How Reliable and Valid is the Japanese Version of the Strategy Inventory for Language Learning (SILL)?

Gordon Robson
Showa Women's University
Hideko Midorikawa
Showa Women's University

This study looks at the internal reliability of the Strategy Inventory for Language Learning (Oxford, 1990), using the ESL/EFL version in Japanese translation. The results of the Cronbach's alpha analysis indicate a high degree of reliability for the overall questionnaire, but less so for the six subsections. Moreover, the test-retest correlations for the two administrations are extremely low with an average shared variance of 19.5 percent at the item level and 25.5 percent at the subsection level. In addition, the construct validity of the SILL was examined using exploratory factor analysis. While the SILL claims to be measuring six types of strategies, the two factor analyses include as many as 15 factors. Moreover, an attempt to fit the two administrations into a six-factor solution results in a disorganized scattering of the questionnaire items. Finally, interviews with participating students raised questions about the ability of participants to understand the metalanguage used in the questionnaire as well as the appropriateness of some items for a Japanese and EFL setting. The authors conclude that despite the popularity of the SILL, use and interpretation of its results are problematic.

本研究は、Oxford(1990)の外国語学習ストラテジー・インベントリー(SILL)のEFL/ESL用日本語版の内部信頼性及び構成概念妥当性を実験と統計によって検証したものである。クロンバック・アルファ検定による内部信頼性については、インベントリーの全項目は全体としては信頼性が高かったが、6タイプのサブカテゴリーに分類されたストラテジーについては信頼性が低かった。また、インベントリーを用いたテスト・再テストの相関は低く、全項目では平均寄与率19.5パーセント、サブカテゴリーでは25.5パーセントであった。構成概念妥当性検定のための説明的因子分析の結果は、6タイプのストラテジーが15因子に細分化されたこと、さらに、全項目を6因子に分けた結果、それぞれの因子が無秩序に分類される結果となった。最後に、インタビューによって、この実験に参加した被験者学生にインベントリーの各項目の内容理解について確認した結果、日本語がわかりにくく判断しいくい記述、日本のEFLの状況では理解しにくい記述があることが明らかになった。以上のすべてから、SILLの実用的評価にもかかわらず、それを用いること、また、そこから得た結果の解釈には問題が含まれているというのが、本研究の研究者が得た結論である。

The use of self-report instruments to investigate various aspects of individual learner differences is a common and accepted practice in the field of second language acquisition research. As a consequence, a large number of such instruments have been developed and used over the years. These include the Attitude/Motivation Test Battery (A/MTB) (Gardner and Lambert, 1972), the Foreign Language Classroom Anxiety Scale (FLCAS) (Horwitz, Horwitz, and Cope, 1986), and the instrument under discussion here, the Strategy Inventory for Language Learning (SILL) (Oxford, 1990). However, despite the wide acceptance and use of these instruments, issues such as their reliability and validity are often lost in the enthusiasm to find out what students really feel or believe. Although a given instrument may have been rigorously developed and even subjected to various measures of reliability and validity, when it is translated into another language or used in a cultural setting different from the one originally intended, it must once again be rigorously examined, as suggested by Griffee (1999).

This report will present the initial results of the researchers' attempts to provide reliability and validity data on the SILL in a Japanese university setting. This study is grounded in the researchers' numerous other attempts to validate other Japanese translations of measures of individual learner differences, such as motivation, anxiety, learning styles, learning beliefs, and learning strategies. Reliability is typically measured through statistics such as Cronbach's alpha or multiple administrations of a test with the same subjects, both of which are used here. Regarding validity, although in the past other methods of validating have been put forward, recently Chapelle (1994) and Messick (1989) have persuasively argued for validity to be condensed into a single, general approach where the focus is on the instrument as a construct. As the measures typically used in this type of research have been self-report questionnaires in which items were grouped into categories or subscales, researchers have favored factor analytic validation for the various groupings or categories assigned to the questionnaire items. The use of factor analysis to confirm a theorized grouping of items is a long-established practice (Guilford & Fruchter, 1973), especially in the field of personality research, where it has been used to validate self-report questionnaires for over 50 years (see for example Allport, 1937; Guilford, 1940; McCrae, 1989). Therefore, this will be the approach taken in validating the six groups of strategies making up the SILL.

What is Reliability and Validity?

There are various approaches to testing and confirming the reliability

and validity of a given research instrument. We can define the reliability as the proportion of the variation in test scores that is true variation and not error (Bachman, 1990; Brown, 1988). Typically, when measuring reliability, the items on the questionnaire are subjected to one or more types of statistical measurement. The most commonly employed statistic is Cronbach's alpha, which measures the internal consistency of a test. Another approach is to obtain simple correlations between test items of a measure that is given to the same population two or more times. This is referred to as test-retest reliability.

In general, the validation of a self-report instrument is much more difficult and in the past involved several different types of validity such as face validity, content validity, construct validity, factor analytic validity, and criterion-related validity. However, in an insightful article Messick (1989) points out that while it is important to validate the method of data collection, the more crucial area is to validate the inferences, interpretations, and actions taken based on the scores derived from the data. Moreover, Chapelle (1994) argues that "construct validity is central to all facets of validity inquiry," and as an ongoing process, there is no once and forever validity (p. 161).

From a statistical point of view, there are several ways to confirm the construct validity of an instrument. The use of correlation approaches and factor analysis has been noted previously. Typically these approaches involve using several tests or questionnaires that are believed to represent a construct, such as language-learning strategies, and then confirming the validity of the items through high correlations. If the correlations are high enough, then we can infer that they measure the hypothesized construct. Factor analysis can be used when measures for several different constructs are being used, such as for motivation, strategies, and personality. Subsequently, their loadings on distinct factors confirm that they measure separate aspects of learner behavior. A second use of factor analysis is to break a measure into subgroupings, such as the six hypothesized parts of the SILL, and then factor them to see if these divisions are valid. Evidence of a measure's validity can also be confirmed experimentally or quasi-experimentally through related outcomes using, for example, a measure of language learning strategies and scores on some measure of language learning such as the TOEFL Test. This would indicate not only that the measure was validly measuring strategies, but also that such strategies were useful.

The Strategy Inventory for Language Learning (SILL)

The Strategy Inventory for Language Learning (SILL) is a self-report questionnaire for determining the frequency of language learning strategy use. It consists of 50 items with five Likert-scale responses of never or almost never true of me, generally not true of me, somewhat true of me, generally true of me, always or almost always true of me. Based on a factor analysis of an earlier, larger version, Oxford organized the SILL into six strategy subscales: (a) Memory Strategies (9 items), (b) Cognitive Strategies (14 items), (c) Compensation Strategies (6 items), (d) Metacognitive Strategies (9 items), (e) Affective Strategies (6 items), and (f) Social Strategies (6 items). The questionnaire was translated into Japanese as part of the Japanese language version (Oxford, 1990/1994) of Oxford's (1990) Language Learning Strategies: What Every Teacher Should Know. Although Oxford does not directly discuss the process for establishing the reliability and validity of the SILL, a note to Chapter Six explains that an earlier, 121-item version of the SILL was found to have a reliability of .96 based on a 1,200-person sample and .95 with a 483-person sample. She then goes on to state that the reliability of 9 of 10 factors was found to be moderate to high with figures of .60 to .86, although for the 10th factor it was only .31. This is not the typical way of reporting the results of factor analysis, and if the 121-item version was claiming to measure six strategy types, then a 10-factor solution is hardly confirmation. The note goes on to state that the fifty-item version 7.0 of the SILL under discussion here was still being assessed for reliability and validity. Thus, while it would seem that the various versions of the SILL have a proven level of reliability, this does not suggest that the questionnaire is valid. As Bachman (1990) has stated, he primary concern in test development and use is demonstrating not only that test scores are reliable, but that the interpretations and uses we make of test scores are valid (p. 237). If at this point in the SILL's construction it were found to be unreliable, there would be no need to proceed, as an unreliable measure is similarly not valid.

Oxford (1996) discusses the psychometric qualities of the SILL, and in terms of reliability, she cites Watanabe (1990), where a Japanese version of the SILL achieved a Cronbach's alpha reliability of .92, and other studies with similar reliabilities in the .90 range. Following the above-mentioned Messick (1989) and Chapelle (1994) approach to test validity, Oxford examined a number of studies where the SILL correlated significantly with various measures of language learning. In Oxford, Park-Oh, Ito, and Sumrall (1993), a multiple-regression analysis found low but significant predictive relationships between strategies and final

test grades (.20). Takeuchi (1993), also using multiple-regression analysis with language achievement as measured by the Comprehensive English Language Test (CELT), found that four SILL items (17, 21, 22 and 32) positively predicted language achievement while four items (6, 30, 43 and 49) negatively predicted language success. Finally, Watanabe (1990) found low correlations between SILL items and students' self-ratings of their own proficiency. Although these results provide some measure of validation, only a few SILL items are involved, and the correlations are extremely low.

In Brown, Robson, and Rosenkjar (1996) an independently translated version of the SILL was used in a multiple, individual learner differences study. The overall reliability of that translated version was .94 with the reliability for the six strategy types being .74 for Memory Strategies, .84 for Cognitive Strategies, .69 for Compensation Strategies, .88 for Metacognitive Strategies, .63 for Affective Strategies, and .73 for Social Strategies. The factor analysis in the Brown et al. (1996) study was only used to determine if the SILL was measuring something distinct from the other measures of such variables as personality, anxiety, and motivation. The six strategy types were found to load on a single factor, which confirmed that the SILL was measuring a variable distinct from the other instruments. The researchers know of no other published study that has attempted to establish either reliability or validity in this manner using a Japanese version of the SILL.

However, at TESOL 2000 in Vancouver, Canada, Hsiao and Oxford (2000) presented the results of a multi-group confirmatory factor analysis for an 80-item SILL. The factor analysis placed only 17 items into the six hypothesized groupings, leaving 63 items with no relation to the six strategy categories hypothesized. The 17 items were Memory Strategies (4, 5, 8), Cognitive Strategies (26, 27, 28), Compensation Strategies (41, 43), Metacognitive Strategies (49, 53, 55), Affective Strategies (66, 68, 69), and Social Strategies (72, 73, 74). Of these, only items 5, 27, 28, 68, and 72 are the same as or similar to items on the 50-question version of the SILL under study here.

To summarize, the SILL appears to enjoy a high degree of reliability in its various versions and the languages in which it has been employed. However, the reliability has been for the SILL as a whole, with the exception of Brown et al. (1996), where several of the scales were rather low. This still leaves the question of validity, which based on the sources discussed seems far from established, and has led us to ask the following research questions.

Research Questions

1. How reliable is the Japanese language version of the SILL for Japanese university students?
2. To what degree is the Japanese language version of the SILL valid for Japanese university students?

Method

The present study is based on two administrations of the officially translated SILL (Oxford 1990/1994) to the same group of 153 Japanese university students. The group was comprised of 110 first- and second-year females and 43 first- and second-year males studying at a private women's university and a private coeducational university in Tokyo. Their English proficiency level was approximately low intermediate. The first administration was conducted at the beginning of the spring semester. A second administration was conducted during the beginning of the fall semester using a version in which the order of the items had been randomized. There were no changes in the makeup of the group of subjects for the two administrations of the questionnaire. In addition, post-administration interviews were conducted with ten randomly selected students, four males and six females to get feedback on what the students thought about the questionnaire. The interviews were conducted individually in Japanese by the Japanese nativespeaker author of this study with each of the interviewees. They were questioned about their thoughts on each of the 50 items and their responses were taken down in the form of notes.

Analysis

The data collected from the two administrations of the SILL were first analyzed for item statistics followed by descriptive statistics for the six parts as well as the entire SILL. The alpha level for all statistical decisions was set at .05. Both administrations were then examined for internal consistency using Cronbach's alpha for each of the six parts as well as overall reliability for both administrations. Next, each Time One item was compared to its identical Time Two item using the Pearson correlation. The resulting correlations were then squared to determine the degree of shared variance. The squared value of the correlation coefficient can be interpreted as the proportion of similarity between the two items (Hatch & Lazaraton, 1991). This procedure was repeated for the six parts of the SILL and for the entire SILL as well. Finally, the two administrations were

examined using principal component analysis (PCA), which is a type of exploratory factor analysis, with varimax rotation and eigenvalues set at one. These are the typical procedures for carrying out factor analysis. As is common, loadings of .30 and above were considered strong enough for inclusion in a given factor (Hatch & Lazaraton, 1991). In the initial use of PCA, the analysis was allowed to select as many factors as could be found with an eigenvalue over 1.00; however, a second PCA was run on both administrations in which the analysis was forced to choose six factors based on Oxford's theorized grouping. Scree plots for all PCAs were also calculated. These additional procedures were conducted to provide the SILL with as many opportunities as possible to supply support for its theoretical basis. Finally, the notes taken during the interviews were examined to determine the types of difficulties the students had understanding the questionnaire items and how their difficulties compared to one another.

Results

Table 1 shows the items themselves with their groupings, the mean on each item and the standard deviation, with Table 2 showing the means and standard deviations for the items on the second administration. Table 3 provides the descriptive statistics for the six subsections of the SILL and the entire SILL for both administrations. The distributions are all either positively or negatively skewed and those with skewness statistics at 1.0 or greater are problematic (Brown, 1997). These skewed distributions can reduce the test reliability and are violations of the assumptions of normality for the correlation statistics and factor analysis, which could adversely affect these results.

Table 1: Mean Scores and Standard Deviations for the Items and Their Strategy Types, Time One (n = 153)

Item	Statement	Type	M	SD
1	I think of relationships between what I already know and new things I learn in English.	Memo	2.79	0.94
2	I use new English words in a sentence so I can remember them.	Memo	2.56	0.95
3	I connect the sound of a new English word and an image or picture of the word to help me remember.	Memo	3.02	1.09
4	I remember a new English word and an image or picture of a situation in which the word might be used.	Memo	2.63	1.12
5	I use rhymes to remember new English words.	Memo	2.41	1.11
6	I use flash cards to remember new English words.	Memo	2.19	1.42

7	I physically act out new English words.	Memo	1.80	0.88
8	I review English lessons often.	Memo	2.66	0.66
9	I remember new English words or phrases by remembering their location on the page, on the board, or on a street sign.	Memo	2.56	1.24
10	I say or write new English words several times.	Cog	3.98	0.99
11	I try to talk like native English speakers.	Cog	3.09	1.23
12	I practice the sounds of English.	Cog	3.40	1.05
13	I use the English words I know in different ways.	Cog	2.89	0.96
14	I start conversations in English.	Cog	2.17	0.86
15	I watch English language TV shows spoken in English or go to movies spoken in English.	Cog	3.25	1.09
16	I read for pleasure in English.	Cog	2.77	0.97
17	I write notes, messages, letters or reports in English.	Cog	2.19	1.06
18	I first skim an English passage (read over the passage quickly) then go back and read carefully.	Cog	3.39	1.09
19	I look for words in my own language that are similar to new words in English.	Cog	2.39	1.16
20	I try to find patterns in English.	Cog	2.81	1.07
21	I find the meaning of an English word by dividing it into parts that I understand.	Cog	2.70	1.18
22	I try not to translate word-for-word.	Cog	2.96	0.99
23	I make summaries of information that I hear or read in English.	Cog	1.97	0.92
24	To understand unfamiliar English words, I make guesses.	Comp	3.44	0.92
25	When I can't think of a word during a conversation in English, I use gestures.	Comp	3.65	1.16
26	I make up new words if I do not know the right ones in English.	Comp	2.23	1.11
27	I read English without looking up every new word.	Comp	3.07	1.05
28	I try to guess what the other person will say next in English.	Comp	2.35	0.99
29	If I can't think of an English word, I use a word or phrase that means the same thing.	Comp	3.81	0.94
30	I try to find as many ways as I can to use my English.	Meta	2.60	1.01
31	I notice my English mistakes and use that information to help me do better.	Meta	3.37	1.01
32	I pay attention when someone is speaking English.	Meta	3.60	0.98
33	I try to find out how to be a better learner of English.	Meta	2.73	1.07
34	I plan my schedule so I will have enough time to study English.	Meta	2.31	0.89
35	I look for people I can talk to in English.	Meta	2.19	1.03
36	I look for opportunities to read as much as possible in English.	Meta	2.50	0.97
37	I have clear goals for improving my English skills.	Meta	2.94	1.29
38	I think about my progress in learning English.	Meta	3.09	1.04
39	I try to relax whenever I feel afraid of using English.	Aff	2.80	1.07
40	I encourage myself to speak English even when I am afraid of making a mistake.	Aff	3.07	1.16
41	I give myself a reward or treat when I do well			

	in English.	Aff	3.43	1.09
42	I notice if I am tense or nervous when I am studying or using English.	Aff	3.08	1.16
43	I write down my feelings in a language learning diary.	Aff	1.48	0.86
44	I talk to someone else about how I feel when I am learning English.	Aff	1.99	0.99
45	If I do not understand something in English, I ask the other person to slow down or say it again.	Soc	4.14	0.88
46	I ask English speakers to correct me when I talk.	Soc	2.65	1.19
47	I practice English with other students.	Soc	2.24	1.01
48	I ask for help from English speakers.	Soc	2.69	1.24
49	I ask questions in English.	Soc	2.44	1.09
50	I try to learn about the culture of English speakers.	Soc	3.03	1.21

Note: The statement for each item is in the English original from which the Japanese translation was made.
Key for Strategy Type: Memo = Memory, Cog = Cognitive, Comp = Compensation, Meta = Metacognitive, Aff = Affective, Soc = Social

Table 2: Mean Scores and Standard Deviations for the Items, Time Two (n = 153)

Item	M	SD	Item	M	SD	Item	M	SD
1	3.95	0.89	2.99	1.14	1.06	16	3.39	1.04
2	2.51	1.04	3.25	1.13	3.15	17	2.08	1.22
3	3.24	1.17	3.52	0.90	1.07	18	2.88	1.04
4	2.22	1.06	2.36	0.82	2.97	19	3.47	1.22
31			46			32		
33			48			34		
47								

Table 3: Descriptive Statistics for the SILL and Subsections, Times One and Two (n = 153)

Measure	M	SD	Min	Max	Range	Skew
SILL, Time One	139.00	24.60	66	207	141	-.24
Memo, Time One	22.64	4.66	11	36	25	-.04
Cog, Time One	39.95	7.65	18	61	43	-.20
Comp, Time One	18.33	3.60	8	28	20	-.04
Meta, Time One	25.03	6.34	9	40	31	-.16
Aff, Time One	15.84	3.64	7	27	20	.07

Soc, Time One	17.18	4.89	6	30	24	.11
SILL, Time Two	144.58	25.22	63	229	166	-.26
Memo, Time Two	26.67	4.94	11	41	30	-.29
Cog, Time Two	37.84	7.91	16	66	50	.13
Comp, Time Two	18.85	3.41	8	27	19	-.42
Meta, Time Two	26.21	5.71	9	44	35	-.16
Aff, Time Two	17.45	3.73	6	27	21	-.38
Soc, Time Two	17.54	3.65	6	26	20	-.25

Key for Strategy Type: Memo = Memory, Cog = Cognitive, Comp = Compensation, Meta = Metacognitive, Aff = Affective, Soc = Social

Table 4 gives the reliability for the six parts and the overall reliability for both administrations. While the SILL as a whole for both times one and two has very high reliability at .93, several of the subsections are very low. In particular, the Time One reliabilities for Memo, Comp, and Aff are unacceptably low. The same is true for Memo, Comp, Aff, and Soc in Time Two. The results for the second measure of reliability, test-retest, are shown in Tables 5 and 6. The degree of shared variance for the items does not exceed 46 percent with some as low as 3, 4, 5, and 7 percent. The average for all the items is just 19.5 percent. For the subsections, the shared variance is similarly low, with the only exception being for the SILL as a whole at 58 percent.

Table 4: Internal Consistency for the SILL and Subsections, Times One and Two (n = 153)

Measure	Alpha	Measure	Alpha
SILL, Time One	.93	SILL, Time Two	.93
Memo, Time One	.63	Memo, Time Two	.66
Cog, Time One	.80	Cog, Time Two	.83
Comp, Time One	.67	Comp, Time Two	.58
Meta, Time One	.85	Meta, Time Two	.79

Table 5: Percentage of Shared Variance Between SILL Items, Times One & Two (n = 153)

Items	R Squared	Items	R Squared
1	.03	26	.28
2	.16	27	.10
3	.16	28	.10
4	.21	29	.14
5	.18	30	.24
6	.07	31	.18
7	.18	32	.18
8	.07	33	.24
9	.13	34	.18
10	.18	35	.41
11	.34	36	.25
12	.26	37	.42
13	.12	38	.16
14	.29	39	.24
15	.29	40	.22
16	.29	41	.34
17	.27	42	.08
18	.14	43	.21
19	.14	44	.05
20	.14	45	.14
21	.27	46	.46
22	.07	47	.24
23	.11	48	.28
24	.07	49	.07
25	.36	50	.04

Table 6: Percentage of Shared Variance Between the SILL & Subsections, Times One & Two (n = 153)

Measure	R Squared
Memo	.25
Cog	.36
Comp	.14
Meta	.35
Aff	.17
Soc	.26
SILL	.58

Key for Strategy Type: Memo = Memory, Cog = Cognitive, Comp = Compensation, Meta = Metacognitive, Aff = Affective, Soc = Social

Tables 7 and 8 show the results of the first PCAs with a 15-factor solution for Time One and a 13-factor solution for Time Two. We would expect the factor analysis to group items 1 through 9 in one factor, items 10 through 23 in a second factor, items 24 through 29 in a third factor, items 30 through 38 in a fourth factor, items 39 through 44 in a fifth factor, and items 45 through 50 in a sixth. However, the results for the Time One PCA show very few items loading together. The greatest group of items loading together is in factor 14 with items 46 through 49 together; however, beyond this, there are no greater groups of loadings than just two or three items together. Factor one takes up 23 percent of the total variance with the other factors accounting for considerably less, which is confirmed by the eigenvalues. In addition, the communalities, which show the degree to which the factors are accounting for each item, are not particularly high except for items 24, 25, 35 and 36. A similar state of affairs is found for the Time Two PCA; however, there are no groups of loadings greater than two, making the results appear even less systematic than with those of the Time One analysis. Again, almost all the total variance is being accounted for by factor one. Also, with the exceptions of items 13 and 14, the communalities are not particularly high.

Tables 9 and 10 show the attempt to force the SILL into a six-factor

Table 7: Principal Component Analysis, Time One (n = 153)

Item	Factor Loadings					
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Communalities
35	0.85					0.91
36	0.85					0.91
23	0.64					0.47
30	0.60					0.58
50	0.44					0.61
32		0.72				0.57
29		0.66				0.51
41		0.59				0.53
31		0.56				0.62
40		0.53				0.58
39			0.45			0.50
20			0.70			0.57
9			0.63			0.35
19			0.58			0.44
21			0.39			0.45
8				0.79		0.49
34				0.49		0.54
16				0.45		0.57

Table 7 cont...

Item	Factor Loadings					
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Communalities
44					0.79	0.38
1					0.32	0.38
Eigenvalues	11.54	3.19	2.40	2.02	1.98	
Percent of Total Variance	23.08	6.37	4.81	4.04	3.95	

Item	Factor Loadings					
	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10	Communalities
24	0.91					0.91
25	0.91					0.91
27	0.43					0.59
26		0.79				0.39
33			0.63			0.53
38			0.56			0.55
37			0.52			0.51
12			0.38			0.53
10				0.78		0.38
45				0.41		0.49
6					0.79	0.43
42					0.53	0.35
Eigenvalues	1.69	1.60	1.45	1.33	1.22	
Percent of Total Variance	3.39	3.20	2.91	2.66	2.45	

Item	Factor Loadings					
	Factor 11	Factor 12	Factor 13	Factor 14	Factor 15	Communalities
18	0.68					0.47
22	0.65					0.50
13	0.37					0.45
3		0.74				0.46
4		0.58				0.51
7		0.46				0.41
5		0.46				0.44
17			0.73			0.49
14			0.59			0.65
15			0.58			0.42
11			0.48			0.59
48				0.70		0.57
46				0.64		0.65
47				0.52		0.63
43				0.52		0.46
49				0.51		0.61

Table 7 cont...

28				0.74	0.42
2				0.47	0.36
Eigen values	1.19	1.18	1.12	1.08	1.04
Percent of Total Variance	2.38	2.36	2.24	2.16	2.09

Note: Only items with loadings equal to or over 0.30 are indicated in the table

Table 8: Principal Component Analysis, Time Two (n = 153)

Item	Factor Loadings					
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Communalities
17	0.72					0.66
35	0.71					0.76
26	0.65					0.59
41	0.61					0.56
30	0.58					0.52
19	0.53					0.55
22	0.46					0.48
12	0.43					0.41
8	0.39					0.51
21	0.38					0.54
9		0.69				0.58
38		0.64				0.48
44		0.62				0.49
40		0.56				0.54
37		0.55				0.56
10		0.54				0.51
32		0.49				0.55
11		0.48				0.55
42			0.66			0.61
28			0.65			0.53
16			0.63			0.62
45			0.63			0.47
29			0.54			0.51
5			0.38			0.42
15				0.67		0.42
47				0.59		0.53
31				0.51		0.55
1				0.51		0.53
3				0.46		0.49
4				0.38		0.37
13					0.85	0.91
14					0.85	0.91
Eigenvalues	11.94	3.02	2.70	2.09	1.89	
Percent of Total Variance	23.88	6.04	5.41	4.18	3.79	

Table 8 cont...

Item	Factor Loadings					
	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10	Communalities
2	0.76					0.42
36	0.45					0.53
46		0.85				0.48
18			0.73			0.51
20			0.52			0.66
24			0.49			0.47
25			0.42			0.46
43				0.77		0.46
50				0.49		0.41
49				0.39		0.51
6					0.62	0.38
7					0.43	0.42
Eigenvalues						
	1.58	1.46	1.39	1.36	1.20	
Percent of Total Variance						
	3.16	2.91	2.79	2.73	2.40	

Item	Factor Loadings			
	Factor 11	Factor 12	Factor 13	Communalities
48	0.72			0.37
39	0.43			0.64
34	0.37			0.49
23		0.75		0.53
33		0.40		0.70
27			0.72	0.31
Eigenvalues				
	1.16	1.10	1.05	
Percent of Explained Variance				
	2.31	2.20	2.09	

Note: Only items with loadings equal to or over 0.30 are indicated in the table.

solution. Here we have a clearer picture of why the first factor is taking up so much of the total variance, although with Time Two, there is less of a concentration of items in the first factor. Nonetheless, the loadings for both PCAs show a combination of related and unrelated items from the six subgroups loading together. Figures 1 and 2 give visual representations of the eigenvalues through scree plots, which, if we count the number of factors to the left of the point where the line turns strongly to the right, seem to indicate that a one factor analysis of the SILL would be most appropriate.

The interviews revealed some very interesting problems the question-

Table 9: Principal Component Analysis with Six Forced Factors, Time One (n = 153)

Item	Factor Loadings				
	Factor 1	Factor 2	Factor 3	Factor 4	Communalities
35	0.79				0.91
36	0.79				0.91
14	0.69				0.65
47	0.68				0.64
49	0.68				0.61
30	0.67				0.59
46	0.67				0.65
48	0.60				0.57
23	0.55				0.47
17	0.51				0.49
50	0.49				0.62
40	0.47				0.58
28	0.44				0.42
16	0.43				0.57
4	0.42				0.51
15	0.39				0.42
13	0.35				0.45
26	0.31				0.39
20		0.72			0.57
21		0.66			0.45
19		0.59			0.44
22		0.49			0.50
18		0.44			0.47
39		0.43			0.50
9		0.43			0.35
7		0.40			0.42
3		0.39			0.46
5		0.35			0.44
32			0.69		0.57
11			0.62		0.59
45			0.57		0.49
12			0.56		0.53
29			0.55		0.51
38			0.51		0.55
33			0.51		0.53
31			0.49		0.62
37			0.44		0.51
10			0.39		0.38
8				0.69	0.49
34				0.49	0.53
2				0.48	0.36
6				0.44	0.43
Eigenvalues					
	11.54	3.19	2.40	2.02	
Percent of Total Variance					
	23.09	6.37	4.81	4.04	

Table 9 cont...

Item	Factor Loadings		
	Factor 5	Factor 6	Communalities
43	0.53		0.46
44	0.53		0.38
1	0.43		0.38
41	0.37		0.53
24		0.86	0.91
25		0.86	0.91
27		0.44	0.50
Eigenvalues	1.98	1.69	
Percent of Total Variance	3.96	3.39	

Note: Only items with loadings equal to or over 0.30 are indicated in the table.

Table 10: Principal Component Analysis with Six Forced Factors, Time Two (n = 153)

Item	Factor Loadings				
	Factor 1	Factor 2	Factor 3	Factor 4	Communalities
38	0.63				0.48
44	0.58				0.49
32	0.57				0.55
39	0.57				0.64
37	0.57				0.56
31	0.56				0.55
24	0.53				0.47
40	0.52				0.54
4	0.47				0.37
25	0.47				0.46
15	0.40				0.42
35		0.68			0.76
22		0.66			0.48
26		0.65			0.59
17		0.59			0.66
13		0.57			0.91
14		0.57			0.91
6		0.54			0.38
7		0.51			0.42
41		0.51			0.56
30		0.51			0.52
33		0.44			0.70
23		0.39			0.53
21		0.39			0.54
10		0.38			0.51
16			0.72		0.62
42			0.65		0.61

Table 10 cont...

20			0.65		0.66
28			0.64		0.53
29			0.61		0.51
45			0.49		0.47
18			0.47		0.51
5			0.47		0.42
3			0.37		0.49
34			0.37		0.49
47				0.59	0.53
49				0.57	0.51
36				0.57	0.53
1				0.52	0.53
48				0.49	0.37
50				0.48	0.41
8				0.44	0.51
12				0.32	0.41
Eigenvalues	11.94	3.02	2.70	2.09	
Percent of Total Variance	23.88	6.04	5.41	4.18	

Item	Factor Loadings		Communalities
	Factor 5	Factor 6	
9	0.62		0.58
11	0.58		0.55
19	0.56		0.55
43	0.39		0.46
2		0.69	0.42
27		-0.52	0.31
46		-0.42	0.48
Eigenvalues	1.89	1.58	
Percent of Total Variance	3.79	3.16	

Note: Only items with loadings equal to or over 0.30 are indicated in the table.

naire posed for the respondents. The majority of students interviewed had difficulty understanding items 1, 5, 6, 7, 14, 19, 20, 22, 26, 43, 44, and 47. The most commonly cited reason for their lack of understanding was unfamiliar Japanese or English expressions. This was particularly true for items 5, 6, 22, and 43. Another reason respondents gave for their difficulty in understanding was that they could not imagine the situation.

Discussion

The results reported above provide a high level of reliability for the SILL

Figure 1: Scree Plot, Principal Component Analysis,
Time One (n = 153)

Figure 2: Scree Plot, Principal Component Analysis,
Time Two (n = 153)

as a whole, which is problematic, as the SILL should be measuring six different types of strategies, not one grand strategy type. Moreover, the alphas for the subsections show a similarly low level of reliability as was found by Brown et al. (1996). One reason for this could be the number of items in the subsections, where the longer subsections such as Cognitive Strategies have higher reliability. Length is an important factor in reliability, as longer measures tend to be more reliable (Bachman, 1990). Moreover, as was noted previously, all the distributions are skewed, which must also be affecting the level of reliability. For example, Social Strategies Time One has fairly high reliability, but at Time Two it drops to .59. However, there is also an increase in the skew between times one (.11) and two (-.25). Nevertheless, these skewed distributions cannot fully explain the relatively low reliability as the Cognitive Strategies subsection has a consistent level of reliability from Time One to Time Two, but skewed distributions of -.20 and .13. The test-retest reliability as indicated by the percentage of shared variance for the items, subsections and entire measure show that the SILL is highly unreliable. It is important to remember that reliability can be measured several different ways, and that dependence on a single approach can be risky. One reason for the low figures has been found in other studies looking at either beliefs or strategies (for example Gaies & Sakui, 1999), where the students were found to change over time. Although it is difficult to determine the exact reasons for change without conducting extensive post-administration interviews, students may interpret the questions on a given measure in light of their current learning situation and not learning situations in general. Moreover, the effects of training and learning must also be taken into account. In addition, it is important to remember that strategies are not personality traits, which have been shown to remain stable over time and across situations (see Angleitner, 1991). Thus, it is hardly surprising that the percentage of shared variance should be so low between the two administrations. However, there are other possible explanations for the low levels of reliability. Again, the skewed distributions could be adversely affecting the results, or it is possible that the population surveyed was too homogeneous. The subjects are all from a single language background and culture with close similarities in age and possibly educational experience. The skewed distributions are likely part of the explanation. Nonetheless, the Japanese version of the SILL was designed to examine just this type of population. Moreover, the educational background of this group of subjects is probably not all that homogeneous. The students at the women's university come from a wide area north and west of Tokyo and attended both private and public high schools where there are educational differences from one

school to the next. The co-ed school subjects are similar in this regard. It also seems reasonable to expect that most administrators or teachers will use the SILL under similar conditions in Japan.

The factor analysis results do not confirm Oxford's six strategy categories even when attempting to force the analysis into a six-factor solution. In fact, the SILL is either measuring 15 or 13 different types of strategies, or even just one as indicated by the eigenvalues and scree plots. There are a number of potential reasons for this. The low reliability is an important factor as is the size of the population. Hatch and Lazaraton (1991) recommend at least 35 subjects per variable for PCA, which in this study would necessitate an n size of about 1,750 subjects. With a sample size of only 153, there is considerable loss of statistical power. Nonetheless, other studies with larger samples have shown similar results (Hsiao & Oxford, 2000) and based on those found here as well as in Brown et al. (1996), it would seem safer to limit the SILL to one grand language learning strategies factor instead of trying to break it into theorized groups.

Attempting to label each of these strategy types is very difficult. There seems to be almost no system to the factor loadings, although, some of the factors can be tentatively labeled. For example, Time One factor 14 seems to be related to Oxford's Social Strategies, while factor 2 contains items from the Analyzing and Reasoning subgroup within Cognitive Strategies. Factor 12 seems to be the Memory Strategy subgroup Applying Images and Sounds. The factor solution for the second time shows an even greater mixing of items from different strategy groups almost necessitating a complete abandonment of Oxford's categories. However, by looking at the wording, we can apply tentative labels. For example, factor two can be interpreted as various speaking strategies. In addition, there seem to be groupings of items in both Time One and Time Two based on the type of action expressed by the verb in Japanese. An example would be Time One factor four, where the subjects seem to place emphasis on such actions as "review," "read," and "plan." The attempt to force the SILL into six factors for Time One resulted in what looks like a one-factor solution including some items from each subsection. If these results had been repeated in Time Two, there would have been an opportunity to support a one-factor solution based on this data. Unfortunately, the items in the first factor differ.

The problems students had understanding the questionnaire were partially revealed by the post-administration interviews conducted with a very small sample. These students were unfamiliar with such expressions or situations as kokoro ni egaku (making a mental picture) in item 4, in o tsukau (use rhymes) in item 5, "flash cards" in item 6, and karada

de hy n shite (physically act out) in item 7. For example, during the administration of the questionnaire, the majority of students could not read the character in, which means rhyme in Japanese, and did not know its meaning when it was read to them. Moreover, other items that were incomprehensible were ones that reflect a more Western approach to learning strategies than one with which Japanese students are familiar, such as with items 4 and 7. In addition, students had difficulty relating many of the situations presented in the questionnaire to their own learning. First, the questionnaire presumes an ESL learning situation, where the situation in Japan is clearly EFL. Thus, these students have few opportunities for target language use outside of the classroom. Of the 10 interviewees, 3 had experienced studying in an English-speaking country and had few comprehension problems with the learning situations presented. However, for the remaining 7 students, target language study and use was limited to the classroom, library, home, train, or their English Speaking Society (ESS) meetings and they found many of the learning situations in the questionnaire unimaginable or strange. Moreover, as was noted above, the interviewees responded to items based on their current learning situation and not learning situations in general.

Conclusion

The simplest conclusion one can draw from this initial attempt at determining the reliability and validity of the Japanese language version of the SILL is that it is neither reliable nor valid based on this student sample. Although the SILL has shown a high degree of internal reliability for the entire questionnaire, it claims to measure six different strategies and thus must be analyzed as six different measures. In fact, the high degree of reliability for the entire SILL, as noted above, is not necessarily a good thing. The subsections have a generally low and unacceptable alpha level. Moreover, there are serious questions about how reliable the results are when given to the same group more than once and how valid the categories used to group the items on the questionnaire are. In other words, while the SILL may indeed be measuring language-learning strategies, it does not seem to be measuring groups of strategies in the manner Oxford has claimed, at least for these learners. It would seem reasonable, based on the high reliability for the entire SILL, the eigenvalues, and scree plots, to describe the SILL as a general measure of language-learning strategies and not a measure of six different strategy types. The researchers believe that these conclusions can be drawn based on these data in spite of potential problems with n size,

the possibly homogeneous population, skewed distribution, and low reliability.

As discussed previously, the methods Oxford used to validate an earlier version of the SILL are somewhat suspect. Taken together with the lack of established reliability or validity for the later versions (despite claims by Yamato, 2000, p. 142, to the contrary), those using the English version of the SILL will not be able to rely on the results. Moreover, Hsiao and Oxford's (2000) confirmatory factor analysis does not provide much confidence either. Cautionary use becomes even more necessary with the Japanese translation, as it is now a new questionnaire that has not gone through a rigorous reliability and validation process. These issues and problems are not just about strategies, but relate to any use of a self-report questionnaire. It should be clear from this analysis that simply taking a questionnaire, translating it into another language, administering it to a group of students and then using the results for making educational policy decisions are very unwise practices. Moreover, any questionnaire must reflect the actual learning situations of the target population, their strategy use, the type of language with which they are familiar, and any cultural differences that might affect the outcome. The researchers hope that through this initial attempt at validating Oxford's questionnaire other researchers and language-teaching professionals will take a more cautious approach to questionnaire use and interpretation.

Acknowledgements

This is a much revised and expanded version of a paper delivered at AILA 99 in Tokyo. The authors would like to thank the two anonymous reviewers for their suggestions and valuable comments.

Gordon Robson is Professor at Showa Women's University. His research interests include reading strategies and individual learner differences.

Hideko Midorikawa is Professor at Showa Women's University. Her research interests include reading strategies and teacher education.

References

Allport, G. W. (1937). Personality: A psychological interpretation. New York: Holt, Rinehard & Winston.

Angleitner, A. (1991). Personality psychology: Trends and developments. European Journal of Personality, 5, 156-171.

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

Brown, J. D. (1988). Understanding research in second language teaching: A teacher's guide to statistics and research design. London: Cambridge University.

Brown, J. D. (1997). Statistics corner, questions and answers about statistics: Skewness and kurtosis. Shiken: JALT testing & evaluation SIG newsletter, 1 (1), 16-18.

Brown, J. D., Robson, G., & Rosenkjar, P. (1996). The motivation, anxiety, learning styles and proficiency of Japanese college students. University of Hawai'i Working Papers in ESL, 15 (1), 33-72.

Chapelle, C. (1994). Are C-tests valid measures for L2 strategy research? Second Language Research, 10 (2), 157-187.

Gaies, S., & Sakui, K. (1999). Symposium on learners' beliefs about language learning. Paper presented at the 12th World Congress of Applied Linguistics, Tokyo, Japan.

Gardner, R. C., & Lambert, W. E. (1972). Attitudes and motivation in second-language learning. Rowley, MA: Newbury House.

Griffee, D. (1999). Translating questionnaires from English into Japanese: Is it valid? In A. Barfield, R. Betts, J. Cunningham, N. Dunn, H. Katsura, K. Kobayashi, N. Padden, N. Parry & M. Watanabe (Eds.), On JALT 98: Focus on the classroom (pp.176-180). Tokyo: Japan Association for Language Teaching.

Guilford, J. P. (1940). An inventory of factors S T D C R, manual of directions and norms. Beverly Hills: Sheridan Supply Co.

Guilford, J. P. & Fruchter, B. (1973). Fundamental statistics in psychology and education. New York: McGraw Hill.

Hatch, E. & Lazaraton, A. (1991). The research manual: Design and statistics for applied linguistics. New York: Newbury House Publishers.

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. The Modern Language Journal, 70, 125-132.

Hsiao, T.-Y., & Oxford, R. L. (2000). Learning strategy use and target languages: confirmatory factor analysis across languages. Paper presented at the annual meeting of the Teachers of English to Speakers of Other Languages, Vancouver, Canada.

McCrae, R. R. (1989). Why I advocate the five-factor model: Joint analyses of the NEO-PI and other instruments. In D. M. Buss & N. Cantor (Eds.), Personality psychology: Recent trends and emerging directions (pp. 237-245). New York: Springer-Verlag.

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

Oxford, R. L. (1990). Language Learning Strategies: What Every Teacher Should Know. New York: Newbury House.

Oxford, R. L. (1994). Gengo gakushu sutoratejii gaikokugo kyoushi ga shitte okankereba naranai koto [Language Learning Strategies: What Every Teacher Should Know] (M. Shishido & N. Ban, Trans.). Tokyo: Bonjinsha. (original work published 1990)

Oxford, R. L. (1996). Employing a questionnaire to assess the use of language learning strategies. Applied Language Learning, 7, (1 & 2), 25-45.

Oxford, R. L., Park-Oh, Y., Ito, S., & Sumrall, M. (1993). Factors affecting achievement in a satellite-delivered Japanese language program. American Journal of Distance Education, 7, 10-25.

Takeuchi, O. (1993). A study of language learning strategies and their relation to achievement in EFL listening comprehension. Bulletin of the Institute for Interdisciplinary Studies of Culture, 10, 131-141.

Watanabe, Y. (1990). External variables affecting language learning strategies of Japanese EFL learners: Effects of entrance examination, years spent at college/university, and staying overseas. Unpublished master's thesis, Lancaster University, Lancaster, U.K.

Yamato, R. (2000). Awareness and real use of reading strategies. JALT Journal, 22 (1), 140-164.

(Received August 5, 2000; revised May 3, 2001)