given *kanji* using relevant symbols as clues. The answers for tasks five and six are b) and a) respectively. The meaning of the target character, 聞, should be inferred from 耳, the meaning symbol which conveys the meaning "ears". The pronunciation of 校 would be /kou/ as this character has the pronunciation symbol 交 that is pronounced as /kou/.

4. Instruments for analyses

The data were analysed by two test analysis applications, GENOVA (Crick and Brennan, 1984) and QUEST (Adams and Khoo, 1993).

GENOVA is a test analysis program developed on the basis of Generalizability theory or G-theory (see Lynch and McNamara, 1998; Shavelson and Webb, 1991). G-theory is a statistical theory developed by Cronbach et al. in the 1960s (Cronbach et al., 1972) and is becoming increasingly more widely used in language testing. It has overcome some of the limitations of classical test theory, which could not identify each of separate sources of errors that affect test scores: G-theory estimates relative effects on test scores that arise from different factors, eg. the number of test items, that of raters etc., or "facets" as they are called in G-theory. In its practical application, G-theory has a great advantage over its classical predecessor. A single administration of a set of test tasks to a fixed number of test-takers under one certain occasion/setting will, by applying G-theory, provide the investigator with the data to estimate the optimal combinations of facets by supplying test reliability to each of the possible combinations. In short, a single administration of a test could produce sufficient information for test analysis.

In analysing a test, GENOVA executes two phases, generalizability study (G-study) and decision study (D-study). G-study corresponds to the estimation of the relative effect of each facet on a test score mentioned above. For each facet, it provides a percentage figure that indicates the magnitude of its effect. This estimation is based on data obtained from a single observation and is generalised from the test-taker's observed test performance. Using the data obtained from G-study, D-study allows investigators to design a new combination of facets (eg. adding some more tasks with fewer raters), and see how the test results might be affected by the changes. D-study does this by providing reliability to each of the possible test conditions with several different combinations of facets. Investigators could then compare them and determine which version is most useful in a given test setting. In G-theory, this estimate of reliability is expressed by a statistical figure called Generalizability coefficient or G-coefficient.

QUEST is a computer application for the Rasch model (see McNamara, 1996), one of the two main models of Item Response Theory, or IRT. IRT is another statistical theory now widely used in language testing.

IRT analyses the interaction between person ability and item difficulty, and shows their relationship on a single map. The strength of QUEST is that it can illustrate the relationship between person ability and item difficulty on a single measurement scale. This scale is expressed in a kind of probability index known as logit scores. The logit scale is a true interval scale, which not only shows whether an item is harder or easier than another, but also by how much. This then allows investigators to see which of the items were more difficult and which were less so. Under the IRT assumptions, candidates that have the same underlying ability should obtain the same logit score, independently of influences of particular test item difficulties of different tests. It, in turn, is also capable of demonstrating how much better or poorer one candidate has performed in a test compared to other candidates.

The program also allows test administrators to identify any 'problematic' items that drew unsystematic responses, or examinees whose performance was not in an expected range. These items or examinees are referred to as "misfit". Misfit analysis also gives test administrators valuable information for revising a test by pin-pointing the items in a given test that did not discriminate the candidates' ability well.

The following section presents the results of the analyses by GENOVA and QUEST on SPOT and SKAT.

5. Data analyses

5.1 Data

The data were obtained from a group of 82 undergraduate students, with a diversity of language backgrounds and Japanese language proficiency levels, who sat for the placement test at the start of the first semester in 2000. The students' scores on SPOT and SKAT were statistically analysed using GENOVA and QUEST in this respective order.

5.2 Procedures

Both tests were administered to the students *en masse*. Prior to the test, the students were instructed to follow the instructions given by the teacher. It was emphasised that they must start and finish each task only when they were told to do so.

Firstly, the SPOT papers were handed out to the students with the example page up, and they were instructed to fill in the gaps with one *hiragana* letter as quickly as possible as the audio-tape gives the answers to each item at the speed of normal speech. They were given 10 trial sentences followed by the test, which consists of 60 sentences. As soon as the tape finished reading the last sentence, the students were told to put their pens down, and the test papers were collected.

As for SKAT, the procedure was similar to SPOT, except explanation was given before moving onto the next task. 30 seconds were allocated to each task, and the timing started only when it was confirmed that all the test-takers understood the example of the task.

5.3 SPOT analysis

5.3.1 SPOT analysis 1: GENOVA

Table 2 shows the generalisability coefficients for SPOT with varying numbers of items.

Table 2. Generalizability coefficient for SPOT

Number of items	Generalisability coefficient
10	0.784
20	0.879
30	0.916
40	0.936
50	0.948
60	**0.956**

The generalisability coefficient for the current 60-item version is .96. G-coefficient value falls between one and zero and indicates that the closer to one it is, the more reliable a test is at discriminating between students. Thus, a G-coefficient of .96 is very high and indicates that SPOT is very reliable as a testing tool. The coefficients for the versions that have 10, 20, 30, 40 and 50 items respectively are also demonstrated in the table. According to this table, if the number of items is reduced down to 30, for example, the test would still achieve G-coefficient value of .92. In other words, there is now a prospect for effectively downsizing the test without sacrificing its reliability to a great extent. However, it is difficult to decide from the GENOVA data alone, which of the 60 items should be maintained and which ones replaced or discarded.

At this next stage of the study, it was felt that QUEST analysis would be useful for solving this problem as it looks into the characteristics of individual items.

5.3.2 SPOT analysis 2: QUEST

QUEST estimates 1) the difficulty of individual items and 2) the ability of each candidate on the same scale and on a single "map". A summary map generated from QUEST is shown as Figure 3 below.

Figure 3. SPOT difficulty map

```
 5.0                  |
                  XXX |
                      |
                      |
                      |
                      |
                   XX |
 4.0                  |
                      |
                   XX |
                      |
                      |
                    X |
                      |   42
 3.0                X |   15
                  XXX |
                      |
                    X |
                  XXX |
                      |
                  XXX |
 2.0               XX |   20    43
                   XX |    5
                      |    2    32    38
                   XX |    4
                    X |   37
                    X |   17    30
 1.0             XXXX |   39
                  XXX |   40    52
                XXXXX |   21    24    54    59
                      |   58
                 XXXX |   28    47    57
                    X |   50
                 XXXX |   25    45
  .0            XXXXX |    6    13    19    48    56
                    X |    1    11    16    60
                  XXX |
                 XXXX |   12    14    26    33
                 XXXX |   31    44
                  XXX |   36    51
                   XX |    3
-1.0              XXX |   22    49    53    55
                      |
                      |   10
                    X |    8    29
                      |   41
                      |   27    46
                    X |
-2.0                  |    9
                      |    7    18    23    34    35
                      |
                      |
                    X |
                      |
                    X |
-3.0                  |
Each X represents 1 student
```

As mentioned in Section 4, QUEST provides information on both item difficulty and candidate performance. On the far left-hand side of the map is the common scale indicated in logits. Each of the Xs scattered along the scale in the second column represents one student. It indicates that the higher on the scale the student is, the better he or she performed in the test.

The numbers 1 to 60 in the right-hand side column represent the item numbers from SPOT. For the items, the greater the score given to it, ie. the higher up on the scale it is located, the more difficult the item is.

It should be noted that item difficulty and person ability are indicated in terms of probability. For example, the 0 level, highlighted in a bold type on the map, shows that there are five Xs to the left-hand side of the centre axis, and on the right-hand side of it are items 6, 13, 19, 48 and 56. These figures and symbols signify that the five candidates were all given the same logit score of 0 or average level, and that the difficulty of the five items are also 0. It means that these five candidates have a 50% chance of getting the five items right, as both all candidates and all items are given the same logit score of 0. For those five candidates, they have a lesser chance of getting an accurate answer on the items that appear higher up on the map, such as items 20 and 43, which have been allocated the logit score of 2. On the other hand, they will have a much greater chance of success with items such as 7, 18 etc. that are found towards the bottom end of the map.

As evident from Figure 3, the Xs are widely spread along the scale, ranging from a little above - 3.0 up to almost 5.0. This indicates that SPOT was successful in discriminating the candidates well. However, if the spread of the item difficulties is compared with that of the candidates, it is apparent that there are very few items that are difficult enough to match the ability of those candidates who have scored over 2. To help elaborate this, one can look at 0 on the scale again. There are five items (6, 13, 19, 48 and 56) which means there are many items at this level of difficulty. However, at 3 on the scale, there is only one item, #15. This means that the current SPOT has as many as five items to examine candidates' ability at 0 level, but has only one item for level 3. In other words, it is suggested that the current version of SPOT does not have enough difficult items. It would be more desirable if it had more difficult items and fewer easier items, so that the test would have a more even spread of item difficulty. More concretely, some of the easy items could be replaced with more difficult ones.

In summary, QUEST analysis could provide key information to improve the overall test by replacing easy items with more difficult ones.

5.4 SKAT

5.4.1 SKAT analysis 1: GENOVA

The SKAT data were also analysed using GENOVA first. The G-coefficient of the current test with 30 items (5 items x 6 tasks) was .85, indicating that the test distinguished different levels of ability relatively well. GENOVA shows that the reliability can be further increased by adding more items. To increase the reliability of the test to .9 level, for example, GENOVA suggests that SKAT needs to have 45 or more items (see Table 3).

Table 3. Generalizability coefficient for SKAT

Number of items	Generalizability coefficient
30	**0.85**
35	0.87
40	0.89
45	0.90
50	0.91

Whether or not it is wise to increase the reliability by adding items should be considered in relation to test time. As the purpose of this placement project is to establish a timesaving placement procedure, the reliability of the test and time required for the test must be carefully considered.

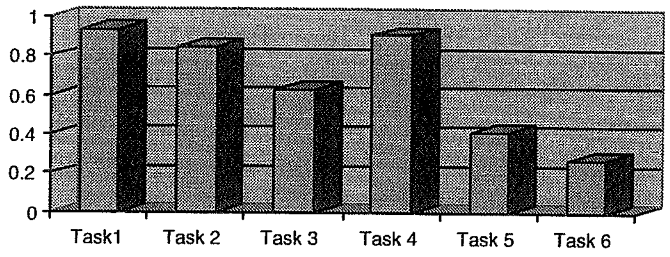
A mean value for each task and the grand mean value or the overall average were also calculated (Table 4).

Table 4. SKAT mean values

	Q1	Q2	Q3	Q4	Q5	Mean
Task 1	0.96	0.96	0.99	0.9	0.89	0.94
Task 2	0.9	0.85	0.94	0.8	0.77	0.85
Task 3	0.95	0.73	0.67	0.35	0.44	0.63
Task 4	0.82	0.91	0.98	0.95	0.91	0.91
Task 5	0.44	0.24	0.49	0.41	0.48	0.41
Task 6	0.28	0.21	0.38	0.26	0.21	0.27
G. Mean						**0.67**

The correct answer mean value of each task indicated an interesting result, that is, each task was different in the level of difficulty, except for an overlap between tasks one and four. The mean values for tasks one to six were .94, .85, .63, .91, .41 and .27 respectively. As a placement test, it is ideal to have tasks of different levels of difficulty in order to discriminate students of various language levels. In this sense, the result was desirable, except for the overlap of the items in tasks one and four in terms of difficulty (Figure 4).

Figure 4. SKAT task difficulty



Here, several possible ways to avoid such overlaps, including deletion of, or modification to the difficulty of one of the tasks could be suggested. In the case of the current test, since tasks one and four are aimed at measuring different aspects of awareness, keeping both, but altering the difficulty of one of the two tasks, would provide a more comprehensive view of test-takers' *kanji* awareness.

5.4.2 SKAT analysis 2: QUEST

QUEST showed that the reliability estimate of the test was .86, which supported the GENOVA result that the test was reliable.

The summary map of QUEST shows the relationship between the difficulty of the test items and the ability of the candidates. The map tells us that the test had some items that were too easy for the students (eg. items 3 and 18). The numbers on the right-hand side of the map show the item numbers (Figure 5).

Figure 5. SKAT difficulty map

```
 5.0               XXX |
                       |
                       |
                       |
                XXXXXX |
 4.0                   |
                       |
                   XXX |
                       |
                       |    27    30
 3.0                   |    22
                    XX |    29
                       |    26
                XXXXXX |
                       |
               XXXXXXX |    14
 2.0                   |    28
                XXXXXX |    24
                       |    15    21
                   XXX |    25
                       |    23
                 XXXXX |
 1.0              XXXX |
                       |
                 XXXXX |
                       |
              XXXXXXXX |    13
               XXXXXXX |
                       |
  .0            XXXXXX |    12
                       |    10
                   XXX |
                  XXXX |     9    16
                       |
                    XX |     7
-1.0                   |
                       |
                       |     4     5     6
                     X |    17    20
                       |
                       |
-2.0                   |     8
                       |    11    19
                       |
                       |     1     2
                       |
                       |
-3.0                   |    18
                       |
                       |
                       |
                       |     3
                       |
-4.0                   |
```

Each X represents 1 student

As each task had 5 items, 1 - 5 belonged to task one, 6 - 10 belonged to task two, and so on. The items from task one are all placed near the bottom of the map, which means that they were all very easy, and task two, relatively easy. As expected from the high correct answer rate for task four, the items in this task were all very easy. The items in tasks five and six were more difficult and appeared near the top of the map. The map also shows a gap near 1.0 level on the right-hand side, indicating a lack of items falling in this range.

Task three showed an interesting distribution. The items of this task were scattered widely, while the items of the other tasks were gathered together closely. This tells us that these items in task three were not at the same level of difficulty, and scores of the test might vary according to the specific *kanji* that students are asked to decompose or break down to components. This is also something that needs to be considered when revising the test.

The major findings here are as follows: firstly, the items within each task banded together, and secondly, that the difficulty level of each task was different. The findings indicate that SKAT has the potential to provide a yardstick for measuring learners' *kanji* awareness at varying developmental stages.

6. Conclusion

This paper illustrated the two newly developed tests, SPOT and SKAT, which have been employed in the Japanese Program at the University of Melbourne, and discussed the findings from the analyses on the two tests focusing on their reliability. In the course of the analyses, the two test analysis packages successfully complemented each other, and showed that they were powerful tools for supplying relevant data in order to revise the tests. In summary, the findings suggest that both tests are highly reliable, and that they could be further improved through revisions. Specifically, SPOT can be downsized without losing its reliability by removing some easy and overlapped items, and the item difficulties could be evenly spread by adding more difficult items. SKAT could be improved by deleting some easy items and adding some items of average difficulty instead, and also by amending some items to comply with the average difficulty level of each task.

In order to use SPOT and SKAT as a comprehensive placement battery, their validity also needs to be further investigated by comparing the SPOT and SKAT scores with the results from other conventional tests, such as listening, grammar, vocabulary and reading tests.

University Japanese programs throughout Australia are now under growing pressure to deal with large numbers of students with a limited number of staff and resources. A combination of SPOT and SKAT has the potential to be developed into a simple and yet credible placement means that would provide a solution for Japanese language course providers who face similar problems.

References

Adams, R. J. and S.T. Khoo. 1993. *Quest: the interactive test analysis system* [computer program]. Hawthorn: Australian Council for Educational Research.

Crick, J. E. and R. I. Brennan. 1984. *GENOVA: a general purpose analysis of variance system*. Version 2.2. Iowa City, IA: American College Testing Program.

Cronbach, L. J., Gleser, G. C., Nanda, H. and Rajaratnam, N. 1972. *The dependability of behavioral measurements: theory of generalizability*. New York: John Wiley.

Dobson, A. 1997. Reading strategies of Japanese L2, Paper presented at JSAA Conference. The University of Melbourne.

Ford-Niwa, J. 1998. The two components of language performance and what SPOT measures. [In Japanese] *Development of SPOT (Simple Performance-Oriented Test) for the Purpose of Placing Japanese Language Students*. Report 3: 53-58.

Ford-Niwa, J. 1997. An attempt to measure language proficiency - on the construct validity of SPOT. [In Japanese] *Development of SPOT (Simple Performance-Oriented Test) for the Purpose of Placing Japanese Language Students*. Report 2: 39-49.

Ford-Niwa, J., Kobayashi, N., and Yamamoto, H. 1995. What does a "Simple Performance-Oriented Test" measure?. [In Japanese] *The Journal of Japanese Language Teaching*, 86: 93-102.

Hashimoto, Y. 1999. Use of SPOT at the University of Melbourne. [In Japanese] Paper presented at a seminar in Melbourne Institute of Asian Languages and Societies, The University of Melbourne.

Hashimoto, Y. 2000. Relations between SPOT scores and course achievement: Analysis of the University of Melbourne students. [In Japanese] *Journal of Japanese Language Teaching*, 15: 87-97.

Tsukuba, Japan: International Student Center, University of Tsukuba.

Hatano, G. 1986. How do Japanese children learn to read?: orthographic and ecocultural variables, In B.R. Foorman & A. W. Siegle (eds.*), Acquisition of Reading Skills: Cultural Constraints and Cognitive Universals.* Hillsdale, Lawrence Erlbaum Associates, 81-114.

Hatasa, Y. and Tohsaku, Y. 1997. SPOT as a placement test. *Development of SPOT (Simple Performance-Oriented Test) for the Purpose of Placing Japanese Language Students.* Report 2: 5-20.

Kobayashi, N., Ford-Niwa, J. and Yamamoto, H. 1996. A new method for testing proficiency in Japanese as a second/foreign language: SPOT. [In Japanese] *Japanese-Language Education Around the Globe,* 6: 201-218.

Koda, 1999. Development of L2 intraword orthographic sensitivity and decoding skills. *The Modern Language journal,* 83 (1): 51-64.

Lynch, B. K. and T. F. McNamara. 1998. Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing,*15 (2): 158-180.

McNamara, T. 1996. *Measuring Second Language Performance.* New York: Addison Wesley Longman Limited.

Murakami, K. 1998. Use of SPOT at the Nagoya University. [In Japanese] *Development of SPOT (Simple Performance-Oriented Test) for the Purpose of Placing Japanese Language Students.* Report 3: 94-99.

Okita, Y. 1995. *Kanji* learning strategies and student beliefs on *Kanji* learning [in Japanese], *Japanese–Language Education Around the Globe,* 5: 105-124.

Shavelson, R. J. and N. M. Webb. 1991. *Generalizability theory: a primer.* Newbury Park, CA: Sage.