<table>
<thead>
<tr>
<th>Title</th>
<th>A Confirmatory Approach to the Structural Validity of Scores Generated by the Strategy Inventory for…</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Isemonger, Ian</td>
</tr>
<tr>
<td>Citation</td>
<td>International Journal of Social and Cultural Studies, 9: 1-14</td>
</tr>
<tr>
<td>Issue date</td>
<td>2016-03</td>
</tr>
<tr>
<td>Type</td>
<td>Departmental Bulletin Paper</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/2298/34461">http://hdl.handle.net/2298/34461</a></td>
</tr>
</tbody>
</table>

熊本大学学術リポジトリ

Kumamoto University
A Confirmatory Approach to the Structural Validity of Scores Generated by the Strategy Inventory for Language Learning (SILL)

Ian Isemonger
Kumamoto University, Japan

Abstract
The hypothesized factor structure for a Japanese-language version of the Strategy Inventory for Language Learning (SILL) authored by Oxford (1990) was examined in this study using Cronbach’s alpha, and most importantly, confirmatory factor analysis (CFA). The results for alpha were arguably satisfactory on all but two scales. Alpha does not demonstrate unidimensionality of scores, which is critical to score interpretation, and therefore the results for the CFA, which indicated rejection of the hypothesized model, take precedence. Non-normal score distributions were also a significant property of the data generated by this Japanese translation of the SILL. It is argued that while these findings are for this particular Japanese translation of the SILL, other translations into Japanese or other languages should also be submitted to similar analyses as part of cumulative work into the validity of scores generated by this instrument, because its presence in the literature, and potentially in pedagogical practice, is significant.

Keywords: SILL, Strategy Inventory for Language Learning, Learning Strategies, Validity

Introduction
Language learning strategies (LLS) research within the second language acquisition (SLA) field has occupied some part of the research agenda, to one degree or another, for more than four decades now. The intellectual fabric of this interest was inherited from well beyond the field of SLA, and in some cases from much earlier work. With respect to the earliest work one might include, for example, Immanuel Kant and his contribution to the notion of autonomy which, although more elemental and bound up with the moral agent rather than the modern learner, contributed to one assumption closely tied up with modern ideas about learning strategies, i.e. that they presuppose an autonomous learner with the capacity for personal
agency in his/her own learning.

More recently, and more proximate to pedagogy, the cognitive revolution which was catalyzed by the transformative ideas of Chomsky (1957, 1965) and his reassertion of a radical mentalism in the explanation of the most quintessential of human behaviors, namely language behavior, came to have a significant impact on the enumeration and elaboration of the capacities which the learner brings to learning. This revolution came in the face of the apparent theoretical poverty inherent in the behaviorist paradigm which presumed intellectual passivity rather than activity; an inheritance from its deep roots in empiricist epistemologies.

While behaviorism was dominant at the time of Chomsky’s intervention, other psychological systems of explanation which had predated Chomsky’s contribution curated notions of psychological agency and autonomy through the years of behaviorist dominance; and these were to be reprised in the post-Chomskyan intellectual milieu which was more accommodating to sources of explanation comfortable with the interior world of the human agent. They were, however, also to present tools of explanation for mental behavior which went beyond the fundamental innatism expounded by Chomsky in the form of the Language Acquisition Device (LAD). For instance, the work of Piaget (e.g. 2001a; 2001b; first published in 1926 and 1951 respectively), conducted over decades and subordinated to his project of outlining a genetic epistemology, had produced accounts of the child’s mental life which were essentially constructivist and adaptive; and which became part of the fabric of cognitive accounts of human intelligence, development, and learning. These included explanatory tools such as assimilation and accommodation, classification, and so forth which were all subordinated to the deeper developmental force of cognitive adaptation.

Also reprised in the post-Chomskyan revolution, a little belatedly due to historical factors such as the relative intellectual isolation of the post-War USSR, were the ideas of Vygotsky (1978, 1986) whose Marxian engagement with mind inverted the question of how individual minds aggregate to comprise society to one of how society constructs the mind. Central to this engagement was the notion of self-regulation which was theoretically located in the more primitive form of ‘regulation by the other’ as part of the process of child development. These ideas were also to have a more specific revival within LLS in the 2000s where socio-cultural concerns started to find expression in models of LLSs (covered below).

Were it not for the unseating of behaviorism and the invigoration of mentalist and then cognitive and constructivist approaches (all three more
comfortable with the notion of mind) to development, intelligence and learning, the interest in LLSs within SLA might not have emerged as it did in the 1970s. After all, approaches to language pedagogy, underpinned by behaviorist thinking had included, for example, the audio-lingual method where pattern practice and drills were the signature features, and where the learner adopted an essentially passive and receptive mental posture. These kinds of approaches afforded no room for the strategic intervention of the learner in his/her own learning. A further issue, which arguably tilled the intellectual soil for the germinating interest in LLSs within SLA, was that Chomsky’s extreme innatism was theoretically elaborated in service to an account of first language acquisition (FLA) where outcome is relatively invariable (all children, other than in cases of pathology, acquire their native language). The case with SLA was quite different however. A lot of variation in outcome is almost always observed, and in this context the Chomskyan position of extreme innatism had limited theoretical potential. Indeed, the central endeavor of SLA theory and its associated pedagogy became the explanation and mastery of this variation in outcome. Into this context emerged early and important contributions to rising interest in LLSs including work centered around the notion of the “good language learner” (GLL). This work put individual differences at the core of explanatory work on L2 outcome (Naiman, Frohlich, & Todesco, 1975; Rubin, 1975; Stern, 1975).

The interest in the GLL, a formulation of the problematic of variable learning outcomes somewhat centered on the person but important nonetheless, was to give way to a research agenda more closely associated with the actual strategies which promoted successful language learning. The shift here is somewhat nuanced because strategies are executed by the person in the end, but suffice to state that theoretical developments in cognitive research related more widely to learning in general were providing conceptual and theoretical tools for the modeling of learning strategies, and SLA became more pre-occupied with this modeling than with persons who were good at learning languages. One important development, of many, which came from within cognitive research was the attention given to the theoretical notion of metacognition (Flavell, 1979; Pressley, Borkowski, & Schneider, 1987) which is critical to self-regulative cognitive adaptation; and which is essentially the superordinate area of executive control for the appropriate and successful execution of subordinate strategies for learning.

In this context, the research agenda for SLA in the area of LLS evolved to become the task of enumerating strategies, and submitting them to
coherent taxonomies and models. Work on this came as early as Bialystok (1978) where one part of the elaborated theoretical model of SLA included LLSs with citation of Stern (1975) and Rubin (1975). An important paper a little later (Bialystok, 1981) gave more direct focus to conscious learning strategies in SLA while another contribution came at the same time from Rubin (1981) in work which followed up on the earlier interest in the GLL. Rubin’s taxonomy saw some adoption in subsequent research (e.g. Vann & Abraham, 1990). Another important trajectory of research (Chamot & Kupper, 1989; Chamot & O’Malley, 1987; O’Malley, Chamot, Stewner-Manzanares, Kupper, & Russo, 1985a, 1985b) culminated in a book (O’Malley & Chamot, 1990) by two of the previously involved researchers which offered a taxonomy of LLSs.

All of the above work was important, but perhaps the taxonomy which gained the most exposure and traction in the literature was that of Oxford (1990). Arguably, the most significant reason for the adoption of the taxonomy by subsequent researchers was the offering of companion instrumentation associated with the taxonomy in the form of the Strategy Inventory for Language Learning (SILL). According to Oxford (2011) the SILL was influenced by Weinstein, Goetz and Alexander (1988). The extensive use of this instrument (for a few examples in more prominent journals see J. Green & Oxford, 1995; Hong-Nam & Leavell, 2006; McCullen, 2009; Nisbet, Tindall, & Arroyo, 2005; Wharton, 2000; Yang, 1999) facilitated a large body of empirical research bound to the taxonomy offered by Oxford; and the use of the instrument, while not quite as prolific as in the past, continues to the present (e.g. Sun, Mantero, & Summers, 2014; Tam, 2013).

The use of the instrument has been attended by concerns with respect to the psychometrics of scores generated by the instrument with most of the concern focused on the structural validity of these scores. Often the more positive evidence for the psychometric properties of the instrument have been based on the value derived for Cronbach’s alpha; but in many cases the value for alpha has erroneously been derived for scores on the entire instrument rather than scores derived on each of the subscales making up the instrument. Nonetheless, even when correctly derived, the value for alpha is an index of limited diagnostic worth (Bentler, 2009; S. B. Green & Yang, 2009a, 2009b; Revelle & Zinbarg, 2009; Sijtsma, 2009a, 2009b) and does not demonstrate the critical property of unidimensionality of scores (Cortina, 1993; S. B. Green, Lissitz, & Mulaik, 1977) generated by a subscale requiring interpretation. On the other hand, the more negative evidence for the psychometric properties of the instrument have often been
based on executions of Exploratory Factor Analysis (EFA), and the outcomes have been negative in the sense that the simple structure arrived at via the EFA in each case bears little correspondence with the measurement model offered by Oxford (1990). These studies have arrived at solutions with differing numbers of factors and differing items representing the factors (e.g. El-Dib, 2004; Nyikos & Oxford, 1993).

In this overall context, few studies have employed confirmatory factor analysis (CFA) which is a more powerful tool than EFA and which allows for a direct and a priori test of the model hypothesized by the author for a psychometric instrument. One notable exception to this was a study conducted by Hsiao and Oxford (2002). This study reported that Oxfo's taxonomy of LLSSs, which involves the six factors or scales comprising the SILL, offered the best fit to data generated with the SILL on a sample of 517 college students in Taiwan when compared with the fit of competing taxonomies derived from the work of other authors. However, these taxonomies from other authors were not used in the construct formulation of the SILL, and therefore it is to be expected that Oxfo's taxonomy should better align with the dimensionality of the data generated by the SILL. Also, it is important to note that while Oxfo's taxonomy provided a relatively better fit than taxonomies proposed by other authors when tested against the data generated by the SILL, it still did not provide adequate fit.

While the Hsiao and Oxford (2002) study represents an important contribution to the literature because it uses CFA, which affords an a priori and direct test of the model and associated scoring regime hypothesized by Oxford for the SILL, evidence for validity is a cumulative process (Messick, 1989) and further studies are always required to allow a larger empirical picture to emerge. In other words, the model hypothesized by Oxford needs to be the subject of further direct and a priori tests using samples from other populations. The study reported in this paper represents this kind of contribution as part of the process of normal science. A relatively large dataset was obtained for the SILL and the measurement model hypothesized by Oxford by virtue of the scoring regime advocated for the instrument was directly tested using CFA as the method.

Method

Data for the Japanese version of the SILL was collected in 2008 at a public university in Japan. The data was entered into a Microsoft Access database and analyzed using SPSS (Version 21) and IBM AMOS (Version 21). The
SILL has more than one translation available in the literature. The particular version selected has been used for research purposes in the Japanese literature (e.g. Iwasaki, 2006) and there are findings premised upon the scores it generates.

Participants and Treatment of the Data

Informed consent was obtained for participation, and the identities of all participants were protected. Participants were freshmen (837 cases). With respect to the treatment of missing data, there were 81 cases where a participant had not responded to one or more items. These cases were inspected together and the omissions were judged to be at random rather than systematic, and therefore they were deleted. The final dataset submitted to analysis numbered 756 cases.

The age range of the sample was highly constrained with 99% of the sample falling within the range of 18 years to 21 years. There were 389 males (52%) and 365 females (48%) with two non-responses.

Instrument

The SILL comprises 50 items and these items are hypothesized (Oxford, 1990) to generate scores which conform to a six-factor structure with each factor underpinning a scoring and interpretive subscale. The following items make up the following subscales: Items 1 – 9 (Memory Strategies), Items 10 – 23 (Cognitive Strategies), Items 24 – 29 (Compensation Strategies), Items 30 – 38 (Metacognitive Strategies), Items 39 – 44 (Affective Strategies), and Items 45 – 50 (Social Strategies). Notably, these items are not placed randomly in the overall instrument but in continuous sequences. While this might not be the normal approach, and allows respondents to detect item patterns more easily, the structure was retained in order to remain consistent with the author’s design.

Each item on the SILL is responded to on a Likert scale under the following semantic anchors (associated score in parentheses): Never or almost never true of me (1), Usually not true of me (2), Somewhat true of me (3), Usually true of me (4), Always or almost always true of me (5).

Results

The results for properties of the data in terms of normal distributions are reported first. Following this the results for Cronbach’s alpha, including
the 95% confidence intervals (Fan & Thompson, 2001), are reported (see Table 1). The results for the most important part of the analysis, the CFA for the hypothesized structure of scores for the SILL (Oxford, 1990), are then reported.

In terms of distribution, the most important features of the data were as follows. There were only 15 items, out of the total of 50 making up the instrument, where the value for the critical ratio for skew fell below the predetermined cutoff value of 3. The results with respect to kurtosis were somewhat better, with 33 items meeting the threshold. Nonetheless, this still leaves 17 items which did not meet the threshold. With regard to multivariate normality, the value for Mardia’s coefficient was derived (116) and was found to be high. These properties of the data influenced subsequent analyses because the Bollen and Stine (1993) bootstrap procedure was adopted to specifically cope with the problem of non-normality in the score distributions.

In terms of the results for Cronbach’s alpha the results are available for inspection in Table 1. The rule-of-thumb criterion offered by Nunnally and Bernstein (1994) is .70 for an acceptable alpha value. The Compensation and Affective scales were on this threshold while the others were above it. The lower bound of the confidence intervals for alpha was below the threshold for two scales; and again these were the Compensation and Affective scales.

### Table 1
*Cronbach’s alphas and confidence intervals (95%) for alpha for subscales of the Strategy Inventory for language learning*

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Alpha</th>
<th>Confidence Interval (95%) for Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Lower bound</td>
</tr>
<tr>
<td>Memory Strategies</td>
<td>.74</td>
<td>.71</td>
</tr>
<tr>
<td>Cognitive Strategies</td>
<td>.81</td>
<td>.79</td>
</tr>
<tr>
<td>Compensation Strategies</td>
<td>.70</td>
<td>.67</td>
</tr>
<tr>
<td>Metacognitive Strategies</td>
<td>.88</td>
<td>.87</td>
</tr>
<tr>
<td>Affective Strategies</td>
<td>.70</td>
<td>.67</td>
</tr>
<tr>
<td>Social Strategies</td>
<td>.79</td>
<td>.76</td>
</tr>
</tbody>
</table>

With regard to the CFA, the model was specified so that items indicated the respective factors hypothesized by Oxford (1990). Items were
permitted to indicate only one factor which is consistent with the specification of the unidimensional model. In addition, and also consistent with the unidimensional model, error was not permitted to correlate. Factors, however, were permitted to correlate consistent with the assumption that orthogonal factors are highly unlikely in psychological measurement (Kline, 1994) and consistent with normal practices within the field. There were 1275 distinct sample moments in the model and 115 distinct parameters to be estimated. There were 1160 degrees of freedom. The criterion of overidentification was met. The method of estimation was most likelihood (ML) which makes normal-theory assumptions and this was the same method used by Hsiao and Oxford (2002). However, the method is still used with Likert scale data (which is ordinal data) in practice provided that the points of discrimination on the scale are not less than five (West, Finch, & Curran, 1995). Non-normal distributions further threaten the normal theory assumptions of ML estimation if these departures are severe, and because non-normality was a property of the data, the further analytical method of the Bollen and Stein bootstrap procedure was adopted.

The model was evaluated using a combination of the chi-square value and its associated probability level as well as a range of four indexes recommended by Hu and Bentler (1999). These indexes, among others, are routinely used within the literature because using the chi-square exclusively, which is a test statistic and not an index, leads to over-rejection of models. This is because the chi-square is very sensitive and will reject models on the basis of only small, or even trivial, differences between the dimensionality of scores generated by an instrument and the model hypothesized for it.

The model was significant at the .01 probability threshold (chi-square 4289). This means that the model hypothesized for the structure of scores generated by the instrument (Oxford, 1990) was significantly different from the dimensionality of scores in the dataset collected for this study. Thus, on the basis of this sample the model hypothesized by Oxford should be rejected. However, because the chi-square tends to over reject, the results for the indexes are also important. With respect to these indexes, the results were as follows (the Hu and Bentler [1999] cutoffs are in parentheses): TLI, .760 (> .95); CFI, .772 (> .95); RMSEA, .060 (< .06); and SRMSR .062 (< .08). In terms of these indexes, the TLI and the CFI represent strong evidence against the model, while the RMSEA produced a value on the threshold, and the SRMSR value which expresses the residuals after the model is fitted was satisfactory.
Finally, the results for the Bollen and Stein (1993) bootstrap procedure, adopted due to the non-normal distributions in much of the data, also indicated a rejection of the model. The probability level associated with the bootstrap was significant at the .01 level, and in the reverse logic of CFA this has to be interpreted as a rejection of the model hypothesized by Oxford (1990).

While the results for the RMSEA and SRMSR indexes were acceptable, all of these results taken in triangulation indicate that the model should be rejected. The Hu and Bentler (1999) indexes are recommended to be used together which means that all of them have to be acceptable for there to be model fit.

**Discussion**

In framing the results for this study, the following context and caution should be noted. First, this study dealt with one of the translations of the SILL present in the Japanese literature, and the results are necessarily bound to this instance. Other translations should be examined on their own merits; and this includes translations into other languages as well. The original English version of the SILL should also continue to be examined on its own merits when used in the English-speaking population for which it was first designed.

Nonetheless, the results obtained for this translation when used in the Japanese population, while not applicable to other instances of the instrument used in other populations, should at least be taken as suggesting the agenda under which other instruments in other populations should be examined. They should also be seen as supportive of an emerging picture if they align with results from other versions used for other populations. Second, and as stated earlier, Messick (1989) argued that demonstration of validity, or lack thereof, is a cumulative project, and each study is an incremental step in this process. This view applies to the results obtained in this study which represent an incremental contribution rather than a definitive one, and a contribution as part of the process of normal science.

With regard to the distributions of scores for items making up the instrument, these distributions were, on the whole, non-normal and while this is found quite often in practice (Micceri, 1989), this could also be seen as a focus for revision of the translation or, indeed, of the original instrument. Excessive non-normal distributions necessarily imply a loss of information, and normal-theory analyses become progressively less tenable as the problem becomes more serious; even if the assumptions around
normal-theory analyses have been relaxed for use with ordinal scales. These non-normal distributions were sufficient to influence the subsequent analysis in this study when the Bollen and Stein (1993) bootstrap procedure was introduced to cope with them.

The results for Cronbach’s alpha were generally acceptable. As reported in the results section, the rule-of-thumb criterion of .70 for alpha, introduced by Nunnally and Bernstein (1994), is typically used to assess whether a returned value for alpha is acceptable or not. According to this criterion the Compensation and Affective scales were the weakest, falling right on this threshold rather than above it. There are two issues which need to be taken into critical consideration when applying this rule-of-thumb however. The first issue is that alpha is positively biased by the number of items in a scale (Cortina, 1993) and while the scales making up the SILL vary in terms of the number of items, none of the scales has less than six items. In this context, both the Compensation and Affective scales produced values for alpha which could be interpreted as actually quite low (both scales comprising 6 items). The second issue is that alpha does not demonstrate unidimensionality (Cortina, 1993; S. B. Green et al., 1977), as has been mentioned above, and in fact the index has come under criticism for this (Bentler, 2009; S. B. Green & Yang, 2009a, 2009b; Revelle & Zinbarg, 2009; Sijtsma, 2009a, 2009b). Therefore, other forms of analyses (such as CFA) should be adopted which do demonstrate unidimensionality of scores.

This analytical weakness with alpha turns the discussion over to the results of the CFA. Overall, the model hypothesized was rejected when all results were taken in triangulation, and this corresponds with Hsiao and Oxford (2002); although the combination of indexes adopted by these authors was not informed by Hu and Bentler (1999). The first index used by Hsiao and Oxford was the CFI for which they obtained a value of .748 and this was very close to the value of .772 obtained in this study for the same index. The second index used was the NNFI. The NNFI (or Non-normed Fit Index) is the same index as the TLI (or Tucker Lewis Index), but under different name. On this index, Hsiao and Oxford obtained a value of .734 which was also similar to the value of .760 obtained in this study. So there is correspondence between the results for the Hsiao and Oxford study and the results for this study.

The Hsiao and Oxford study claimed, on the basis of reported results, that 1) the 6-factor model hypothesized by Oxford (1990) provided a better fit than other models informed by other taxonomies and theory, and 2) that the 6-factor model still did not provide adequate fit in and of itself. It could
be added with respect to the first claim however, that it should not be surprising that models emerging from other theory and taxonomies should provide a worse fit on an instrument explicitly designed for Oxford's theory and associated taxonomy. Thus the most important claim is the second one; i.e. that Oxford's (1990) hypothesized six-factor model for the SILL did not provide a plausible fit to the scores generated by the instrument. The results in this study therefore specifically support this second claim made by Hsiao and Oxford.

One important area of departure for the Hsiao and Oxford (2002) study from the typical administration format for the instrument was the actual response scale for each item. While the original English-language version of the instrument used a five-point Likert scale with semantic anchors on each point of discrimination as reported above, the Hsiao and Oxford study used an eight-point scale semantically anchored at either end of the scale but not on the six intermediate points of discrimination. While this change in scale format did not provide an outcome on the CFA which was very different from the outcome obtained in this study, it is possible that alternative scales might improve score distribution.

In conclusion, the findings from this study have the important implication that practitioners using this translation should treat the scoring regime associated with the hypothesized six-factor model with skepticism. The results reported in this study constitute negative evidence for this model. Furthermore, this study corroborates the Hsiao and Oxford (2002) study across another population and translation, which suggests that other versions of the instrument, whether these are alternative Japanese translations, other-language translations, or the original English version, should be treated with caution until there is clear empirical support for the model in each case.

Acknowledgement

I would like to thank the Durban University of Technology in South Africa where I was a Visiting Fellow during the writing of part of this paper. The data itself was collected before this visiting fellowship.

References

Learning, 28(1), 69-83.
McCullen, M. (2009). Using language learning strategies to improve the writing skills...
Sijtsma, K. (2009a). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika, 74(1), 107-120.


