Individual Differences in Language Education: An International Journal

https://doi.org/10.32038/idle.2024.02.01

Developing an aptitude test based on Pienemann's SLA theory

Peter Skehan*
Institute of Education, University College London, UK
Erica Ying Shan Chan
Independent Researcher, Germany

Correspondence
Email: peterskehan@gmail.com

Abstract

This article reports on a proof-of-concept study regarding the development of a language aptitude test. The test, which measures inductive language learning, is based on Pienemann's account of second language acquisition (SLA). His theory proposes several stages which are followed in second language development, and the aptitude test draws on these stages to build in progressive complexity as the test proceeds. Data is reported suggesting that actual performance is consistent with the Pienemann's stages. Finally, the place of such a theoretically-based test within aptitude measurement is discussed, as well as the way the test could be adjusted to better predict attainment at different levels of proficiency.

ARTICLE HISTORY

Received: 10 July 2024
Revised: 15 September 2024
Accepted: 02 December 2024

KEYWORDS

language aptitude, language aptitude test, Pienemann's SLA theory

How to cite this article (APA 7th Edition):
Skehan, P., & Chan, E. Y. S. (2024). Developing an aptitude test based on Pienemann's SLA theory. *Individual Differences in Language Education: An International Journal*, *2*, 1-20. https://doi.org/10.32038/idle.2024.02.01

Introduction

What could be called the 'modern' approach to aptitude started with the work of J. B. Carroll. From a theoretical perspective, he proposed his celebrated four-factor view of aptitude, but more practically, with Stanley Sapon, and through the publication of the Modern Language Aptitude Test (MLAT: Carroll & Sapon, 1957), he established the most influential foreign language aptitude test of all time. In this way he offered an explanatory account of aptitude, one that was plausible in terms of psychological theorising from that time, and also a practical means of assessing aptitude. The account was also fairly general - not focussed on particular groups of learners, such as high achievers, but relevant to all adult (and young adult) learners.

Carroll's four-factor theory was based on an extensive research project, in which a large number (over forty) of candidate tests were produced. Carroll (1962) explained that these tests were generated under three headings: those derived from more general accounts of human abilities, an area to which Carroll made distinctive contributions

(Carroll, 1993); those developed because they seemed to reflect skills related to the specific nature of language learning; and finally work sample tests, which attempted to mirror the process of language learning. These were administered to large numbers of disparate language learners, and the candidate tests were correlated with one another and then with outcome measures of actual language learning. The intention, with this analysis, had twin aims: to identify aptitude tests which largely duplicated one another (so that the best could be selected, and the rest rejected), and to explore which tests were most effective at predicting higher achievement on tests of language learning. Then, finally, a battery, the MLAT, could be assembled, consisting of a small but effective set of tests that did not correlate particularly highly with one another, but were all strong predictors of success.

As it happens, the relationship between the practical contribution (the development of the MLAT), and the theoretical contribution (the four-factor theory), although it should have been straightforward, warrants some scrutiny. Carroll's (1962) four factors were phonemic coding ability (how unfamiliar sound could be processed in order to aid retention), associative memory, grammatical sensitivity, and inductive language learning ability. The third and fourth of these are the starting point for this article. Grammatical sensitivity (considered a specific-skill test in terms of its origin) probes the capacity to identify the functions that words fulfil in sentences. In other words, it is concerned with grammatical analogies. This particular test, a cornerstone of the MLAT, has proved to be widely used, and generates consistent correlations with achievement scores.

Carroll's final factor, inductive language learning (ILL), is extremely interesting. Within the large exploratory test battery, there was a test of this construct, and indeed, this test played an important role in the development of the four-factor theory, largely through its patterning in the associated factor analytic work. Carroll (1962) proposed that it involves the capacity to notice language patterns, to infer structure, and to be able to generalise and extrapolate. The test in question, very much a work sample test, was based on Tem-Tem, a made-up language. The test presented material to be learned inductively. After the initial instructions, no further use of English was involved. Presentation of the Tem-Tem material was auditory, using a tape recorder, synchronised with a film strip. Both vocabulary and grammar were to be learned. This sub-test alone required half an hour as well as the relatively complex equipment for the time (the 1950s), which was difficult to set up and unreliable. On the basis of this administrative complexity, and despite the very interesting statistical contribution the sub-test made, it was not included within the MLAT (although, as we have seen, it had a major impact on theory).

Writing some seventy years later, one has to say that there were fateful consequences of the decision not to include the inductive language learning test in the MLAT, even though this was done for eminently practical reasons. The time required, together with the equipment implications, sealed the fate of the Tem-Tem test. Of course, now we

are looking back from an era with pervasive audio and visual computer-controlled delivery, and we have to conclude that this was disappointing. As a result, language structure, within the MLAT, was represented only by the Words in Sentences sub-test – fifteen minutes long and entirely paper-and-pencil delivered. Despite its predictive effectiveness, the Words in Sentences sub-test's potential association with outdated audiolingual methodologies, decontextualised learning, and metalinguistic knowledge, has questioned the validity of this measure in relation to more naturalistic and acquisition-oriented learning, and perhaps been the main reason that some have argued that the MLAT is passé.

Since the MLAT's publication, there have been a number of additional initiatives in producing aptitude tests. It is striking that, apart from some translations for administration to L1s other than English, there have not been any attempts to rework the Words in Sentences sub-test, to take its underlying principles, and to try to improve upon it. In contrast, it is commonplace where a new aptitude initiative attempts to assess the ability to handle language structure that some variant of an inductive language learning test will be developed. There are, now, quite a large number of these. Pimsleur (1968), very soon after the development of the MLAT, produced the Pimsleur Language Aptitude Battery (Pimsleur, 1966), which incorporated, as Part 4, an inductive language learning sub-test based on Kabardian. The LLAMA battery, although loosely based on the MLAT, addresses language structure through an inductive language learning test (Meara, 2005). The York Language Aptitude Test (Green, 1975) (created, engagingly, because postal strikes meant that the PLAB was not delivered in time for a research project), unsurprisingly, was modelled on Part 4 of the PLAB. The DLAB (Petersen & Al-Haik, 1976) contains two sub-tests which assess inductive language learning ability: Part 3, Language Analysis, and Part 4, Foreign Language Concept Formation. Grigorenko, Sternberg, and Ehrman (2002), in the CANAL-F built in essentially inductive language learning formats into two sub-tests of the wider battery, Sentential Inference and Language Learning Rules. Pan (2023), in a Ph.D. study, developed an internet-delivered inductive language learning test specifically for Chinese L1 learners of English L2. In general, though, these tests have been paper-and-pencil in nature, and not incorporated any auditory component. This is ironic, given that Carroll and Sapon used auditory presentation precisely because they thought this would be more appropriate for the learning of oral language, a feature which would have made an imagined, reworked MLAT more appropriate to contemporary methodologies!

Strikingly then, in almost all language aptitude tests developed since the MLAT (Hi-LAB is a notable exception, Linck et al., 2013), a concern for morphosyntactic language aptitude has been based on Carroll's inductive language learning construct. It is worthwhile, therefore, to explore more deeply what is involved in these tests, how they are similar, and how they differ. One aspect of this concerns the emphasis on presentation, structured example-based learning and testing. In Carroll and Sapon's (1957) Tem-Tem test, although the focus was on inductive language learning, there

was presentation of material arranged to support this learning. Other inductive learning tests have varied somewhat in the relationship between these elements. The DLAB (Petersen & Al-Haik, 1975) provides quite a lot of exposition, for example, while the York Language Aptitude Test (Green, 1975) focusses on example material and testing. The other examples vary in the balance they strike between these different components.

Most interesting, though, is to focus on the language content itself. Table 1 shows what the main emphases are in different inductive language learning tests.

Table 1

Content Focus in the Different Tests

Test	Detailed Content
Pimsleur	Article, Article plus Plural, Present tense, Strongly morphological
York	Swedish-based (+ explicit with English): Plural formation, Definite article Definite article + Plural, Present tense: Very morphological
DLAB	Part 3: Also explicit: Noun-adjective; Possessive; Sentence structure (quite complex); Reminder of five rules, then tested material.
	Part 3: Foreign Language Concept Formation: In fact, *sentence* formation + Vocabulary learning, Agreement: Quite difficult
Pan	Nominal ending (number, gender); Verbal inflection (Affirmative, Negative: Past, Future); Word order (Adj. position, Adv. position)
Grigorenko, Sternberg, & Ehrman	Section 4: Sentence translation (after some vocab learning) example: Past tense plus Direct + Indirect objects; Section 5: Possessive; Adj-noun modification
LLAMA	Adjectives follow nouns, Numerals precede nouns; Verb initial; Nouns take singular marker but not plural and singular marker follows noun; Affixes have several forms, similar to gender.

Broadly here the different ILL tests seem to focus on morphology, syntax, and vocabulary, but with different emphases. Morphology figures prominently, typically with features such as gender, adjective-noun agreement, and number. Occasionally, verb inflection is included. Syntax is, perhaps, slightly less represented, with features such as word order (but this is typically within a clause), or the contrast between direct and indirect objects. Vocabulary varies widely across tests and a few of the tests do make it a slight feature, such as the York Test, DLAB, and CANAL-F.

Inductive Language Learning, Methodology, and Second Language Acquisition

In order to understand how we can conceptualise and test an aptitude to handle structure in language, a brief digression is necessary. As mentioned earlier, the MLAT was developed at a time when audiolingualism was a very influential methodology. Since then, there has been something of a revolution in approaches to how second and foreign languages are most effectively learned. Far more attention now is paid to communicative methods, and most recently to task-based approaches to instruction. In parallel, the field of second language acquisition (SLA) has emerged and become very influential. Taken together, a more contemporary view of language instruction would be that naturalistic processes are considered more important, and that learners

need to proceed at their own pace, and in their own path. Interaction is seen as a supportive environment for development, feedback is important when timely and personalised, and there is the possibility that language development draws on both implicit and explicit processes.

If we look at aptitude through this sort of lens, there are interesting implications for the contrast between grammatical sensitivity and inductive language learning ability. Centrally, grammatical sensitivity is concerned with the ability to recognise structure in language. In principle, as Carroll (1973) suggests, a relatively small exposure to education would be enough for this to be feasible. However, it has been argued that grammatical sensitivity conflates a basic capacity to understand patterns, on the one hand, with metalinguistic abilities, on the other. This certainly complicates the situation. There may be an advantage for students who have had to do language analysis as part of their education, whether this involved metalinguistic terminology or not. In this view, getting a high score on grammatical sensitivity would be a mix of language pattern ability and previous experience. This might not impair prediction, and could even enhance it. However, it might cloud any interpretation of the nature of the contribution of this sub-test to predicting language learning success, and this might apply more strongly to communicative contexts.

In contrast, inductive language learning ability sub-tests are more consistent with the sorts of processes that SLA research and task-based methodologies highlight. They draw upon noticing, inferencing, attention management, generalising and extrapolating, all processes integral to SLA research. So, while the predictive qualities of the Words in Sentences test are still important, it has to be said that the different versions of ILL test development appear to be more theoretically defensible.

This analysis leads to something of a paradox. Even though an inductive language learning approach is more consistent with the theoretical and methodological developments that have just been discussed, it can nonetheless be argued that the actual ILL tests that have been developed remain somewhat detached from these wider developments. The features outlined in Table 1 are essentially atheoretical, and do not often connect with strands of SLA research. They come across as unsystematic in their coverage, varying across tests, and opportunistic rather than principled. There is no connection with any acquisitional data, even though a great deal of such data is available. As a result, given the greater consistency of an inductive language learning approach to current theory and methodology, it seems appropriate to explore how such an ILL test can have a more theoretical grounding.

An immediate problem is that, while SLA research has shown huge vitality, there are not many theories which provide developmental detail of a general sort. Insights abound, but are often linked to particular language sub-systems. After examining a number of possible approaches, and finding them lacking in generality, we decided to base an approach to inductive language learning aptitude test construction on the

work of Manfred Pienemann (1998), which outlines general, even universal macro stages of development.

The approach has a number of advantages. First, it is extremely detailed, and provides clarity about the various stages of L2 development. Second, it is supported by a wealth of empirical data. Third, it has been shown to be valid cross-linguistically, and apply to many L1-L2 combinations (Pienemann, 2005). Finally, it is an effective mix of cognitive and linguistic perspectives. It is grounded in linguistic development, but simultaneously, it has strong cognitive processing foundations. In this way, it is relevant to one of the central debates within aptitude theorising: is second language development linguistic in nature, or cognitive? The Pienemann approach allows both accounts to be influential within the same system.

But most important of all, from these advantages, it is remarkably clear about a series of developmental stages. It suggests that six macro stages are fundamental:

- Formulaic language
- Canonical Order
- Adverb Preposing
- Verb separation
- Inversion
- Verb-end

They represent the stages that Pienemann proposes will be followed in the acquisition of a second language. This particular sequence is justified because each stage introduces progressively greater complexity, of word order, of connections between elements (and their distance from one another), and of movement of elements (and their distance of movement). More importantly, each stage presupposes all the preceding stages. In other words, to arrive at a particular stage, the operations and processing required in lower stages have to be in place. The capacity to handle adverb preposing, for example, is required before verb separation can be mastered, and so on. Although the theory seems to be dealing with relatively few elements, it has been applied effectively to the range of language structures that need to be acquired within second language acquisition. It has also been examined from a cross-cultural perspective and shown to be relevant for different L1-L2 combinations (Pienemann, 2005).

Most fundamentally, the cumulative nature of this account renders it accessible even to artificial languages. It provides a sampling frame and a sequence of increasing difficulty. In this respect, it contrasts quite markedly with the other theories of second language learning which are available. Universal Grammar, Functionalism, and the like offer suggestions about development in particular areas. They do not, though, clarify an entire sequence of learning. For that reason, Pienemann's Processability

Theory will serve as the basis here for exploring how an aptitude test focused on pattern recognition can be developed..

Even so, the existence of a theory of development does not translate in any obvious way to actual test development. Pienemann's sequence operates over an extended time period, and each stage needs to be established before learners can realistically move on to the next. The evidence certainly suggests that this takes time. Yet time is of the essence in aptitude measurement, and part of the challenge is not simply to measure relevant abilities, but to do so relatively quickly. We saw that this was one of the reasons for the abandonment of the very interesting Tem-Tem test. So, using Pienemann's theory as the basis for aptitude test construction (and this would probably be only one sub-test in a longer battery) does have its problems.

Even so, the different stages Pienemann proposes can be used as the basis for developing different sections of a test which probe the learner's ability to handle these progressively greater levels of complexity. First, each stage covers quite a lot of sub-areas, and within these sub-areas, choices can be made to select the most accessible aspects. This gives respondents a reasonable challenge in what they have to do. The range of difficulty within each stage enables the possibility of a progression within that stage. Second, the movement from one stage to the next enables the process to start afresh, to some degree, as one moves to the next stage. This gives learners the chance to return to easier items, and give them a new opportunity to re-engage with the overall sequence.

Typically, in testing, a distinction is made between speed and power. The former emphasises what can be achieved *within a particular timeframe*. The latter is less concerned with timeframe and more concerned with the maximum difficulty level that can be achieved (Cronbach, 1970). Most other inductive language learning tests emphasise the speed alternative. Using a Pienemann framework, as we have outlined it, enables the combination of a speed test, and also a strong 'power' component, in that moving through the items of the test is not simply accumulating a higher score – it also reflects an ability to handle progressively more searching aspects of the language system. We argue that this approach leads to a higher level of construct validity.

Developing the Test

The choice in developing a new inductive language learning test lies between using a little-known language or an artificial language. In the present case, the decision was to take the second of these alternatives, and the invented language of Latejami was chosen (Morneau 1995/2006). This language is extensive and extremely well-documented. It also contains a great deal of linguistic sophistication with the result that implementing Pienemann's views on developing language structure was feasible. There are dictionaries and grammars available (www.eskimo.com/~ram/latejami), and the language purports to provide a complete communicational system,

particularly adapted for machine translation. Its comprehensiveness was attractive because of the way this enables test material to be constructed without undue constraints. It is also a strongly morphological language, with extensive agreement marking, allowing considerable scope to implement different features of grammar.

Following Pienemann (1998), it was decided to have six levels within the test. These were:

- Word: lemma and basic word structure
- Word: category procedure, emphasising plurality
- Syntax: simple word order
- Syntax: noun agreement
- Syntax: verb formation
- Syntax: sentence generation

It has to be said that the mapping between these six levels and Pienemann's developmental stages is not exact. The first two levels here correspond to a broad interpretation of Pienemann's first stage, formulaic language, in that they are concerned with the word level of analysis, though not particularly emphasising formulaic language *per se*. Then the next three levels do roughly correspond to Pienemann's second to fifth stages, while the final level here is not particularly based on Pienemann's sixth stage. Instead, it is rather an amalgamation of the three preceding levels.

Test Development

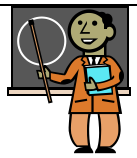
In each section, material is first provided illustrating the principles that are relevant to that section, and examples of correctly formed responses are given. Then there are trial items to probe what has been learned. Typically, there are five to ten trial items per section. Throughout the test, the verbal material is accompanied by visuals which serve as cues for meanings that need to be expressed in written form. Most sections are multiple choice in nature with Latejami words as alternatives and illustrations to indicate the target item. The following sections provide examples of each of the sections of the test. It should be noted that the L1 of the test taker is irrelevant for the completion of this test. The only L1-specific elements are some general orienting instructions at the outset (which could easily be modified for any different L1 populations), and the use of one or two words, such as 'name' in certain questions. These elements, too, can easily be adapted for speakers of different L1s.

Section One

This section is aimed at testing whether a test-taker can choose the right word in response to a picture, based on the given examples concerning the patterns of Latejami word formation. The processing procedure, word/lemma, as suggested by Pienemann, is the basis for this section.

A Latejami noun is formed by putting a modifier, the morpheme root which indicates what category a particular noun belongs to, to the left of a classifier, the root which modifies the word semantically. The word for bicycle (xetimi), for example, is formed by a modifier (xe) meaning two, a classifier (tim) referring to vehicle and a suffix (i) referring to a noun.

Based on some examples, a test taker is required to choose an answer out of four options in each question. There are, in total, nine questions with four of these testing classifiers, two testing modifiers and three testing the combination of both. Here is one trial item from Section One.

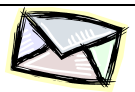
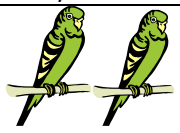
			
joybegi	kobaybegi	zetimi	xetimi

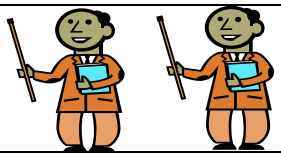
1. A. febegi
   B. byetimi
   C. bujisi
   D. kobiji

Section Two

After a noun is activated, a category procedure is necessary to indicate the number of a noun. This section, therefore, is aimed at testing whether a test-taker can change the form of a singular noun to mark its plurality based on the examples given.

To make a singular noun plural, a plural prefix (ly or li) is necessary. In the original Latejami language system, only “li” is used to mark all types of plural. In this test, however, “ly” was added as an extra prefix in pursuit of a greater level of discrimination between language learners. A test-taker has to figure out the rule which determines in which case a particular prefix is suitable. There are, in total, eight questions. Four options are provided in each question. Here are two trial items from Section Two (although in the actual test more items were needed to show the range of the prefix ‘li’).

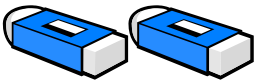
	
fafimi	lyfafimi
	

	
kobaybegi	lykobaybegi
	

tedami	lytedami	cunbegi	licunbegi

1. A. lykawcapi
   B. lykawcapo
   C. likawcapi
   D. likawapi

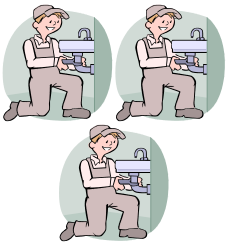
kawcapi



?

2. A. lybobegi
   B. lybegi
   C. libobegi
   D. libegi

bobegi



?

Section Three

In this section, examples of some Latejami sentences are presented in a simple word order, verb preceding the subject and the subject preceding the object. From the examples, a test-taker needs to figure out this pattern and choose the right sentence describing the picture in each question.

There are, in total, five questions in this section. To avoid the extra burden of new vocabulary, the nouns in the trial items were carefully selected and were presented before the testing part. Here is one of the trial items:

1. A. Cakopa byetimi bobegi.
   B. Byetimi cakopa bobegi.
   C. Bobegi byetimi cakopa.
   D. Cakopa bobegi byetimi.

Section Four

After activating a lemma and marking its number, a phrasal procedure is necessary in order to achieve noun phrase agreement. In this section, a test-taker needs to learn how to apply either a definite article (domo) or an indefinite article (tomo) to the noun to form a grammatically correct noun phrase. In the original Latejami language system, all nouns are definite by default, so “domo” does not exist, but having only one form was problematic for test-takers – hence the inclusion of ‘domo’.

In order to find the correct answer, a test-taker does not only need to choose the article correctly, but also to mark the number of the noun correctly. If a plural noun is needed, a decision about how to change the noun into its plural form has to be made. Knowledge learnt from Section Two is needed in order to make the right decision. Five options were provided in order to discriminate better. There are ten questions in total. Here are two trial items from Section Four.

		
Name	Laryabegi	Lamaybegi & Lapetebegi
Occupation	tekabegi	lykokwabegi

1. Kava Laryabegi _____.
   A. tekabegi domo
   B. lytekabegi domo
   C. tekabegi tomo
   D. lytekabegi tomo
   E. tekabegi

2. Ximunza tekabegi domo 2 _____.
   A. likokwabegi domo
   B. likokwabegi tomo
   C. lykowabegi domo
   D. lykowabegi tomo
   E. lykowabegi

Section Five

Section Five was designed to test the S-procedure, which requires the processing of inter-phrasal information, as suggested by Processability Theory. After completing the noun phrase, a language user needs to make some decision on the verb form which agrees with the noun phrase.

In the version of Latejami in this test, the verb form changes in accordance to features of the subject noun phrase, keeping the tense constant in the present. Regardless of the person, all the plural noun phrases (we, they, you) take the same root. The singular first person (I), the singular second person (you), the singular human third person (he, she) and the singular non-human third person (it) all take different roots. These features are also different from the original version of Latejami, but such changes were made, once again, to achieve better discrimination.

A test-taker needs to figure out the formation pattern based on the examples provided. A verb with its original form is provided and a test-taker needs to change it correctly according to the rule acquired after learning from the examples. There are six questions in this section. Here is one trial item.

1. _____ (fikicala) kokwabegi tomo.

Section Six

After acquiring the basic Latejami word and sentence formation rules, a test-taker is supposed to be able to generate a simple sentence. This section aims at testing if a test-taker can form a sentence while paying attention to the noun agreement, subject verb agreement and the sentence order.

There are four questions in this section. In each question, the noun(s) and verb necessary are provided. A test-taker is instructed to form a sentence using these words in reference to the picture in each question. Here is one of the items from this section.

1. (Cunbegi, Bucala)

"_____."

Trialling of the Test: General Considerations

Research Aims

The present article is a report on the first stage of a longer research programme. Given the discussion of the development of the test, the first challenge has been to establish that the theoretical base can be grounded in some empirical data. To that end, the test was used to explore whether:

- there is some sort of progression in difficulty across the test as a whole
- each section demonstrates a progression in difficulty, with the possibility that the final items in one section are slightly more challenging than the initial items of the next section

For the present, we are not concerned with any other forms of validity, and so in this study, we will not be correlating the test performance with foreign language learning achievement, or with other aptitude sub-tests.

Preliminary Work

Pienemann's approach aims at characterising development over extended periods, i.e. months and years. This test, in contrast, is of thirty-to-forty minutes duration, and so requires an intensity of learning that is not typical of what happens over longer periods. While this allows for efficient sampling, it does present challenges in test construction – ensuring that the test is neither too difficult nor too easy while maintaining its ability to discriminate between different levels of performance and ensuring its construct validity. Fundamental testing problems, obviously, had to be solved in a satisfactory manner. Examples had to be clear, visuals had to be unambiguous, and the test procedure had to be accessible to those who took the test. These problems were real, but soluble in a fairly clear manner with careful work. Far more difficult was the problem of achieving the right level of difficulty, with associated

effective discrimination. One needs a range of performance, with some test-takers doing reasonably well, and others not so well. Our approach was to engage in rounds of trialling, first with very small numbers of participants in early cycles, followed by a larger group, described below, for the main data collection.

Each preliminary cycle was followed by considerable debriefing of test-takers to explore what difficulties they had encountered. The typical outcome, at the early stages was to produce a test which was too difficult. This had the effect of generating relatively low scores which did not discriminate particularly well. It also had the problem that it produced a test which provoked some degree of irritation, as test-takers had to wrestle with extremely searching items. Several revisions of the test were required at this developmental stage, with each revision aimed at removing ambiguities or misunderstandings in the presentation materials, and also making the sub-tests from the different stages slightly easier. The results reported here, as indicated earlier, are based on the final version of the test, after completion of earlier developmental cycles.

Participants

Nineteen adults (fourteen males and five females) participated in the revised version of the test, ranging in age from nineteen to thirty-five years old. They had all at least completed high school. Three of them were studying in university at time of testing. Ten of them had obtained B.A. degrees while three had master's degrees. One of them had a doctoral qualification.

The language background of the participants varied. Nine of them had Cantonese as their mother-tongue. All of these were born and raised in Hong Kong except one who grew up in Guangzhou, China. The L1 of the other participants are all Germanic languages (English, Swedish, and German) and they all grew up in countries where their L1 is the official language there. All the participants speak at least one additional language and one of the participants had learned six additional languages.

Results

The basic statistics obtained when the final version of the test was trialled are shown in Table 2.

Table 2

Time and Score Statistics for the Test

Section of Test	Time (mins)	Score
Part One	4.84	7.65
	(1.61)	(2.00)
Part Two	5.16	5.47
	(2.16)	(1.81)
Part Three	3.61	4.65
	(1.93)	(1.00)
Part Four	9.84	8.53
	(3.20)	(2.48)

Part Five	8.63	2.18
	(4.91)	(2.38)
Part Six	7.75	6.24
	(4.16)	(4.29)
Total	39.83	34.71
	(11.40)	(9.75)

**Standard deviations are shown in parentheses.*

These figures are the raw figures, and take no account of the number of items in each section of the test. Even so, they do provide some insights. It is clear that significantly more time is spent on Parts Four, Five, and Six. In addition, the standard deviations show that there is reasonable discrimination between the participants. Finally, the percentage scores themselves do not have any clear progression in level. Parts Three and Five seem to generate particularly low scores. The issue of the number of items is very important, as the data based on the number of items in each section can provide more meaningful insights, as shown in Table 3.

Table 3

Time and Score Statistics corrected for No. of Items

Section of Test	Time per Item	Percentage Score
Part One	0.54	85.0
	(0.18)	(22.2)
Part Two	0.65	68.4
	(0.27)	(22.6)
Part Three	0.72	92.9
	(0.39)	(19.9)
Part Four	0.62	53.3
	(0.20)	(15.5)
Part Five	1.44	36.3
	(0.82)	(39.6)
Part Six	0.46	36.7
	(0.25)	(25.3)
Total	0.65	57.0
	(0.19)	(0.16)

Table 3 shows more clearly that the test is progressive, but not in a straightforward linear way. There are two patterns in play. Regarding time, one could generalise, tentatively, and say that for five of the levels, (Parts One to Four, and Part Six), values are not hugely different from the mean of these five scores, 0.60. Typically, that is, participants take a little over half a second per item, and this is not particularly affected by the progression through levels. But then we have the considerable contrast with Part Five, with a mean item time of 1.44 seconds, more than twice the sort of times found elsewhere. Syntax: verb formation, seems to require significantly more time than the other levels. So we are dealing with one distinctly different performance, while otherwise, Level has little impact on time taken.

Regarding scores, there is a clear trend for later sections to be associated with lower scores, as evidenced by the scores for Parts One, Two, Four, Five, and Six. The clear exception to this is Part Three, which is the first syntax-based section and had the longest time per item except for Part Five. Parts One and Two were concerned with

words and morphology, with morphology focussed mostly on the categorisation of words following a system unlike most other languages. Parts One and Two do reflect an increase in difficulty, but it seems that participants found their first encounter with syntax relatively straightforward, with the emphasis in this section being on word order. Later syntax sections are significantly more difficult, and this is reflected in the lower scores. In fact, one could generalise here and say that the first three sections, which have a mean of around 82 seem clearly easier than the last three sections, with a mean of around 43.5.

Broadly, these results follow Pienemann's account of complexity, and suggest that a theory-driven account of difficulty within an inductive language learning test can be effective. The performance of the participants as the levels increase in complexity is reflected in the scores obtained (if not particularly in the time-per-item). This suggests considerable potential for efficient use of time. The results also suggest that, while the overall score could be the most valid overall index to deal with, there is considerable research potential in treating individual or subsets of the levels independently. Since each level generates a distinct score, there may be value in exploring the relationships between these individual scores and language learning achievement. In other words, they may be contributing unique variance which can be useful in prediction. One facet of this might be the potential the approach has to deal with particular levels of aptitude. Parts Four to Six are particularly demanding, and these sections may be specifically appropriate for measuring high-level aptitude (the target that motivated the development of Hi-LAB). Broadly, that is, there is scope for more fine-grained information to link particular test levels to different aptitude levels.

Discussion

In general, these results provide supportive validating evidence for the ILL test which was constructed. Most fundamentally, the test discriminates effectively. The average percentage score is around 56%, with a standard deviation of 16%. The scores range from 24% to 87%, suggesting that the test captures wide variations in achievement, even with a relatively talented group of language learners. In addition, the way in which mean scores generally decline as the test develops is gratifying. It suggests that the participants are having more difficulty, in general, as the test continues, and as higher levels of Pienemann's Processing Hierarchy are encountered. We can restate the hierarchy and relate it to the results, as in Table 4.

Participants cope with the sub-sections concerned with words, and also show a decrement in performance as the word tasks become more demanding. Then, with the arrival of syntax, they achieve clearly higher scores initially, but these decline, in a fairly regular manner. As syntax becomes more demanding, it leads to greater difficulty, and the scores for the last two stages of syntax, verb formation and sentence generation, are the lowest of all. Those participants who still score well at this stage have responded to the challenge extremely well. Roughly one-third of the participants, (bearing in mind that this was a fairly selective group), scored above fifty per cent in

Parts Five and Six. We can reasonably assume that this group are talented in their language learning abilities, at least as far as pattern identification is concerned.

Table 4

Processing Hierarchy and Difficulty Scores

Processing Hierarchy Stage	Percentage Score
Word: lemma and basic word structure	85.0
Word: category procedure, emphasising plurality	68.4
Syntax: simple word order	92.9
Syntax: noun agreement	53.3
Syntax: verb formation	36.3
Syntax: sentence generation	36.7

On the basis of the evidence so far, it seems likely that Pienemann's Processing Hierarchy provides a good basis for sampling the inductive language learning component of language aptitude. In fact, it redefines this component somewhat. Previously, a miscellaneous collection of puzzles, intuitively identified as progressively more complex, were used to structure the aptitude tests that were produced. On this occasion, the theory came first, and its instantiation in sub-groups of items came second. What is interesting is how predictive the theory has been regarding the difficulty level of the different stages in the processing hierarchy. Equally interesting is the split between word operations and syntax operations. Although word: categorisation is at a lower level than syntax: word order, the latter proved to be the easiest of all in the test that was developed. Word order, at least in a simple form, seems to be within the grasp of many learners. Syntax quickly became more complex, but it is interesting that in its first level, it proves quite accessible. As a result of these figures, we can now have more confidence that a second language acquisition theory is at least a promising approach to structuring an inductive language learning test.

The results also provide an interesting possibility for the future. Far and away the most dominating foreign language aptitude test is the MLAT (Carroll and Sapon 1957), which, as pointed out earlier, does not include an inductive language learning sub-test. The Pimsleur test did, but this was targeted at high-school students, as was the York Language Aptitude test. The Defense Language Aptitude Test (Petersen and Al-Haik 1976) was an interesting reaction to this work. Following Culhane (1970), these authors regarded the MLAT as ineffective at discriminating between high level language learners. The inductive language learning test that was included in the DLAB was therefore targeted at a higher level of learner. In the event, this targeting seemed to be rather limited in effectiveness and the DLAB did not generate validity coefficients any better than the MLAT, even with higher-level learners (Skehan 1989).

The findings reported in this chapter give some cause for optimism in the production of aptitude tests for different levels of foreign language achievement. As Table 4 suggests, there is a relationship between theoretical views of what is advanced, and

the empirical results regarding difficulty scores. The implication is clear. If one wants an easier test, one can sample from the earlier levels to produce such an easier test. If, in contrast, one wants a more difficult test, then Levels One to Three or even Four might be weighted less strongly, and a larger weighting of the items in the test could come from Levels Five or Six. In this way, it may be possible to produce not so much a specific aptitude test, but rather a methodology for the production of such tests, which could then be adapted or calibrated to deal with local circumstances. This has never been done with aptitude tests in the past and it is an exciting prospect that follows from the research reported here.

But of course, a lot remains to be done. What we have reported on is the development of a test of inductive language learning ability following certain principles. What is needed next is a validation study which explores how this test can predict second language learning success. Only with this evidence can we see whether the promise of using second language acquisition theory is realised in actual language performance. This will need to be the focus of future research.

Conclusions

A first point to consider is what has been achieved with this research. Essentially, all we can claim is proof of concept. The research may not be extensive, and is indeed quite narrow in nature. However, it has taken a theoretical starting point, shown how an inductive language learning ability can be assessed in a manner consistent with the starting point, and reported performance data consistent with what the ILL theory would predict. This contrasts with previous approaches to the construction of tests, and in itself, opens up a new avenue of aptitude research. The current research is hardly comprehensive, but it is very encouraging.

Another achievement, following from this, is the demonstration that Pienemann's levels of development have a sufficiently clear relationship with the reported data. In particular, this suggests possibilities for a multi-level approach to measuring ILL. In other words, while a general score on the test as a whole has the potential to be very informative, one can also think in terms of more detailed scores for the individual levels. Bearing in mind that each level functions within a cumulative system, this could mean that diagnostic information is potentially available, even for different levels of aptitude. In other words, it may be possible to modify the test, or adapt scoring procedures, to make the test more effective for different levels of ability - the sort of task that Hi-LAB attempted. The result could be better discrimination between participants at the levels where such discrimination is desired.

The approach is also consistent with Skehan's Stages account of aptitude, which proposes a sequence of stages, from auditory processing, through language engagement, to automatisation. Skehan (2019) proposes sub-processes of noticing, pattern identification, generalising, and complexifying. The current test complements this account of aptitude. It even holds out the possibility that later stages of

automatisation might be tapped very slightly if such learning speeds up test performance.

It is important to say, though, that we are dealing here with work in progress. Establishing proof of concept is a crucial first step, and while the initial results are promising, they are not sufficient on their own. What is needed now is more extensive research to expand this beginning. Two obvious lines of this research will be to explore the functioning of our ILL test with (a) other aptitude sub-tests and (b) with actual language learning performance. Regarding the former, it is important to establish relationships with established tests such as MLAT 4, Words in Sentences, as well as with alternative ILL tests such as LLAMA F. There may also be merit in exploring relationships with implicit learning tests, as well as memory tests, including working memory, more broadly (Chan et al. 2011). Some overlap would be expected here, but a great deal of discriminant validity could also emerge (Skehan, 2023a). Equally important is the need to explore predictive validity in actual learning contexts. A diverse range of ages, proficiency levels, and L1-L2 combinations is needed. A particular need is to explore test effectiveness with different methodologies (more conventional and teacher-centred compared to communicative and task-based), as well as contexts where naturalistic learning is possible. A test such as the one reported on here could also be very effective in the sort of micro studies which have grown in number in recent years (Skehan, 2015, 2023b). In principle, an ILL test should be appropriate for wider, acquisitional contexts, as well as in classroom learning.

There are also other interesting lines of research to pursue. Our work with the test took a power perspective - there were no time constraints on the participants. But it could of course be interesting to introduce a speed element, partly to explore people's capacity to learn under a little pressure, partly perhaps to increase discrimination, and definitely to make the test more efficient, in relation to time spent. It would also be useful to conduct more qualitative studies with the test. We did monitor performance at the piloting stage, and explore what participants found most difficult. However, a more extensive qualitative study using the current version of the test could provide valuable insights into participants' interpretations, the aspects they find most difficult, and possibly, how the test could be modified to enhance its effectiveness. Such data might also give insights into the relationship between implicit and explicit processes for people doing the test.

ORCID

 https://orcid.org/0000-0003-1429-8859
 https://orcid.org/0009-0007-2277-6802

Acknowledgements
Not applicable.
Funding
Not applicable.

Ethics Declarations

Competing Interests

No, there are no conflicting interests.

Rights and Permissions

Open Access

www.EUROKD.com

References

Carroll J.B. (1962). The prediction of success in intensive foreign language training, In R.Glaser (Ed.), *Training, research, and education*. University of Pittsburgh Press.

Carroll J.B. (1973). Implications of aptitude test research and psycholinguistic theory for foreign language teaching. *International Journal of Psycholinguistics*, *2*, 5-14.

Carroll, J.B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Carroll J.B. & Sapon S. (1957). *The Modern Languages Aptitude Test*. Psychological Corporation.

Chan, E., Skehan, P., & Gong, G. (2011). Working memory, phonemic coding ability, and foreign language aptitude: Potential for the construction of specific language aptitude tests – the case of Cantonese. *Ilho do Desterro: A Journal of English Language, Literatures, and Cultural Studies, 60*(1), 45-73.

Cronbach L. (1970). *Essentials of psychological testing* (3rd Edition). Harper and Row.

Culhane T. (1970). *University of Essex Language Centre Occasional Paper No. 7.* Colchester, England: University of Essex Language Centre.

Green P. (1975). Aptitude testing: an ongoing experiment, *Audio-Visual Language Journal*, 12, 205-210.

Grigorenko E., Sternberg R.J., and Ehrman M. (2000). A theory based approach to the measurement of foreign language learning ability: The Canal-F theory and test. *The Modern Language Journal*, 84, 390-405. https://psycnet.apa.org/doi/10.1111/0026-7902.00076

Latejami: www.eskimo.com/ram/latejami.

Meara, P. (2005). *LLAMA Language Aptitude Tests*. Lognostics.

Morneau R. (1995/2006). *Lexical semantics*. Available at: www.eskimo.com/ram/lexical semantics.

Pan, J. (2023). *Developing and validating an internet-based battery of Tests of Aptitude for Language Learning (TALL).* Unpublished Ph.D. dissertation, University of York.

Petersen, C. R., & Al-Haik, A. R. (1976). The Development of the Defense Language Aptitude Battery (Dlab. *Educational and Psychological Measurement*, *36*(2), 369-380. https://doi.org/10.1177/001316447603600216

Pimsleur P. (1966). *The Pimsleur Language Aptitude Battery*. Harcourt, Brace, Jovanovic.

Pimsleur, P. (1968). Language aptitude testing. In A.Davies (Ed.). *Language Testing Symposium: A Psycholinguistic Perspective* (pp. 98-106). Oxford University Press.

Pienemann M. (1998). *Language processing and second language development: Processability theory*. John Benjamins.

Pienemann M. (2005). *Cross-linguistic aspects of processability theory*. John Benjamins.

Skehan P. (1989). *Individual differences in second language learning*. Arnold.

Skehan, P. (2015). Foreign language aptitude and its relationship with grammar: A critical overview. *Applied Linguistics, 36*(3), 367-384. https://doi.org/10.1093/applin/amu072

Skehan, P. (2019). Language aptitude implicates language and cognitive skills. In Z.Wen, P.Skehan, A. Biedron, S.Li, & R.Sparks (Eds.). *Language aptitude: Advancing theory, testing, research, and practice* (pp. 56-77). Routledge.

Skehan, P. (2023a). Testing language aptitude: A commentary on batteries and reanalysis of constructs. In Wen, Z., Skehan, P., & R. Sparks (Eds). *Language aptitude theory and practice* (pp. 208-245). Cambridge University Press.

Skehan, P. (2023b). Reflections on aptitude: Theory, research, and measurement. In Wen, Z., Skehan, P., & R. Sparks (Eds.). *Language aptitude theory and practice*. (pp.443-467). Cambridge University Press.

Wen, Z., Skehan, P., Sparks, R. (2023). *Language aptitude theory and practice*. Cambridge University Press.