The predictive validity of SPOT and self-assessment questionnaire

Yoshinori Sasaki
Ochanomizu University

Abstract

This study intends to assess the predictive validity of the Simple Performance Oriented Test (SPOT) and self-assessment questionnaire as against Japanese language learners' academic performance. These two measurement tools ("tests") were repeatedly administered to second-year Japanese language students at an Australian University at the beginning and end of the academic year. While most students increased their score in SPOT after a year's study, many students attained a lower score in the second self-assessment questionnaire than in the first. Moreover, the correlations between the first and second self-assessment levels were surprisingly low. The SPOT score had a high correlation with the academic score, whereas the self-assessment and the academic performance scores mostly had only non-significant correlations. A factor analysis also revealed that self-assessment and objectively-measured Japanese language skills each represent different mental traits.

1. Introduction

Traditionally, Japanese course placement tests were often developed by each individual institution based on its curriculum. Such curriculum-dependent tests alone will not allow for a long-term accumulation of incoming students' proficiency data assessed against a commonly accepted criterion. This precludes comparative quantitative analyses over years and/or across institutions.

This lack of comparability is regrettable, because in the absence of an objective measure to evaluate a curriculum's efficacy, teaching methods tend to regress into "cliques" which merely advocate proponents' unsubstantiated "beliefs". In order that a teaching methodology can continue to progress as a solid empirical science

Melbourne Papers in Language Testing 2000 Volume 9.2 pp. 30-55. The Language Testing Research Centre, The University of Melbourne.

and technology, it is essential to prepare opportunities to objectively assess instructional efficacy so that a dialectic mutual facilitation will be established between discussions in teaching philosophy, inventions and descriptions of teaching skills, and qualitative/quantitative measurements of instructional effectiveness.

Fortunately, an increasingly large number of institutions incorporate curriculum-independent proficiency tests as a part of their placement test battery. Many of those proficiency tests are based on learners' "objective performance". Among them, Simple Performance-Oriented Test (SPOT) recently developed at Tsukuba University attracts attention among professionals of Japanese as a second language (JSL) by virtue of its high empirical validity reported so far (Hatasa & Tohsaku, 1997) along with its ease to administer in group within 10~15 minutes. This test also stands out above its competitors because its development process has been reported in detail.

SPOT requires testees to listen to a series of mutually unrelated tape-recorded sentences, and to complete their transcriptions on the answer sheet by filling in a one-syllabary gap in each sentence. Table 1 presents the ten SPOT exercise sentences. (Many real SPOT sentences are more complex than these.)

（1）　どう（　）よろしく。

（2）　ここは静（　）ですね。

（3）　おはよう（　）ざいます。

（4）　わたし（　）たなかです。

（5）　ごはんを食（　）ました。

（6）　どこから（　）ましたか。

（7）　あしたここに来ます（　）。

（8）　ぜんぶ（　）いくらですか。

（9）　タクシーでいきま（　）ょう。

（１０）あたらし（　）車を買いました。

Table 1: SPOT item examples

Another new trend in placement tests is self-assessment (Leblanc & Painchaud, 1985), namely, learners' self-assessed proficiency level, an approach which is motivated by the autonomous learning movement.

Out of this spirit, currently the University of New South Wales (UNSW) in Australia employs Japanese Self-Assessment Questionnaire (Kinoshita-Thomson, 1995), an adaptation of ISLPR[1] (International Second Language Proficiency Ratings) (Wylie & Ingram, 1993; Quinn & McNamara, 1987) to JSL learners, as a part of its Japanese course placement test battery.

This self-assessment questionnaire presents descriptions of eight levels[2] of Japanese language proficiency, each in four macro-skill areas ("speaking", "listening", "writing" and "reading") (See Table 2 for an example). A learner is required to choose the best description that most closely approximates her/his own level of proficiency in each macro-skill area. These scales do not suppose that the interval between two adjacent levels are identical: in other words, self-assessment scales are, at best, ordinal scales, but not equal interval scales.

Level 2. [Level 1. in Kinoshita-Thomson (1995)]
I understand in very simple face-to-face conversation in Japanese about very familiar things (eg. how long and where I have studied Japanese) provided the other person uses simple sentences, speaks slowly, repeats or rewords things to help me, and the utterance isn't confounded with socio-culturally unique context.

Table 2: An example of a "listening" proficiency level description in the Japanese self-assessment questionnaire

The present study primarily aims to assess how accurately these two measurement tools (SPOT and self-assessment questionnaire) can each predict a learner's subsequent academic score in a Japanese course. This provides an extension of past efforts to validate SPOT as a placement test.

[1] ISLPR was originally called ASLPR (Australian Second Language Proficiency Ratings)

[2] Level 0 (zero proficiency) of ISLPR was excluded from the questionnaire because it is an unlikely score in a placement test.

The present study also examines how much progress learners reveal when they take these tests repeatedly at the beginning and the end of one academic year. In other words, it explores their uses as a tool to measure instructional effects or proficiency progress.

Meanwhile, it is natural to expect a certain degree of correlation of the test scores conducted at the beginning and the end of an academic year, given the fact that foreign language learning involves accumulation of skills and knowledge over academic terms. Therefore, the present study also examines their test-retest correlations.

Research Questions

- How high are the correlations between learners' SPOT and self-assessment scores on the one hand and their subsequent academic scores on the other?
- How much do learners improve their test scores after an academic year's instruction? How much test-retest correlations of scores are there?

2. Method

Location/Institution

The study was conducted at the University of New South Wales (UNSW) in a suburb of Sydney, which hosts the highest ratio of foreign students in Australia. Its Japanese program offered two academic terms of instruction every year, and each term was 14-weeks long. (An academic year and a calendar year mostly coincide in Australia.)

Procedure

Self-assessment questionnaire (Week 1) and SPOT (Week 3) were conducted in class in the first (autumn) semester of the 1998 academic year (Table 3). Absentees took them later in an instructor's office. In Week 12 of the second (spring) term, SPOT and Self-Assessment Questionnaire were conducted again. In other words, students took the same proficiency test after one academic years' instruction (i.e., half a calendar year).

	Timing	1998 academic year		
		Term 1		Term 2
		Beginning	End	End
Proficiency assessment	SPOT	V		V
	Self-assessment	V		V
Academic test	Term-final oral test		V	V
	Term-final written test (reading/composition/others)		V	V

Table 3: Data Collection Schedule

SPOT

Students listened to SPOT version 3B tape, which is the easiest among the 4 versions which were available in those days. It took approximately 10 minutes to conduct, including distribution and collection of test sheets, instruction, and replaying the tape.

Japanese Self-Assessment Questionnaire

A Japanese Self-Assessment Questionnaire comprising eight levels of proficiency descriptions each in four macro-skill areas was distributed to each student. The description and format of the questionnaire largely followed Kinoshita-Thomson (1995), with minor modifications as follows:

- Instructions on the cover page were slightly paraphrased.
- The sequence of macro-skills in the questionnaire was reordered.
- Proficiency levels (1-, 1, 2, 3, 3+, 4, 4+, 5 from the lowest) were recoded, each with an integer (1~8 from the lowest), so that statistical analyses are easier.

Most students spent 15~20 minutes to answer, whereas no specific time limit was set in advance.

Data Collection and Analyses

Results from the above-mentioned procedures along with students' biographical and academic performance data were recorded in a student record database (FileMakerPro for Macintosh). To perform statistical analyses, data were exported from this database to packages such as SPSS for Mac.

Student Backgrounds

As mentioned above, two types of proficiency assessments (SPOT and self-assessment) were conducted each twice. Namely, there were four assessment sessions altogether within the academic year, each of which some students missed. Of the 97 students who participated in the first proficiency assessments and later received the first-term course grade, we successfully collected the full second-term data from 43. The results summarized below represent these 43 participants who participated in all four sessions and received two Japanese course grades, unless otherwise specified.

3. Results

What follows is the report of (1) comparison of first and second assessment results, (2) predictive validities against students' Japanese course performance (multiple regression analyses), and (3) internal correlations and factor analysis results, in this order. In some parts non-parametric analyses are cited along with parametric analyses on the grounds that self-assessment scales are not equal interval scales.

Longitudinal Comparisons of Assessment Scores

Table 4 shows SPOT and self-assessment scores in the first and second trials. Overall, there were gains of average and median scores in both. At an individual level, however, statistically significant correlations between the first and second trial scores were found only in SPOT ($r = 0.807$, $p < 0.0001$) and Self-Assessment "Listening" ($r = 0.499$, $p < 0.005$).

		Average score (S. D.)		Tests of change		Correlation Coefficients	
		Term 1	Term 2	*t*-test Score increase *t* *p* (one tailed)	Wilcoxon test *z* *p* (one tailed)	Pearson *r* *p* (two tailed)	Spearman Rho z *p* (two tailed)
SPOT (3B)		25.186 (12.223)	36.698 (10.643)	11.512 9.787 0.01%	5.3907 0.01%	0.781 0.01%	0.807 5.232 0.01%
SA (8 is highest)	Speaking	2.581 (1.074)	2.977 (0.938)	0.395 2.004 5%	2.0661 5%	0.179 n.s.	0.186 1.205 n.s.
	Listening	2.721 (1.054)	3.395 (1.072)	0.674 3.921 0.02%	-3.2630 0.5%	0.437 1%	0.499 3.235 0.5%
	Writing	2.953 (1.29)	3.93 (0.985)	0.977 3.94 0.02%	-2.0191 5%	-0.059 n.s.	0.036 0.231 n.s.
	Reading	2.884 (1.159)	3.326 (0.919)	0.442 2.018 2.5%	-3.3845 0.5%	-0.0026 n.s.	0.062 0.401 n.s.

Rho and *z* values are tie-adjusted.
SA: Self-assessment

Table 4: Comparison of Test scores: the first-term beginning vs. the second term end (N=43)

Table 5 indicates the number of participants whose score improved, decreased, and stayed the same in their second assessment trial in each scale. Whereas more than 90% participants performed better in the second SPOT trial, only 46.5% ("speaking") ~ 62.8% ("listening") placed themselves at a higher level in the second self-assessment. In fact, quite a few participants placed themselves at a lower level in the second self-assessment. In particular, 25.6% of participants gave themselves a lower "writing" self-assessment score in the second trial. It is somewhat of a relief that in all 4 scales the number of participants who scored better in the second trial outnumbered those who performed worse, by a statistically significant margin (sign-test).

		Change of score/level			Sign test
		Up	Same	Down	Z *p* (one-tailed)
SPOT		39 (90.7%)	2 (4.7%)	2 (4.7%)	5.6223 .0000
Self assessment	Speaking	20 (46.5%)	16 (37.2%)	7 (16.3%)	2.3094 5%
	Listening	27 (62.8%)	11 (25.6%)	5 (11.6%)	3.7123 0.05%
	Writing	22 (51.2%)	10 (23.3%)	11 (25.6%)	1.7408 5%
	Reading	26 (60.5%)	11 (25.6%)	6 (14.0%)	3.3588 0.2%

Table 5: Increase/decrease of the test score from the first-term to the second-term

Predictive Validity

What follows is the regression equation, resulting from a multiple regression analysis to predict the term-final accumulative academic score in the first semester based on the proficiency assessment scores (SPOT and four self-assessment scales) conducted earlier in that semester:

$$y = 61.597 + 0.382\ x\ (R = 0.382, N = 97)$$

where
y: term-final accumulative academic score (0~100)
x: term-initial SPOT score (0~60).

Initially both SPOT and self-assessment score data were plugged into the statistical package, but the stepwise calculation procedure excluded self-assessment scores because they do not add significant contribution to improve the accuracy of prediction.[3] It should be

[3] Self-assessment scores had high correlations with each other (Table 12), as will be reported later. With an intention to avoid possible resultant multicolinearity (Hirose & Sasaki, 1994), we ran another multiple regression analysis, with the SPOT score and the total of the four self-assessment scores as two predictor variables. (We owe this insight to SASAKI Miyuki.) The result replicated the original one, where only the SPOT score again appeared as the sole statistically significant predictor.

noted that this particular statistical analysis is based on scores from all the 97 participants in the first-term assessments, to maximize the sample size.[4]

Tables 6~8 show details of the regression analysis.

Source	Degrees of freedom	Sum Squares	Mean Square	*F* test
Regression	1	2126.264	2126.264	16.359
Residual	96	12477.379	129.973	
Total	97	14603.644		

Table 6: Analysis of variance of the multiple regression equation to predict the first-term final academic score

Variable	Coefficient	Standard error	Standard coefficient	*F* to remove
Axis	61.597			
SPOT	0.382	0.094	0.382	16.359

Table 7: Variables retained in the regression equation to predict the first term final academic score

Variable	Partial correlation	*F* to enter
Self-assessment "Speaking"	0.096	0.882
Self-assessment "Listening"	0.016	0.025
Self-assessment "Writing"	-0.032	0.100
Self-assessment "Reading"	0.01	0.009

Table 8: Variables excluded from the regression equation to predict the first-term final academic score

We also ran a similar multiple-regression analysis where the score of the first-term final written examinations is the criterion variable. Again, the SPOT score was the only significant predictor variable. The regression equation is as follows:

[4] Results were similar when the same regression analysis was conducted for the 43 students who participated in all assessment sessions throughout the academic year.

y = 92.623 + 1.448 x (R = 0.614, N = 95)

where
y: term-final written examination score (0~200)
x: term-initial SPOT score (0~60).

Source	Degrees of freedom	Sum Squares	Mean Square	*F* test
Regression	1	30271.103	30271.103	56.869
Residual	94	50035.533	532.293	
Total	95	80306.637		

Table 9: Analysis of variance of the multiple regression equation predicting the first-term final examination score

Variable	Coefficient	Standard error	Standard coefficient	*F* to remove
Axis	92.623			
SPOT	1.448	0.192	0.614	56.869

Table 10: Variables retained in the regression equation predicting the first term final examination score

Variable	Partial correlation	F to enter
Self-assessment "Speaking"	-0.052	0.252
Self-assessment "Listening"	-0.06	0.336
Self-assessment "Writing"	0.022	0.043
Self-assessment "Reading"	0.018	0.031

Table 11: Variables excluded from the regression equation predicting the first-term final examination score

Internal Correlations and Factor Structure

Table 12 shows the correlations between scores of SPOT tests, self-assessments, and term-final academic test scores. Here "speaking" represents the score of an oral test performed toward the end of each semester, "writing" represents scores of short essay questions in term-final written examinations, and "reading" represents scores of passage-reading sections in the term-final written examinations.

			Academic test score: AT						SPOT		Self-assessment: SA							
			T1			T2			T1	T2	T1				T2			
			S	W	R	S	W	R			S	L	W	R	S	L	W	R
AT	T1	S	**1.0**	**.374**	.269	**.484**	**.490**	**.359**	**.432**	**.449**	.061	-.093	.036	.050	.009	.053	.051	.167
		W	**.374**	**1.0**	**.612**	**.582**	**.691**	**.623**	**.733**	**.693**	.256	.170	.162	.198	**.443**	**.310**	**.392**	**.341**
		R	.269	**.612**	**1.0**	**.552**	**.516**	**.698**	**.690**	**.564**	**.316**	.063	.277	.213	**.393**	.188	**.344**	**.334**
	T2	S	**.484**	**.582**	**.552**	**1.0**	**.654**	**.626**	**.504**	**.593**	.276	.167	.183	.134	.238	.114	.186	.021
		W	**.490**	**.691**	**.516**	**.654**	**1.0**	**.536**	**.636**	**.724**	**.363**	**.314**	**.357**	.258	**.301**	**.323**	**.375**	.120
		R	**.359**	**.623**	**.698**	**.626**	**.536**	**1.0**	**.620**	**.656**	**.308**	.032	.237	.108	**.357**	.089	.090	.015
SPOT		T1	**.432**	**.733**	**.690**	**.504**	**.636**	**.620**	**1.0**	**.781**	.238	.095	.286	.280	.233	**.360**	**.313**	**.416**
		T2	**.449**	**.693**	**.564**	**.593**	**.724**	**.656**	**.781**	**1.0**	**.318**	.270	.256	.115	**.381**	**.441**	.266	.171
SA	T1	S	.061	.256	**.316**	.276	**.363**	**.308**	.238	**.318**	**1.0**	**.736**	**.610**	**.673**	.180	**.333**	.190	.186
		L	.093	.170	.063	.167	**.314**	.032	.095	.270	**.736**	**1.0**	**.636**	**.656**	.162	**.437**	.121	.111
		W	.036	.162	.277	.183	**.357**	.237	.286	.256	**.610**	**.636**	**1.0**	**.793**	.156	.115	.059	.091
		R	.050	.198	.213	.134	.258	.108	.280	.115	**.673**	**.656**	**.793**	**1.0**	.119	.134	.114	.003
	T2	S	.009	**.443**	**.393**	.238	**.301**	**.357**	.233	**.381**	.180	.162	-.156	-.119	**1.0**	**.530**	**.534**	**.307**
		L	.053	**.310**	.188	.114	**.323**	.089	**.360**	**.441**	**.333**	**.437**	.115	.134	**.530**	**1.0**	**.446**	**.410**
		W	.051	**.392**	**.344**	.186	**.375**	.090	**.313**	.266	.190	.121	.059	.114	**.534**	**.446**	**1.0**	**.447**
		R	.167	**.341**	**.334**	.021	.120	.015	**.416**	.171	-.186	-.111	-.091	-.003	**.307**	**.410**	**.447**	**1.0**

Table 12. Correlations between academic and placement test scores

S: Speaking; L: Listening; W: Writing; R: Reading; T1: Term 1; T2: Term 2; AT: Achievement test score; SA: Self-assessment test score
Correlations significant at the 5% level (two-tailed) are bold-faced

The first and second SPOT scores are each highly correlated with written and oral academic test scores in both academic terms. On the other hand, first and second self-assessment scores have only low correlations with others, whereas self-assessment scores conducted in each term provided moderately high internal correlations. These imply that Japanese proficiency measured by objective performance and by self-assessment questionnaire each represent distinct traits, which do not share a lot.

Factor analysis results supported this interpretation. Tables 13 and 14 show the results of factor analysis of the above-mentioned data, followed by varimax rotation.

	Eigen value	Percent of variance	Cumulative percent of variance
Factor I	6.20874	33.8%	33.8%
Factor II	2.90982	18.2%	52.0%
Factor III	1.90822	11.9%	63.9%
Factor IV	1.07669	6.7%	70.6%

Table 13: Eigenvalues and explained variance

			Factor loading			
			Factor I	Factor II	Factor III	Factor VI
Academic score	T1	Oral test	**0.58062**	-0.11086	-0.00775	**-0.61689**
		Writing test	**0.77788**	0.08474	0.35149	-0.00499
		Reading test	**0.74242**	0.12375	0.27090	0.09660
	T2	Oral test	**0.82189**	0.09229	-0.09057	0.00560
		Writing test	**0.76675**	0.26388	0.19633	-0.09610
		Reading test	**0.87333**	0.06145	-0.07796	-0.17234
SPOT score	T1		**0.74735**	0.15537	0.40720	-0.25348
	T2		**0.82016**	0.14650	0.14650	0.14650
Self assessment	T1	Speaking	0.28052	**0.83422**	-0.04277	0.18544
		Listening	0.03877	**0.88264**	0.08735	0.16645
		Writing	0.17066	**0.89219**	-0.00247	-0.12241
		Reading	0.06630	**0.88966**	0.08203	-0.19277
	T2	Speaking	0.38696	-0.14719	**0.51115**	**0.63762**
		Listening	0.14614	0.23682	**0.73128**	0.17016
		Writing	0.20098	0.06361	**0.71572**	0.22497
		Reading	0.06845	-0.17945	**0.84413**	-0.28514

Factor loadings greater than +/-0.5 are bold-faced
T1: Term 1; T2: Term 2

Table 14: Factor loadings (after varimax rotation)

Factor I is heavily loaded with academic test scores and SPOT, which is appropriate to interpret as "Japanese language ability measured by objective performance". Factors II and III are each heavily loaded with Self-Assessment scores conducted in the first and second semesters, which suggest that they represent "first-semester self-assessment" and "second-semester self-assessment" respectively. These three factors collectively account for nearly 65% of the entire variance.

Meanwhile, Factor IV, which is correlated negatively with Oral Test scores in the first semester and positively with self-assessment "speaking" level in the second semester, is difficult to interpret.

Summary of Results

SPOT

- The score of SPOT conducted early in the first semester provided a fairly accurate prediction of academic achievements in the same and the following semesters.
- First and second SPOT scores, which were repeatedly conducted with an interval of roughly half a year (i.e., one academic year's instruction), provided a high correlation of approximately 0.8.
- More than 90% of participants scored better in their second SPOT trial, than in their first performance half a year earlier.
- Group statistical comparisons also indicate that the average score of the second, academic-year-end SPOT trial was higher than for the SPOT trial at the beginning of the academic year.

Self-Assessment Questionnaire

- Self-Assessment scores conducted early in the first semester were little correlated with academic scores in the same and the following semesters.
- In most instances, self-assessment scores of individual participants conducted with an interval of half a year (i.e., one academic year's instruction) were not significantly correlated. The only exception was the self-assessed "listening" score, which provided a significant correlation larger than 0.4.
- Self-assessment scores of different macro-skills conducted at the same time provided correlations ranging approximately from 0.4 to 0.8.

- Roughly half of the participants placed themselves at a higher level in their second self-assessment trial, in comparison with their first self-assessment earlier in the same academic year. However, there were also more than a negligible number of students who placed themselves at a lower level in the second assessment.
- In group statistics, the average self-assessment scores in the second trial were higher than the first by statistically significant margins.

Relation between SPOT and self-assessment

- In general, correlations between SPOT scores and self-assessment scores were low.

Factor analysis yielded the following three factors which were deemed interpretable:

- Factor I) A factor of Japanese language ability measured by objective performance (highly correlated with SPOT and academic test scores)
- Factor II) A factor of self-assessment conducted in the first semester
- Factor III) A factor of self-assessment conducted in the second semester

4. Discussion

SPOT

SPOT, despite its ease to implement in group within a short time period demonstrated its high predictive validity. This result is consistent with earlier studies which support the practical utility of SPOT. Whereas Kobayashi (1997: p. 6) says that SPOT "is not likely to correlate with in-class achievement test scores" (translation is by this researcher), the actual correlation was unexpectedly high (Table 12). Regression analyses showed that SPOT provided a significant predictor variable of the first-term accumulative academic score (Tables 7 and 10). Although the correlation of R = 0.382 of the regression equation is not amazingly high, this is not surprising because the accumulative academic score took into account some elements beyond the scope of SPOT, such as Kanji (Chinese characters) knowledge, attendance, and homework submission. In fact, a regression analysis with the first-term final written examination score as the criterion variable yielded a high correlation of R = 0.614. Therefore, we can consider that the predictive validity of SPOT is fairly high; the test yielded valuable information as a predictor of learners' subsequent academic success in Japanese language learning.

For the sake of fairness, however, it needs to be noted that the primary mission of a placement test is discrimination ("discriminatory predictive validity"): namely, a test ought to predict whether a certain testee is prepared to smoothly study (in particular, whether s/he will fail the course or not) when s/he is allocated to a certain level. Teachers do not normally expect a placement test to predict students' term-final score per se, as was attempted in the present study. For example in the typical Australian grading system, a placement test has to discriminate students who will end up with the final score of 45 (failure) versus 70 (pass), but it is not essential to discriminate between those who will end with 70 versus 95.

For this reason, it would be ideal from a research viewpoint to run discriminatory analyses, with the SPOT score as a predictor variable and the term-final "pass" or "failure" as the criterion variable, so that the discriminatory predictive validity of SPOT as a placement test can be fully attested. However, this design was not feasible in the present research because all 43 students who were subject to analyses passed the courses.

At any rate, it would be rare, if ever, that a test whose score ranges widely (Table 4) has no correlation with the term-final academic score and still provides high predictive validity for placement purposes only. Thus we can at least take the results from the present study as "circumstantial evidence" for the discriminatory predictive validity of SPOT.

It should also be noted that most placement test takers in the real world are those who have not completed the lower-level courses at the institution, whereas the majority of participants in the present study had studied Japanese at the same institution. This condition difference needs to be taken into account to interpret the results outlined above.

SPOT also yielded a rather high test-retest correlation despite the intervening Japanese language instruction for two academic terms, experience which should have resulted in students' progress which in turn should have yielded a certain amount of variance. The high correlation which was nevertheless observed suggests that learners' ranks of Japanese language ability measured by SPOT is rather stable over time.

At the same time, most participants scored better in the second, year-end SPOT trial, which hints that this progress in some way reflects the instructional effects of the curriculum. However, lack of a control group in the present study makes it impossible to tease apart genuine instructional effects and other factors, such as an increased familiarity with the SPOT test format in the second trial.

Also, if the gain in the SPOT score reflects some form of progress in their Japanese language ability, it still needs further research to see whether it is an index of competence which the majority of Japanese teachers and Japanese language learners consider important. This reminds us of the lack of agreement among researchers on the construct validity (Ford-Niwa, 1997) and face validity (Spence-Brown, 1997) of SPOT, as opposed to their generally high regards on its statistical reliability and practical usefulness. This gap presents another important research topic for SPOT researchers.

Self-Assessment Questionnaire

On the contrary, the Self-Assessment Questionnaire yielded unexpectedly low test-retest correlation and predictive validity. This provides a sharp contrast with Leblanc and Painchaud (1985), who report high validity of their self-assessment questionnaire for placement.

One possible reason for this discrepancy is the difference of the two questionnaires' fidelity. Leblanc and Painchaud (1985) required learners to answer a number of mutually independent (discrete) question items whose total score indexed their proficiency. Bachman and Palmer (1989) and Ross (1998) also report high reliabilities of multiple-item self-assessment tests. On the other hand, Japanese Self-Assessment Questionnaire requires a learner to choose one of eight levels in each scale (macro skill), ranging from novice to near-native, which best fits her/his own level of proficiency. From a psychometric viewpoint, the former style (multiple items) collects a larger amount of information, and thus it is expected to provide a finer measurement. Incidentally, SPOT Version 3B (for beginning learners), whose high predictive validity the present study supports, can also offer a fine-grained score ranging from 0 to 60 points.

However, these "coarse-grained" self-assessment scales still revealed higher internal correlations (Table 12). This fact suggests that the above-outlined psychometric reason alone does not fully account for the discrepancy between Leblanc and Painchaud (1985) and the present study.

Another possible reason for the low predictive validity of Japanese Self-Assessment Questionnaire in the present study involves the difference of specificity of performance descriptions. Leblanc and Painchaud (1985) provide very specific descriptions of the target performance. For example, the three items cited in Table 15 specify genre/media ("over the phone"), situation ("an advisor explains to me in French how to register for an elective course") and target behavior ("understand what page to open") as criteria of assessment.

1. I understand a professor who tells me in French what page to open my book.
2. Over the phone I can understand some basic information in French such as the name of the caller and the number where he can be reached.
3. If an advisor explains to me in French how to register for an elective course, I will understand provided some details are repeated.

Table 15: Examples of "Listening" skill descriptors in Leblanc and Painchaud (1985: 684)

In comparison with this and some other self-assessment questionnaires (Table 16), descriptions in the Japanese Self-Assessment Questionnaire are generic (Table 2), which could be subject to different interpretations across individuals and/or over time. In fact, a sizable number of testees gave themselves lower self-assessment scores in the second trial (Table 5), which suggests that learners interpreted level descriptions differently at each time. Ross (1998) states that: "... episodic memory of using particular skills in the classroom experience would enhance the accuracy of self-assessment" (p. 16), and such episodic memories can be more easily and consistently invoked in response to context-and-task-specific performance descriptors.

	Test settings		Test format		Result: Self assessment as an indicator of objective L2 performance
	Target language	Linguistic environment surrounding testees	Description of proficiency	Number of test items	
Bachman & Palmer (1989)	English	Salt Lake City, Utah, USA (English monolingual)	Generic	Multiple items	OK
Blanche & Merino (1989)	French	Quebec, Canada (French English bilingual)	Task/ context specific	Multiple items	OK
Pierce, Swain & Hart (1993)	French	Toronto (Predominantly English)	Task/ context specific	Multiple items	Low correlation
Ross (1998)	English	Japan (Japanese monolingual)	Task/ context specific	Multiple items	OK
Present study	Japanese	Sydney (English monolingual)	Generic	One item per macro-skill (scale level choice)	Low correlation

Table 16: Summary of some self-assessment validation studies

The difference of the environment surrounding the testees should also be noted. Leblanc and Painchaud (1985) was conducted in Quebec, a homeland of a huge French-English bilingual population, where learners have daily opportunities to be exposed to the target language (French) outside of the classroom. On the other hand, the present study was conducted in Sydney, Australia. Even though there are many Japanese tourists and residents in Sydney, exposures to real-life Japanese are not easily available on many streets of the city. In this environment, learners may not have had communicative experiences against which they could assess their Japanese language proficiency. In this connection, Peirce, Swain and Hart (1993) in their Toronto-based study also report low correlations between learners' French self-assessment questionnaire scores and objectively tested proficiency, and they too invoke learners' lack of communication experiences in French as a possible source of this low validity.

In sum, the three studies summarized in Table 16 which report a reasonable or higher correlation between self-assessment and objective performance score satisfy at least one of the following two conditions: 1) Testees had much exposure to the target language outside of the class; and 2) performance descriptions are task/context-specific. Two among those three studies (Bachman & Palmer, 1989; Blanche & Merino, 1989) satisfy both. Ross (1998) is seemingly an exception (the English self-assessment was conducted in Japan), but it should be noted that his version of self-assessment questions somehow resembled an achievement test: his self-assessment descriptors were specifically based on the teaching materials the learners studied, and thus the learners had plenty of chances to assess their own proficiency in those particular skills without leaving the classroom.

Also, there has been a debate over the construct validity of ISLPR (Davies, 1995), based on which scales the Japanese Self-Assessment Questionnaire (Kinoshita-Thomson, 1995) was developed. In response to Quinn and McNamara's (1987) concern about ISLPR's rating variability, Ingram (1995: 27) cites some studies which support its reliability and validity. In making a case for the scales, he claims that "... the authors of the ASLPR [= an earlier name of ISLPR] consider that training in the scale and its use is essential". This rater training is crucially missing in self-assessment procedures. Indeed lack of rater training in self-assessment triggers a doubt about the consistency of scale description interpretations, which doubt is reinforced by the above-reported score fluctuation over time (Table 5).

Also, some testees expressed difficulty choosing a level which best fits their current proficiency, on the ground that some descriptors in one level and some in another level seemed to characterize their performance most properly. In other words, the ISLPR to their minds did not appear to provide fully ordinal scales.

To make the matter worse, powerful modern test theories (e.g., Rasch models) are difficult to apply to analyze a self-assessed (i.e., a testee and his/her rater are the same person) multiple-proficiency-level-choice (i.e., only one "item" per macro-skill) questionnaire, because the testee, rater and item factors in the rating process are, by design, inseparable. This makes it more difficult for researchers to pinpoint the source of these variabilities on the strength of statistical tools.

All in all, it is necessary to continue development efforts to improve measurement tools and accumulate interpretation know-how, to establish self-assessment questionnaire as a valid Japanese language placement test. One possible approach is to develop an item bank of target performance descriptors, which supposedly represent (i. e., operationalize) each level of ISLPR scale, from which statistically reliable and valid items should be selected to yield a coherent questionnaire. Its field trials, in turn, will provide a psychometric validation of the ISLPR scales.

Beside such psychometric arguments, it is warranted to consider using self-assessment as a part of learning activities (Kinoshita-Thomson, 1997). It is particularly important for more advanced learners to distinguish between "what one can do", "what one cannot do" and "what one is about to be able to do", so that they can set their own next feasible and worthwhile learning goals. This requires a capacity to assess one's own level of language proficiency in relation to their own needs. Blanche and Merino (1989: 313) claim that:

"Self-assessment accuracy is a condition of learner autonomy. If students can appraise their own performance accurately enough, they will not have to depend entirely on the opinions of teachers, and at the same time, they will be able to make teachers aware of their individual learning needs."

In this connection, the present study showed that many learners assessed their language proficiency differently after one academic year's instruction. On the other hand, both the first and second self-assessment scores had only low correlations with objectively-measured proficiency (SPOT) and academic test scores. In other words, it is premature to expect that progress in language learning will automatically bring about an accurate self-assessment.

An extensive list of target performance descriptors each in a specific context, as are exemplified in Leblanc and Painchaud (1985), will motivate us to imagine a syllabus/curriculum in which those descriptors of target performance in relation to a certain requirement ("task") provide a building block of classroom activities. This proposal would deserve serious consideration if some day a sequence of tasks reflecting a certain scale (e.g., ISLPR) is established which indeed accurately illustrates a pathway of communicative skill development. (For further discussion on sequencing tasks or performance descriptors, see Bachman & Cohen, 1998; Brindley, 1998; Long & Crookes, 1993.) This possible tight coordination between assessment and instruction represents a potential merit of self-assessment, as opposed to some performance-referenced tests such as

SPOT, whose misuse might trigger unfavorable backwash effects[5] (Table 17).

		SPOT	Self-Assessment Questionnaire
Number of skill domains assessed separately		One (listening, reading and writing skills are involved in one task type)	Four (listening; speaking; reading; writing)
Number of versions (as of 1998)		Four (beginner ~ advanced)	One for each macro skill domain
Rating		Objective performance on a tape-based task	Self-evaluation by the learner herself/himself
Predictive validity		High	Low
Face validity		Questionable	Reasonable
Construct validity	Criterion	Not specified	ISLPR scale
	Validity of criterion	N/A	Debatable
Pedagogical applications		Difficult	Promising
Negative backwash effects		Possible	Not likely

Table 17: Comparison of SPOT and Self-Assessment Questionnaire

5. Future Issues and Concluding Remarks

The lack of accuracy of a placement test becomes most apparent when students find that a class which they are allocated to is too difficult, to the effect that they drop out or fail the course. On the other hand, if the class level they are allocated to is too low, the problem may lie under the surface even if they feel discouraged or take it too easy. In other words, it is difficult to quantify the opportunity cost of an under-placement, namely, "the progress the learner would have achieved if s/he had been allocated to a higher level course." To address this issue, it will be necessary to devise a more suitable research procedure where, for example, students of different course levels take the same placement tests.

[5] I owe this insight to Yukiko Hatasa.

Unfortunately, the present study only reports on students of a second-year Japanese course, which sampled learners with only a narrow range of proficiency. The number of samples (N = 43) is not very large as a study of this nature. To confirm the tenuous findings from the present study, it is necessary to run a large-scale replication which incorporates more advanced level learners. It should also be noted that the criterion variable of the present study, i. e., the academic course score, reflected the curriculum of the institution where the study was conducted. To demonstrate that the results are generalizable beyond that boundary, it is desirable to collect similar data at other institutions.

The statistical analyses employed in the present study are restricted to simple inferential statistics (e.g., simple correlations; t-tests) and rudimentary multivariate analyses (multiple linear regressions; exploratory factor analysis). To clarify what role the mental traits measured by SPOT and self-assessment questionnaire each play in the development of overall foreign language proficiency development, it will be necessary to employ more sophisticated statistical techniques, including confirmatory factor analysis based on a structural equation model. Moreover, it is desirable to confirm the outcomes of parametric statistical analyses with non-parametric analyses, given the fact that the ISLPR scales are (at best) ordinal scales, not equal interval scales.

So far, our discussions have been predominantly psychometric, centering around the predictive validity of SPOT and self-assessment against academic scores. On the other hand, it is equally important to view learners' self-appraisal from a cognitive angle (Wenden, 1986). Indeed the discrepancy between self-appraisal and actual performance has been reported in various domains of cognitive/behavioral sciences. (For example, Matsuura (1999) reviews research into drivers' tendency to over-estimate their own driving skills.) In particular, the tradition of meta-cognition and self-regulated learning research in cognitive psychology (Hacker, Dunlosky & Graesser, 1998; Elbaum, Berg, & Dodd, 1993) provides various research techniques to tackle this issue, which have so far been under-utilized by L2 self-assessment researchers. For instance, it would be intriguing to collect think-aloud protocols from learners assessing themselves against ISLPR, to see how they relate those descriptors to their own communicative experiences, and where rating fluctuations over time can stem from.

With all these pieces of evidence, some institutions which consider that course allocation policy is a manifestation of their "philosophy" or "beliefs" may still want to decide their placement procedure irrespective of its empirical validity. Indeed science cannot prescribe value judgement: it can at best help people to make a well-informed decision.

Nevertheless, it should be noted that an institution's language program management will depart from a rational policy-making process to approximate the practice of ideology or metaphysics, if results from empirical validations are totally disregarded.

Acknowledgements

I would like to express my gratitude to YAMAMOTO Hirofumi who introduced SPOT to UNSW, KOBAYASHI Noriko who helped us in the process of its implementation, HAYAKAWA Harumi who helped us process the data, and SASAKI Miyuki, IWASHITA Noriko and HATASA Yukiko who provided helpful comments on an earlier version of this manuscript. The Japanese Self-Assessment Questionnaire was used in the present study with the consent of THOMSON-KINOSHITA Chihiro.

References

Bachman, L.F. & Cohen, A.D. (1998). Language testing-SLA interface: An update. In Bachman, L.F. & Cohen, A.D. *Interface between SLA and Testing Research.* New York: Cambridge. 1-31.

Bachman, L. & Palmer, A. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing 6,* 1, 14-29.

Brindley, G. (1998). Describing language development? Rating scales and second language acquisition. In Bachman, L.F. & Cohen, A.D. *Interface between SLA and Testing Research.* New York: Cambridge. 112-140.

Blanche, P. & Merino, B.J. (1989). Self-assessment of foreign language skills. *Language Learning 39,* 3, 313-340.

Davies, A. (1995). Introduction: measures and reports. *Melbourne papers in Language Testing* 4, 2: 1-11.

Elbaum, B.E., Berg, C.A. & Dodd, D.H. (1993). Previous learning experience, strategy beliefs, and task definition in self-regulated foreign language learning. *Contemporary Educational Psychology, 18,* 318-336.

Ford-Niwa, J. (1997). An attempt to measure language proficiency: On the construct validity of SPOT (Simple Performance-Oriented Test). [In Japanese] Development of SPOT (Simple Performance-Oriented Test) for the purpose of Placing Japanese Language Students Report (2). 38-49.

Hacker, D.J., Dunlosky, J. & Graesser, A.C. (1998). *Metacognition in Educational Theory and Practice.* Mahwah: Lawrence Erlbaum.

Hatasa, Y. & Tohsaku, Yasu-Hiko. (1997). SPOT as a placement test. Development of SPOT (Simple Performance-Oriented Test) for the purpose of Placing Japanese Language Students Report (2). 38-49, A5-20.

Hirose, K. & Sasaki, M. (1994). Explanatory variables for Japanese students' expository writing in English: An exploratory study. *Journal of Second Language Writing, 3,* 203 - 229.

Ingram, D. E. (1995). Scales. *Melbourne papers in Language Testing 4,* 2: 12-29.

Kinoshita-Thomson, C. (1995). Japanese Self-Assessment Questionnaire. University of New South Wales.

Kinoshita-Thomson, C. (1997). Self-assessment as a Japanese language placement instrument: learner diversity and curriculum implications. Paper presented at the Japanese Studies Association conference, July 1997, Melbourne.

Kobayashi, N. (1997). Development of SPOT (Simple Performance-Oriented Test) for the purpose of placing Japanese Language Students. [In Japanese] *Proceedings of the fourth International Conference on Testing JFL ability: The SPOT project and related issues.* August 30, 1997. 1-7.

LeBlanc, R. & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly 19,* 3, 673-687.

Long, M. & Crookes, G. (1993). Unit of analysis in syllabus design: the case for task. In Crookes, G. & Gass, S. (Eds.) *Tasks in a pedagogical context: Integrating theory and practice.* Bristol: Multilingual Matters.

Matsuura, T. (1999). Drivers' overestimation of their own skill. [In Japanese] *Japanese Psychological Review 42,* 4, 419-437.

Peirce, B.N., Swain, M. & Hart, D. (1993). Self-assessment, French immersion, and locus of control. *Applied Linguistics, 14,* 1, 25-42.

Quinn, T. J. & McNamara, T. F. (1987). Australian Second Language Proficiency Ratings. In Alderson, S.C., Krahnke, K.J. & Stanfield, C.W. (Eds.) *Review of English Language Proficiency Test.* New York: TESOL.

Ross, S. (1998). Self-assessment in second language testing: a meta-analysis and analysis of experiential factors. *Language Testing 15,* 1, 1-20.

Spence-Brown, R. (1997). The real world and the language tester: considerations of authenticity and interactiveness in the design and assessment of language tests. *Proceedings of the fourth International Conference on Testing JFL ability: The SPOT project and related issues.* August 30, 1997. 73-86.

Wenden, A. (1986). What do second-language learner know about their language learning? : a second look at retrospective accounts. *Applied Linguistics 7,* 2, 186-205.

Wylie, E. & Ingram, D.E. (1993). Australian Second Language Proficiency Ratings (ASLPR): Self-assessment Version. Griffith University.