Poznań Studies in Contemporary Linguistics 44(4), 2008, pp. 569–595

doi:10.2478/v10010-008-0027-6

MEASURING FOREIGN LANGUAGE LEARNING APTITUDE. POLISH ADAPTATION OF *THE MODERN LANGUAGE APTITUDE TEST* BY CARROLL AND SAPON

JACEK RYSIEWICZ
Adam Mickiewicz University, Poznań
rjacek@ifa.amu.edu.pl

ABSTRACT

This article sets itself two main aims. The first is to describe the rationale behind the decision to adapt for Polish learners the Modern Language Aptitude Test (MLAT) by Carroll and Sapon (1959), rather than to develop a new measure. The reasons behind the decision are discussed in the context of the relevant individual differences (ID) research in Poland and the need for a reliable and theoretically valid measure of foreign language (FL) aptitude for L1 Polish is articulated.

The other aim is to describe the development, piloting and initial validation of the Polish MLAT-based adaptation of a new measure of FL aptitude. Two methods of test adaptation (translation and paraphrase) are discussed and justified with relation to the current project. It was decided that all four components of FL aptitude, as proposed by Carroll (1981), would be represented in the Polish adaptation of the MLAT. The piloting was done on approximately 200 secondary school learners aged 19, while the data for the initial validation study, in the form of second language (L2) English proficiency test results as well as simple measures of motivation, length of study, social background and others, came from ca. 250 subjects, aged 18–22.

KEYWORDS: Testing; test adaptation; aptitude; measurement; individual differences.

1. Introduction

Any piece of quantitative research faces the problem of an adequate description of the variables it sets out to investigate. In second language acquisition (SLA) research, foreign language learning aptitude has been one of the most frequently investigated individual difference (ID) variables and, as extensive research has shown, it is responsible for quite a portion of the variance in foreign language (FL) learning achievement and/or proficiency. This research would not be possible without valid and reliable measures of

FL aptitude. From a number of tools created to quantify "the knack for languages", two are widely known: The Modern Language Aptitude Test – MLAT (Carroll and Sapon 1959) and Pimsleur Language Aptitude Battery – PLAB (Pimsleur 1966), which were written for English. Since the time of MLAT publication, more or less successful efforts have been made to produce similar measures of FL aptitude for other languages (Japanese, Swedish, Hungarian). However, no such test has been created for Polish. The purpose of this study is to fill the need for a valid and reliable test of language aptitude for the L1 Polish population.

The overall aim of the study presented in this article is to develop, pilot and partially validate an FL aptitude test for Polish adult learners of foreign languages. In the first, theoretical part of the article, the key concepts involved in aptitude measurement are overviewed (Section 2); one operationalisation of the construct, i.e. the MLAT, is described (Section 3); the Polish context of aptitude research is presented (Section 4); and the problems and issues in mental test adaptation are considered with the aim of justifying the particular adaptation procedure adopted by the present author (Section 5). The second, empirical part of the article (Section 6) is a description of the methodology of the study. First, the rationale behind the type of the adaptation procedure adopted is briefly discussed (Section 6.1), and the stages of the adaptation and the characterization of the target population are described (Section 6.2). Then, a detailed presentation of the adapted tool is given, together with a brief description of the proficiency criterion measures (Section 6.3). Section 6.4 presents the results of the initial, predictive validation of the tool, which is followed by a discussion (Section 6.5). The article ends with concluding remarks where some implications for further study involving the newly developed test are considered.

2. Foreign language learning ability – basic conceptual issues

The key concepts which are invariably mentioned in connection with FL learning ability are "ability", "aptitude" and "achievement". Although, theoretically as well as operationally, there is no clear-cut difference between the three notions, it has been customary to distinguish between abilities which are available here and now, versus those which are potentially possible. Such a distinction is intuitively acceptable and has been argued for by, for example, Reber (1985), who claims that the term "ability" refers to "an individual's potential *to* perform", whereas "aptitude" describes "an individual's potential *for* performance". In other words, abilities are available **now** to an individual for task performance with no further training needed, while aptitudes characterise **potential** for achievement given appropriate instruction. Consequently, achievement tests are tests "designed to evaluate a person's current state of knowledge or skill (ability), whereas aptitude tests are designed to evaluate potentialities for achievement ideally independent of current knowledge" (Reber 1985: 6). This understanding of the aptitude–ability distinction stresses the dynamic relationship between the two and fully articu-

lates the idea of temporal priority of aptitudes in relation to abilities. What is more, by treating aptitude as a potential ability which develops into fully fledged ability with appropriate training, this view assumes psychological priority of aptitudes over abilities. However, this distinction is not as neat as we might hope because, as Reber remarks, "tests of aptitude are, in reality, tests of performance (ability) and interest. The distinction in usage comes from the notion of making prediction about future achievements" (1985: 50). The author suggests that the two are basically the same construct and that the only difference between aptitudes and abilities lies in the function they perform: a diagnostic function in the case of abilities; a prognostic function in the case of aptitudes.

Carroll, while being one of the most ardent proponents of the separateness of the two concepts, sees them as separate but still related notions. In his seminal survey of factor analytic studies, he has this to say on the aptitude–ability–achievement distinction: "I regard the term *ability* as entirely neutral and even uninformative as to whether any given ability is an *aptitude* or an *achievement*" (Carroll 1993: 16). He then goes on to say: "Here, I used the term *aptitude* in a relatively narrow sense, i.e., to refer to a cognitive ability that is possibly predictive of certain kinds of future learning success; I exclude the notion of aptitude as interest in and motivation for a particular activity" (Carroll 1993: 16). Apparently, Carroll understands the two terms as overlapping to a certain degree. For him, *aptitude* is *ability* used in order to predict future learning: "aptitude measures are always measures of some kind of achievement, because the responses of individuals taking these tests are dependent to some degree on past learning" (Carroll 1981: 85). Consequently, the two notions are logically distinguishable only in the context of a specific classroom situation, i.e. teaching methodology, pedagogical aims of the curriculum, specific achievement requirements, etc. Aptitude is defined by Carroll in cognitive, rather than in personal, affective or motivational terms. According to him, aptitude and achievement are clearly distinguishable on theoretical and logical grounds. On this he remarks: "I reserve the term achievement to refer to the degree of learning in some procedure **intended** to produce learning, such as a formal or informal course of instruction [...]. Tests designed primarily to measure such degree of learning are measures of achievement, in addition to being measures of ability" (Carroll 1993: 17).

Generally speaking, FL aptitude is understood to refer to a set of primary capacities, propensities at an individual's disposal, available to him prior to learning and, to a certain degree influencing his potential level of achievement. Carroll (1973, 1981) speaks of the time needed to achieve a given level of competence. Viewing FL aptitude as a function of time, the claim he is making is that practically anyone is predestined to learn an L2 provided the instruction is appropriate, the learner is motivated and able to profit from instruction.

An impressive volume of FL aptitude research that has accumulated since the increased interest in the IDs began in the early 1960s has resulted in a deepened and crystallised understanding of the structure of aptitude as well as its role in the process of FL learning/acquisition. The detailed accounts of the history of research and theorizing as

well as current research agendas in FL aptitude research are beyond the scope of this article and can be found in, for example, Skehan (1989, 1998, 2001), Spolsky (1995), Sasaki (1996), Sparks and Ganschow (2001), Dörnyei and Skehan (2003), Ellis (2004), Dörnyei (2005) or Robinson (2005, 2007). The findings of those, and many other, studies that have accumulated over the years point to the following facts about FL aptitude:

Foreign Language Learning Ability (FL aptitude) is

(1) an autonomous dimension independent of both, affective (anxiety, motivation, attitudes) as well as general cognitive factors;

(2) independent of academic ability or intelligence, although it partially overlaps with these domains;

(3) relatively stable over longer periods of time; not dependent on prior learning experience; not easily modifiable through training;

(4) not a single, unitary capacity but a composite of several relatively independent cognitive abilities (componential/multi-factor structure);

(5) always a better prognostic of L2 learning success than any other ID taken singly or in combination with each other (Gardner and Lambert 1965; Carroll 1981; Sparks and Ganschow 1991; Skehan 1992; Dörnyei and Skehan 2003; Dörnyei 2005).

The classical model of language aptitude (Carroll 1968; Carroll and Sapon 2002 [1959]) is a four-component model developed by Carroll and Sapon as a result of a series of factor analysis studies done on the whole gamut of variables operationalizing abilities that were thought to be important in FL learning (Stansfield and Reed 2004). Eventually, the following set of factor-analysed abilities were identified and interpreted:

– **Phonemic coding ability:** the ability to segment and identify (code) distinct foreign sounds, to form associations between them and graphemic symbols representing them for later use.

– **Grammatical sensitivity:** the ability to identify grammatical functions of words/phrases in sentences.

– **Inductive language learning ability:** the ability to infer linguistic rules and patterns from limited linguistic evidence for example, a fictitious language.

– **Rote learning ability:** the ability to form associative links between form and meaning of language material presented visually, retain the links and recall the remembered meanings (Skehan 1989).

Carroll's four-factor model of FL aptitude was an empirically established model and it served as the basis for the construction of the Modern Language Aptitude Test.

3. The MLAT – a test of language learning ability

The four components of Carroll's model of FL aptitude were operationalised in five specific performance tasks in such a way that the postulated construct displayed itself best in the nature of the abilities called for to complete it. As a result of this test-writing process, which meant turning the theoretical construct into its concrete ability-based form, the following tasks made up the final MLAT test battery:

Part I. "Number learning".
In this part of the test, a test-taker learns a numerical system from a fictitious language which includes the numbers from 1 to 4, the "tens", and the "hundreds". Then, he is asked to perform a task of recognizing numbers which are presented to him aurally as combinations of the items just learned, e.g. 142, 32, 13. The posited construct tapped onto by this task is the rote learning and inductive learning ability. Apart from that, the phonetic coding ability is implicated in the task, in the formation of the link between the numbers presented through sounds and their graphemic representations.

Part II. "Phonetic script".
This test requires the examinee to form a link between symbols (graphemes) and English sounds and to retain those associations throughout a word-recognition task in which he chooses an appropriate grapheme representation of a stimulus from four options provided. The test is an operationalisation of the phonemic coding ability.

Part III. "Spelling clues".
In this test, the task is to recognize a word's meaning by selecting the meaning which is the closest to the meaning of a stimulus word despite that word's distorted (phonetic) and/or incomplete spelling as in the sample item below:

ple 'plea' (A. gift; B. meadow; C. tree; D. appeal; E. circumstance)

The success in this task depends on both phonetic coding abilities as well as on the knowledge and/or scope of L1 vocabulary.

Part IV. "Words in sentences".
Here, the examinee is required to select one word from the five underlined words or group of words in a test item sentence, such that matches the grammatical function of a word or group of words in a stimulus sentence as in a sample item below:

A CALCULATING machine is useful to mathematicians.

Skiing is a fine sport during the winter months./ Seeing is believing.
A **B** C D E

This part of the MLAT measures grammatical sensitivity, i.e. the ability to identify the grammatical functions of words in a sentence without reference to overt grammatical knowledge and/or terminology.

Part V. "Paired associates".
The last test in the battery involves an examinee in a word-learning task in a new language, whereby he is required to memorize a set of 24 words and their L1 equivalents in a limited time, practise them in a follow-up, short exercise stage, and then perform a word recall task. This part tests rote learning abilities.

The five tasks that entered the battery were selected from a long list of tasks used in trial versions. The overriding inclusion criterion was the psychometrically determined functioning of each task in the battery, such that would minimize the relationship among the individual tasks but maximize each task's role in the joint score of the whole battery. As a result of this selection, a battery of five tasks was produced (cf. above) in which the individual tasks (tests) show a relative degree of independence from each other (correlation coefficients in the range from .20 to .40) but a considerable degree of the relationship between themselves and the total score of the battery (*r*'s from .50 to .80). This pattern of the correlations is to be understood as an indication that the abilities tapped onto by the five parts of the MLAT, while sharing some characteristics typical of all cognitive tasks, are generally independent of each other thus capturing different facets of the aptitude construct.

The MLAT is a timed test which means that time limits are set for each sub-test of the battery. In this sense, the MLAT in not a "power test", whereby test takers are given no time limits on the task completion and are left to decide when they want to finish it. However, not all the sub-tests have the characteristics of a speeded test where time allocated for task performance forms a part of the construct description. For example, in the case of "Number learning" and "Phonetic script", ample time is given for the completion of the tasks; however they are tape- or CD-administered, which ensures that all test takers begin and finish at the same time. On the other hand, the "Spelling clues" and the "Paired associates" tasks are speeded, and time restrictions necessarily constitute a factor in the description of the constructs tapped onto by those tasks.

Even a cursory analysis of the MLAT reveals the following two problems with how the battery operationalises the four-component model by Carroll:
– not all the components of the four-factor model are represented in the tasks;
– some of the components get a better coverage in the MLAT than others.

Of the four abilities from the Carroll model, phonetic coding, grammatical sensitivity and rote are represented well in parts 2, 4 and 5 respectively, while phonetic coding, rote and inductive learning are represented weakly in parts 3 and 1 respectively. In other words, inductive learning receives only marginal representation in the "Number learning" task, grammatical sensitivity is represented well in one task, while the other two

components, phonetic coding and rote, are represented twice in two tasks – phonetic coding in "Phonetic script" (well) and "Spelling clues" (marginally); rote in "Paired associates" (well) and "Number learning" (marginally). Table 1 below summarizes the problem.

Table 1. How aptitude components are represented in the tests of the batteries.

Component	Well	Weakly
phonetic coding	MLAT 2	MLAT 3
grammatical sensitivity	MLAT 4	0
rote learning	MLAT 5	MLAT 1
inductive learning	0	MLAT 1

Before the implications of the incompatibility between Carroll's model and its MLAT opeartionalization for the actual task of adapting the test for Polish FL adult learners are discussed in the section below, two more proposals of FL aptitude tests will be briefly discussed.[1]

The first of those is Pimsleur Language Aptitude Battery – PLAB (Pimsleur 1968), in which the author takes a roughly similar approach to that of Carroll and Sapon's, although Pimsleur's is less broad than theirs in that he basically regards FL aptitude as being composed of a language analytic ability and an auditory ability. By including a question on the student's interest (Part 1) and by recording the student's grades in English, Mathematics, History and Science (Part 2), Pimsleur seems to be postulating that a cognitive factor of academic achievement and a personality factor of motivation/attitude are parts of FL aptitude. The remaining parts of PLAB are:

– Vocabulary (Part 3) – a knowledge component: a test of native language vocabulary.
– Language Analysis (Part 4) – inductive language learning ability (rule inference).
– Sound Discrimination (Part 5) – an ability to attend to a new foreign language auditory material.
– Sound-Symbol Association (Part 6) – a phonetic coding ability task.

On Pimsleur's view then, FL aptitude is primarily defined in terms of cognitive abilities of sound identification, coding of meaning and inductive rule acquisition; however, he also includes the knowledge of L1 lexis (verbal intelligence) in his understanding of FL learning ability, a domain which is only implicitly present and weakly represented in

[1] For a comprehensive discussion of other aptitude tests see, for example, Dörnyei (2005) and the references there.

MLAT ("Spelling clues"). Unlike any other conceptualization of the construct, Pimsleur includes motivation as a component of FL aptitude. This is contrary to a generally held belief that the two are rather separate learner factors. But, despite Pimsleur's assertion that he conceives of FL aptitude as consisting of verbal intelligence, auditory ability and motivation for learning, it is to be doubted if the few-item test of interest designed to elicit data on a learner's motivation is capable of yielding any reliable information in this respect. What is surprising, Pimsleur himself seems to exclude his motivation factor from the discussion of the uses of his battery. For example, in Pimsleur (1968: 104), he illustrates the diagnostic (and indirectly prognostic) value of his "empirical theory" on the example of four cases of learners. Each learner profile has a differential level of scores obtained in the four parts of PLAB: Vocabulary, Language Analysis, Sound Discrimination and Sound-Symbol Association. No mention is made of the motivational factor. The two tests (MLAT and PLAB) have been the most frequently used tools in research and educational prognosis. Their predictive power expressed in multiple correlations with achievement measures under intensive training conditions with homogenous groups has stabilised between 0.40 and 0.65.

Grigorienko, Sternberg and Ehrman (2000) developed an entirely new test of FL aptitude based on a particular cognitive theory of FL acquisition the so called CANAL-F Theory (**C**ognitive **A**bility for the **N**ovelty in **A**cquisition of **L**anguage – **F**oreign). The main tenet of the theory, as based on acquisition processes rather than on products of these processes, is that "one of the central abilities required for FL acquisition is the ability to cope with novelty and ambiguity" (Grigorenko et al. 2000: 392). The authors propose the following components of their CANAL-F Theory: (1) Knowledge Acquisition Processes; (2) Levels of Processing; (3) Modes of Input; (4) Memory Processes.

The processes that the learner makes use of during knowledge acquisition include the following:

- Selective encoding – to distinguish between the relevant and irrelevant information in the stream of incoming data.
- Accidental encoding – to encode less important, secondary information outside the learner's conscious focus.
- Selective comparison – to determine the relevance of old information for the needs of a current task.
- Selective transfer – to apply decoded/inferred rules to new contexts/tasks (analysis and induction).
- Selective combination – to synthesise new information with old.

In FL learning those five processes operate at four levels of processing: lexical, morphological, semantic and syntactic (phonological? – JR), in the visual or oral modes of input and rely on immediate and delayed memory processes.

The CANAL-F aptitude test was created to address and gauge the above mentioned four sources of individual differences. According to the authors, the test has at least two

features that make it unique when compared to other traditional tests of FL aptitude. The test is designed to simulate a continuous learning situation and it taps the processes of knowledge acquisition at the time of the test. A comprehensive description of the test would go beyond the scope of the present brief review (cf. Grigorenko et al. 2000 for examples of tasks).

The CANAL-F test was construct-validated against the following measures: Cattell's Culture Fair Test of *g* (Cattell and Cattell 1973); Terman's Concept Mastery Test (Terman 1970); MLAT and prior language experience questionnaire. The test's construct validation yielded two broad first-order factors: an intelligence-related factor and a language-specific factor. While all the CANAL-F subtests loaded equally highly on the two factors, only two of MLAT's subtests ("Paired associates" and "Spelling clues") loaded on the language specific factor. This is to be interpreted as construct homogeneity of the CANAL-F test.

The study of the criterion-related validity of the CANAL-F test against the subjects' (N= 63) instructors' ratings of relevant language skills did not produce outstanding improvement in the magnitude of correlation coefficients: r=0.40. However, Ellis (2004: 533) remarks that although the correlations of CANAL-F test scores with measures of language learning are much the same as those reported for the MLAT, "the test does afford the possibility of achieving a closer match between specific aptitudes and specific psycholinguistic processes".

4. FL aptitude – the Polish context

It is because of the reasons discussed above that the author of the present paper made a decision to adapt for the Polish FL adult learners the most widely used and known measure of FL learning ability, The Modern Language Aptitude Test (MLAT) by Carroll and Sapon (1959). For almost fifty years since its publication, the MLAT has been cited in at least a thousand articles published in high profile professional journals abroad such as *The Modern Language Journal*, *Studies in Second Language Acquisition*, *Language Testing*, *Second Language Research*, *Journal of Applied Linguistics* and the like. It has been used for research purposes in numerous studies as a data collection tool, and has been adopted as the most reliable and valid diagnostic instrument by many educational and research institutions for selection purposes. It is this astounding record that makes the MLAT the most desired model on which to base a language specific adaptation of an instrument aimed at gauging the construct of FL learning ability.

The lack of a reliable, valid, standardised and workable test of FL aptitude is, to the current author's mind, responsible for the faint interest of Polish applied linguists in this aspect of learner individual differences (IDs). The survey of publications devoted to learner IDs in Poland clearly shows that this aspect of applied linguistics has been grossly underrepresented in the research. From the total of some 100 dissertations cited in The National Record of Academic Dissertations and Projects, which has been re-

cording dissertations since 1999 online, only about 10 are devoted to the study of IDs – of those the majority deal with affective IDs such as motivation, anxiety, beliefs and personality factors. Very few investigate the role of cognitive factors in foreign/second language (FL/SL) learning. A similar profile of interest emerges from the survey of papers read at national conferences in Poland. The programmes of three annual conference events (since 1988) contain over 2000 papers read at the International Conference on Foreign/Second Language Acquisition, Szczyrk (since 1988), at the conference of the Modern Language Association of Poland (Polskie Towarzystwo Neofilolgiczne, PTN) and the Polish Association for the Studies of English (PASE, since 1995), of which only about 20 were on IDs, including 17 on affective factors. The same ratio could be observed in professional journals published nationwide or by Polish universities. Similarly, for the last thirty-five years, very few publications devoted to the study of the role of IDs in FL learning have appeared in print. The better known works are: Ozga and Tabakowska (1973), Tabakowska and Ozga (1975), Lewowicki (1975), Niżegorodcew (1975, 1979), and Figarski (1984). In light of the above, the present author hopes that making the Polish adaptation of the aptitude test freely available to researchers will contribute to validity of primary-data generating quantitative research in the country.

5. Issues in mental test adaptation

Adaptation of mental tests from one culture and/or language to another has been a long established practice in educational, psychological and applied linguistics testing and/or research. This trend is going to intensify as there's "considerable evidence to suggest that the need for multilanguage versions of achievement, aptitude, and personality tests, and surveys is growing" (Hambleton 2006: 3). The same author anticipates that "substantially more test adaptations can be expected in the future as a) international exchanges of tests become more common, b) more exams are used to provide international credentials, and c) interest in cross-cultural research grows".

In considering which option to choose – adapt or develop a new test – several reasons for test adaptation are invariably mentioned in the literature. Hornowska and Paluchowski (2004) as well as Hambleton (2006) mention the following considerations in favour of test adaptation:

- test adaptation is very often simpler, faster and less expensive than the design and development of new tests, and saves invaluable time, effort and money especially when the test is to be used for research purposes rather than commercially;
- it is safer to use tests which have been extensively tried and which are well-known, and of established theoretical and practical validity;
- lack of local experts in the mental dimension measured;
- desire to relate research to international state-of-the-art in the measured dimension thus making international and intercultural comparisons possible.

Two of the above mentioned reasons, lack of expertise and international comparisons, are worth considering in some more detail in the context of FL learning ability whose measurement is of main interest to the present author.

First, it seems that the lack of expertise locally for developing a new tool for aptitude measurement should be a decisive factor in favour of the adaptation path. There are two aspects of this to be considered. One is related to the content matter i.e., the knowledge of relevant theories of FL learning ability and the processes implicated in it, and the other, to the "how" of test design procedures and construct operationalizations: writing items, piloting tests, item analysis and validation analysis. The lack of the former may result in reliable tests intending to measure FL aptitude but lacking construct validity, while the lack of the latter may yield theoretically valid tools, but of little psychometric value.

Another consequence of the lack of relevant expertise and research tradition in FL aptitude is that the construct may fall victim to non-professional attempts to create a tool to measure it – attempts which are based on the hunches and commonsense ideas of a prospective test designer, rather than on principled theories of the construct. This is especially true in the case of the language aptitude construct which appears appealing to the general public, language learners and teachers alike, and, at the same time, is felt to be intuitively simple and straightforward to anybody who has ever tried to learn a foreign language and succeeded in the task. Who else can be more predestined to make wise guesses about a "knack" for languages, than experienced teachers or practitioners and/or successful learners? Unfortunately, as good drivers don't necessarily make good driving instructors, so similarly, good language learners or experienced teachers don't necessarily make good aptitude test writers. Such is the case with two measuring instruments which have been published in this country over the last six years (Kuliniak 2002; Wojtowicz 2006),[2] and which make claims of being tests of language learning potential, but in fact are little more than tests of verbal intelligence.[3] Those, are invariably included in many tests of intelligence, such as Wechsler Adult Intelligence Scale-Revised (Wechsler 1981), Primary Mental Abilities Test (Thurstone 1938) or Otis-Lennon School Ability Test (Pearson 1995).

The other of the two reasons in favour of the test adaptation is the desire to relate research done in one country to mainstream theorizing and work in the field that has been done elsewhere. Without an equivalent or, at least, comparable, instrument which would be an operationalisation of a long-established and accepted theory of the construct of FL aptitude any international and/or intercultural comparisons are not possible. When there is weak or non-existent local tradition of research in the domain of lan-

[2] For a critical appraisal of Kuliniak (2002) and Wojtowicz (2006), see Rybacka (2008).

[3] Verbal intelligence is understood here as reasoning ability, word fluency and/or verbal comprehension operationalised in language-based tasks involving the processing of aural and visual language input. The understanding of the relationships between language concepts (analogies, comparisons, antonyms, synonyms, contextual meaning etc.) also constitutes the construct of verbal intelligence.

guage aptitude, it is wisest to use a tool which has received a lot of trialling and which functions as a benchmark test worldwide, rather than to attempt writing *ad hoc* tools of doubtful validity and of questionable theoretical background. Otherwise, that is when theoretically non-equivalent instruments are used, or worse, when the instruments used to investigate FL aptitude have no construct validity (do not measure what they hope to be measuring), it is difficult to maintain any degree of methodological rigour in researching SLA. Such tests used for research purposes mean in effect the loss of time and effort, and any results arrived at with their help must be seriously doubted.

In short, the use of aptitude tests which are not equivalent to like tests used internationally will well prevent any comparisons with work done in this area worldwide. In this connection, Hornowska and Paluchowski (2004: 190) remark: "If our research results are to be generalizable from the country-specific context then a test must be used which will enable us to compare individuals not only in the context of our country but also across countries". This can only be achieved through the use of culturally equivalent tests which are both, reliable and valid.

The requirements of test validity and test reliability are of paramount importance for the development of any mental test. A mental test is "a procedure designed to elicit certain behaviour from which one can make inferences about certain characteristics of an individual" (Carroll 1968). To do its job well, a test must display the following characteristics (Anastasi and Urbina 1997; Hornowska 2003):

(1) consistently define the levels of a measured dimension (be reliable);
(2) measure what we want it to measure and almost nothing else (be valid);
(3) the interpretation of its results cannot depend on who, where and when interprets them (objectivity);
(4) display uniform testing conditions (standardised);
(5) allow for the interpretations of the results across groups (age norms);
(6) be adapted to a receiving country's specific conditions.

Those characteristics can only be achieved if the following steps in test development and/or adaptation are rigidly observed:

(1) the analysis of the test's construct(s) is carefully carried out;
(2) translation of the materials of a test (instructions, stimuli, rubrics, etc.,) or other forms of adaptation;
(3) the piloting of the preliminary version of the test on a sample target population;
(4) equivalence analysis of the pilot version in the form of:
   – psychometric equivalence of: reliability, validity, facility indices, inter-item correlations, means, variances and factorial structure;
   – theoretical (construct) equivalence: achieved via factor analysis or multidimensional scaling;
(5) the construction of the final test version;

(6) testing with the final version;
(7) test validation analysis: confirming test validity against a criterion (prognostic validity), similar tools (convergent validity), or different tools (divergent validity).

When the test writing and/or test adaptation guidelines are not followed, either because of ignorance and lack of understanding of the demands of a quantitative research paradigm, or because of "sloppiness", scientific rigour is relaxed and vagueness, arbitrariness and lack of precision in construct definition sets in. Unfortunately, it is not uncommon that in actual research practice, those criteria are not adhered to, yielding instruments with doubtful construct validity and with little or no theoretical support. What is worse, a "sloppy" instrument, once badly written and published, or used in a research report, starts living the life of its own and gets quoted in other researchers' reports, or even is used as a measure of a construct it purports to tap onto in other research or adaptations. Paraphrasing Brzeziński (2002: 590), the following practice is not uncommon in L2 research:

– For the needs of his study a researcher "constructs" a test in the form of a collection of tasks which he proudly labels "a test of foreign language learning ability" and publishes the research based on it in a scholarly journal or presents it at a conference.
– Another researcher uses the instrument for data collection in his/her primary or replication research saying that the tool used is a Test of Foreign Language Learning Aptitude by a researcher X.
– Certain results are obtained, interpreted and pedagogical implications drawn. A question of their meaningfulness, however, is seriously to be doubted.

As a result of the above procedure, the pseudo-instrument acquires the status of a "valid" and "reliable" tool not because of its theoretical and psychometric characteristics, but merely because it has appeared in print. As Brzeziński (2002) further remarks, it is half of the problem when such a poor test is used for research purposes, as it is only the researcher who risks ridicule and loss of credibility. The real harm of "bad" instruments begins when they are used for diagnostic purposes which result in high stake decisions such as course or learning programmes selection criteria. The damage brought about by the decisions based on invalid and poor instruments is inestimable and difficult to predict.

6. Methodology of the adaptation

6.1. Initial decisions prior to the adaptation

Before the test adaptation procedure could begin, two general questions had to be answered. The answers to those questions were of paramount importance as they would

determine the shape, nature and future uses of the tool. In view of the apparent discrepancy between 'theory' and 'practice' a decision had to be made regarding the following issues:

(1) Should the adapted MLAT be a "replica" of the original test?
(2) Or, should all the components of the Carroll's model be represented in it?

In an article meant as guidelines for those wanting to adapt MLAT for other languages, Stansfield and Reed (2003: 7), enumerating five different adaptation strategies, consider the above mentioned options and remark on the first that "(it) is the least practical strategy, because it would be the most labour intensive and involve uncertainty regarding the effectiveness of adapted versions of the English-based tests". Additionally, this option, while replicating the original battery would also suffer from the same deficiency as the original, namely, a weak representation of the inductive language learning ability factor. Commenting on the other alternative, i.e., representing all four components of language aptitude, Stansfield and Reed (2003: 6) say the following:

> An audio recording would have to be done for MLAT 2 ('Phonetic script'), and [...], a special adaptation to the non-English language would have to be created for MLAT 4 ('Words in sentences'). Otherwise, this combination looks to be feasible and potentially very effective for prediction purposes and for diagnostic purposes.

Choosing this strategy would also mean writing a new task tapping onto the inductive learning component of aptitude, as this was missing from the original test.

The author of the current adaptation decided on a "hybrid" strategy, and in so doing he was advised by the following considerations:

– the desire to represent all four components of Carroll's model of FL aptitude;
– the desire to retain as much as possible from the original;
– the practicality criterion, meaning mainly testing time.

The joint effect of the three considerations was the compromise between full coverage of the aptitude components, task-construct overlap, and test length. In effect the compromise meant:

(1) creating a new task to operationalise inductive language learning ability;
(2) eliminating "Number learning", as it duplicated constructs measured by other tasks, but still leaving "Spelling clues" as it measured an "additional" dimension of L1 vocabulary span (indirectly);
(3) shortening the longer tasks by reducing the number of items in them ("Spelling clues" and "Words in sentences").

Finally, it was decided that five tasks (tests) would enter the Polish version of the battery, which was named the Test of Aptitude for the Learning of Foreign Languages (*Test Uzdolnień do Nauki Języków Obcych – TUNJO*). Four tasks were adapted from the MLAT, and one new task was created (see Table 2 below). The tasks taken from the MLAT were adapted either by recreating the construct of the original task using the Polish material ("Spelling clues" and "Words in sentences") or by translating the task's content and retaining the form of the original as close as possible, i.e. stimulus material, instructions and layout. This was done for the L1-free tasks i.e., for "Phonetic script" and "Paired associates". The two types of adaptation are referred to in the literature on mental test adaptation as paraphrase and transcription (Brzeziński 2002). The "Artificial language" task had to be written by the adaptation team,[4] as it was not included in the MLAT.

Table 2. MLAT – TUNJO: correspondence between tasks.

MLAT	TUNJO	Adaptation	Ability tapped
Phonetic Script	Phonetic Alphabet (*Alfabet fonetyczny*)	transcription	phonemic coding
Spelling Clues	Hidden Words (*Ukryte słowa*)	paraphrase	phonemic coding + L1
Words in Sentences	Words in Sentences (*Słowa w zdaniach*)	paraphrase	grammatical sensitivity
Paired Associates	New Words (*Nowe słowa*)	transcription	rote learning
not in MLAT	Artificial Language (*Sztuczny język*)	*new task*	inductive learning

6.2. Stages of the adaptation and the participants

The first stage in the process of MLAT's adaptation began in October 2006 with a detailed analysis of the constructs of the tasks chosen for the adaptation. The pilot versions of all five tasks were prepared and field tested on samples of the target population of young adults. Due to the organizational constraints that the study was subjected to,

[4] As it will be mentioned later in the article, this task is an original creation of the author's research assistants and has subsequently been rewritten and remodelled basing on the results of the pilot study.

each task (test) was piloted by a different team of assistant researchers on a different group of participants. Care was taken to make the groups as similar to each other as possible: subjects in all the five groups were 18 years old, attended the same type of a secondary school, and came from the same socio-economic background. The average number of subjects in the pilot group was thirty. All in all, 147 secondary school pupils participated in the pilot.

The second stage (January/February 2007) involved item analysis of all the tasks and the preparation of the final version of the test. Item facility indices (FI) and item discrimination indices (DI) were calculated for each item of each test. The indices were used as the basis for the selection of the best working items from the pilot-tested bank of items in those tasks which underwent paraphrase, or which had to be newly written (see Table 2 above). In the case of the remaining two tests, in which transcription was used as an adaptation method, FI and DI were used to check on the appropriateness of the tasks' items. The discussion of the final version of the tasks that made up the TUNJO test is given in Section 6.3. below.

The third stage (March till May 2007) was an initial validation of TUNJO. The aim here was twofold: to research the psychometric properties of the test, and to determine its validity. The first aim meant estimating the test's reliability and descriptive statistics, and the other – finding out about its predictive potential (criterion validity). To this end, TUNJO was administered to a sample of 250 subjects representing the target population of 18- to 20-year-olds in various types of secondary schools in Poznań, a city in western Poland of ca. 700,000 inhabitants.

The fourth stage of the adaptation project, aimed at norming the test on a stratified sample of the target population, i.e. providing the test with age or social background-related norms of achievement, is still in progress. This stage is necessary if TUNJO is to be used for diagnostic purposes by FL teachers or FL programme administrators.

The participants of the pilot and the proper study stages were all secondary school pupils learning their L2 as part of their school curriculum requirement. They represented different levels of L2 proficiency, roughly ranging from the pre-intermediate (CEF Level B1) to advanced (CEF Level C1) level. By the time they were tested, they had been learning their L2 for an average of eight years, counting only the formal, obligatory L2 courses at school. Since foreign language instruction in Polish schools starts at the age of eleven,[5] it is practically impossible to find subjects who have not had any L2 instruction by the time they reach the age which the Polish adaptation of MLAT targets. In this respect, the study is different from the original procedure followed by Carroll and Sapon, who tested their participants before they entered an L2 course. This fact might be responsible for the somewhat lower correlation coefficients arrived at in the present study due to the restriction of range of the scores for both the independent and dependent variables. This issue will be brought up in Sections 6.4. and 6.5.

[5] This corresponds to grade five, but it varies and some schools start earlier.

6.3. The instruments

6.3.1. TUNJO – the independent variable

As mentioned in Section 6.1, one of the considerations in adapting MLAT was the criterion of practicality in administering the test. What this meant in practice was that the test had to be "doable" within less than 60 minutes, and preferably, within 45 minutes, which is the standard length of a class period in Polish schools, a context where, it was assumed, most of the testing will be done. The time constraint, and the fact that a new inductive language learning task was included in TUNJO, meant that other tasks had to be made shorter. Because two tests were taken "across the board" from the MLAT using the transcription method (see Section 6.1), the reduction of the number of items could only be done for the tests that were adapted using the paraphrase method, i.e. "Hidden words" and "Words in sentences". As for the first test, the number of items was reduced from the original 50 items to 30 in TUNJO, whereas in the other test from the original 45 items to 26.

Similarly to the MLAT, the TUNJO tasks maintain the timing characteristics of the original. The time limits for the timing of individual tests adapted from the MLAT were established empirically during the pilot stage. The time limits that were set for the speeded tasks: *Ukryte słowa* ("Spelling clues" in MLAT) and *Nowe słowa* ("Paired associates" in MLAT) reflected the considerations of speediness of the original, as well as L1 Polish specific task demands. The *Alfabet fonetyczny* task ("Phonetic script" in MLAT) was timed from a CD and is described in more detail below. The task entitled *Słowa w zdaniach* ("Words in sentences" in MLAT) was timed but not speeded, which means that almost all the test takers had enough time to complete it. The *Sztuczny język* task (newly created "Artificial language") was allocated an ample time limit of 15 minutes (cf. Table 3 below). The whole test takes between 51 and 53 minutes to complete as compared to approx. 70 minutes needed to do the MLAT.

Below, a brief description of the changes introduced to the MLAT tasks during the process of adaptation is given. The individual tasks are described in the following order: first the "transcribed" tasks ("Phonetic script" *Alfabet fonetyczny* and "Paired associates" *Nowe słowa*), then the "paraphrased" tasks ("Spelling clues" *Ukryte słowa* and "Words in sentences" *Słowa w zdaniach*) and finally the task that was written for the purpose of the adaptation ("Artificial language" *Sztuczny język*).

There were two main changes that were introduced to the "Phonetic script" task. Firstly, the speed of delivery of the audio-recorded stimulus material was increased by shortening the pauses between individual one-syllable words both in the presentation as well as in testing modes. In the former, the four "words" were recorded at approximately one-second intervals between words (as compared to 1.5 seconds in the MLAT), whereas in the latter the intervals were approximately 2 seconds long (4 seconds in the source test). This shortened the task from 11 to 8.5 minutes, excluding the instructions to the test, which are also recorded. The shortening of the intervals in the exposure and

testing modes was done after the test had been piloted with the original speed of delivery maintained. However, it was found that the "longer" version did not discriminate learners well, as 95 percent of them scored from 90 to 100 percent of points on this version. By a similar consideration, a decision was made to use the original L1 English sound material from the MLAT rather than to replace it with Polish sounds. This decision, however, does not change the construct tapped onto by the test, as intended by Carroll and Sapon, which is the ability to form a link between a sound and a letter, but has no additional advantage of improving the test's discrimination index.

The stimulus material was recorded on a CD by the present author. Secondly, it was decided that the written symbols used in this sound-letter correspondence task had to be made to look more "strange" than they actually did in the MLAT. The original test uses the Kenyon and Knott transcription system for American English to code the sounds recorded, and it was anticipated that given the widespread knowledge of English in Poland, the "letters" might be too transparent. So a decision was made to replace some of the more common symbols, for example θ, ð with less common "shapes": ø and ɓ. The exemplar items are given below:

(1)	tɨk	tɣk	tɨs	tɣs
(2)	tɨs	tɣs	tɨz	tɣz
(3)	kɑs	kɨs	tɑs	tɨs

In the "Paired associate" task one minor change was introduced. The time allowed for the learning stage of the 24 pairs of words was shortened, from 2 minutes in the MLAT to 1.5 minute in TUNJO, for the practice stage it was lengthened by 0.5 min., and for the testing stage it was shortened from 4 to 3 minutes (cf. Table 3 below).

In the "Hidden words" task, despite the already mentioned fewer number of items used, different types of distortions to the spelling of the Polish words were made. Polish spelling is less opaque than English; it has fewer sound-letter alternations, has fewer vowel-sounds and, being predominantly consonantal, allows for heavy consonantal clustering in word initial and final positions. Generally speaking, there are two major sources of sound-letter inconsistencies in Polish. The first one is a phonetic process that neutralizes voiced obstruents word finally i.e., *drób* /drup/ 'poultry', and before or after another voiceless obstruent in a cluster as in: *dróżka* /druʃka/ 'little path'; *chwast* /hfast/ 'weed'. The other is the existence of double spellings for consonantal sounds: /ʒ/ – <rz>/<ż>, as in *rzeka* 'river' / *żołądź* 'acorn'; /x/ – <ch>/<h>, as in *chudy* 'slim' / *harcerz* 'scout', and /u/ – <u>/<ó>, as in *pukać* 'to knock' / *półka* 'shelf'. It was mainly those characteristics of the phonotactics of the Polish language and the sound-letter discrepancies that the task capitalised on. As a result of this, the distortions applied in the words included: removing all or most of the vowels; "phonetic spelling" and deliberate confusion introduced by incorrect spelling, and capital letters used as in the examples below:

OŻŁ [ORZEŁ] ‘eagle’
A. smutek ‘sadness’
B. zwariowany ‘crazy’
C. <u>ptak drapieżny</u> ‘bird of prey’
D. pensja żołnierza ‘(soldier’s) pay’

PRNT [PRĄD] ‘current’
A. emerytura ‘pension’
B. dłoń ‘palm’
C. pole bitwy ‘battlefield’
D. <u>elektryczność</u> ‘electricity’

The reader will be reminded that the test-taker’s task here is to guess the meaning of a “hidden word” by circling a word or a phrase which is closest to it in meaning. In the first version of the task, the number of options to choose from was five (as in the original), however in a later version it was reduced to four, in order to lessen the processing burden and speed up the test. Of thirty words included in the test, 18 were nouns (13 concrete and 5 abstract), 5 were verbs and 7 were adjectives. All the words, with the exception of two (*podarek* ‘gift’ and *chrobot* ‘scraping sound’) were high frequency, everyday words.

No major changes, were introduced to the “Words in sentences” task. The original test’s construct lent itself to a pretty straightforward adaptation to the use of Polish material. As in “Hidden words”, the first version of the “Words in sentences” task contained five response options; this was later limited to four for reasons of time. The stimulus material was selected in such a way as to reflect various functions of typical parts of speech-words such as nouns, verbs and adjectives appearing in various forms (aspect, gerund, participle) standing singly or in phrases. The distractors capitalised on the similarity of either position in a sentence and/or morphology (inflection or conjugation). The items were written by the research assistants and the answers were checked by linguist colleagues of the present author. A sample item is given below:

Po przegranej partii w brydża, **<u>mój przyjaciel</u>** bez słowa poszedł do domu.
‘After the lost game of bridge, my friend without a word went home.’

<u>Za</u> <u>domem</u>, blisko <u>lasu</u>, stała <u>stodoła</u>.
A B C **D**
‘Behind the house, close to the forest, stood a barn.’

The “Artificial language” task was conceived and written for the purpose of the adaptation to Polish. The stimulus language material was vaguely based on the language created by J.R.R. Tolkien in one of his fantasy novels and included a list of lexical items with Polish equivalents (verbs, nouns, adjectives, pronouns and prepositions) and a few phrases to illustrate the internal mechanism of the language. The language rules to be extracted from this material, though, were modified somehow to suit the purpose of the task. The test-taker’s task was to choose, from four options given, an appropriate Polish translation of ten stimulus sentences by analysing the language material available to him/her for the duration of the task, as in the example below:

Arenion a nagotas mue kamenont.

(a) Matka pracuje w szkole.	'The mother works in the school.'
(b) Matka pracowała w szkole.	'The mother worked in the school.'
(c) Matka pracuje w szkołach.	'The mother works in the schools.'
(d) <u>Matki pracują w szkołach.</u>	'The mothers work in the schools.'

In the later version of the task, ten items were added which engaged a test-taker into a more productive task of rule application whereby he/she was required to provide a partial translation of a sentence from Polish into the artificial language used in the task. This was done to improve the psychometric properties of the sub-test of the TUNJO battery. However, the results reported in the next section come from the first version of the battery.

Table 3. MLAT and TUNJO characteristics.

		MLAT	TUNJO
Phonetic Script	time	10 min. 50 sec.*	8 min. 30 sec.**
	no. of items	30	30
	response options	4	4
Spelling Clues	time	5 min.	5 min.
	no. of items	50	30
	time per item	6 sec.	10 sec.
	response options	5	4
Words in Sentences	time	15 min.	6 min. 30 sec.
	no. of items	45	23
	time per item	20 sec.	17 sec.
	response options	5	4
Paired Associates	time	2/2/4***	1.5/2.5/3***
	no. of items	24	24
	response options	5	5
Artificial Language	time	–	15 min.
	no. of items	–	20
	response options	–	–/3

* MLAT Phonetic Script – intervals between "words" in: (A) exposure = 1.5 sec; (B) test = 4 sec.

** TUNJO Phonetic alphabet – intervals between "words" in: (A) exposure = 1 sec; (B) test = 2 sec.

*** Times for: exposure, practice and test respectively.

6.3.2. L2 proficiency measures – the dependent variable

Two types of criterion measures were used for the purpose of the initial validation of the TUNJO test: language achievement and/or proficiency tests administered to the partici-

pants in the course of their study, and teachers' grades collected over a period prior to the administration of the aptitude test.

The achievement/proficiency measures were administered by the participants' respective teachers and were traditional "paper and pencil" tests of vocabulary, grammar, reading and listening comprehension. No standardized measures of L2 proficiency, of the TEFL or UCLES Cambridge sort were used. The tests were either classroom material based (achievement) or, in two or three groups, included final secondary school mock exam papers from previous years (proficiency). Additionally, grades from teachers' class-tests were used for all the participants. Those assessed similar language components to the ones mentioned above. The results of those test were expressed in percentages rather than in a 1-to-6 traditional grading scale. The tests were prepared, administered, proctored and scored by the class teachers and the results were made available to the researcher.

Because different grading scales were used in the various measures of L2 achievement/proficiency, standardized scores (z-scores) were obtained for each participant before they were analyzed statistically. L2 scores were only available for 195 out of 245 participants of the validation study at the time of the calculations. For various reasons, it was not possible to estimate the reliabilities of the L2 achievement measures due to the fact that no item-level data from the L2 tests were available to the researcher. Still, it has to be assumed that the L2 proficiency measures administered by the teachers to their respective groups of learners were appropriate with respect to task, and level-related difficulty, and that they were relatively reliable measures of the learners L2 achievement. For the purpose of this study, a general L2 proficiency "index score" was obtained for each participant, that is to say no specific language component variables were produced. This would have been difficult as no uniform L2 measure was used in all the groups of participants involved in the study. The limitations of this fact are briefly discussed in the next section.

6.4. The results of TUNJO's initial validation

The TUNJO battery's reliability was estimated using two methods. The first was a split-half method whereby two "separate" tests were created by splitting the battery in half using the odd-even principle: odd numbered items formed one test, and even numbered items formed the other. The reliability index is the correlation of the two halves corrected for length using the Spearman-Brown formula. In the case of tasks containing an odd number of items the last item is omitted from the calculations. The other method was Cronbach alpha which was applied to each sub-scale of the battery thus giving a reliability index for each test. As Cronbach alpha is sensitive to scale length; the reliability indices for tests containing few items usually underestimate the scale reliability. Also, this method is applicable only to homogeneous scales, i.e. tests in which the items

operationalise only one construct. Tables 4 and 5 below show the results of reliability estimates.

Table 4. Split-half reliability, N=245.

Split-half reliability: **0.89**	1st half	2nd half
Number of items	65	65
Mean	40.19	37.24
Std deviation	6.83	7.40
Variance	46.66	54.75

Table 5. Cronbach alpha, N=245.

TUNJO part	PhS	HW	WinS	NW	AL
Cronbach's alpha	0.69	0.80	0.66	0.85	0.41

PhS – Phonetic Alphabet
HW – Hidden Words
WinS – Words in Sentences
NW – New Words
AL – Artificial Language

The split-half reliability estimate for the whole battery is more than satisfactory (0.89) and indicates that 89 percent of the variation in observed scores is due to variation in the true scores, and the remaining 11 percent cannot be accounted for, and is due to the imperfection of the tool (Brown 1988: 98). The comparison of the descriptive statistics for the two halves indicates that the subjects' performance on them was very similar, which suggests a good degree of consistency in how the subjects responded to the tasks' items. The difference between the variances in the two halves can be assumed to be of little influence as the samples are equal (Brown 1988: 166). The fact that reliability indices estimated for individual tasks (Cronbach alpha) are lower than for the whole battery means that more caution must be shown when inferences are made on the basis of the individual tests (parts) rather than on the whole battery. The battery's descriptive statistics are shown in Table 6 below.

The inspection of the descriptive statistics shows that the distribution of scores of all the sub-tests as well as of the total TUNJO score approach the normal, "bell-like" shape. This is indicated by the "shape" statistics (skewness), which is close to zero (with the exception of PhS and NW), as well as by the dispersion statistics. The almost perfect fit of the observed Total score to the normal distribution is to be interpreted as

Table 6. TUNJO descriptive statistics, N=245.

Variable	Mean	Min.	Max.	St. D.	Skewness	Kurtosis
PhS	24.3	16	30	3.12	−0.395	−0.344
HW	18.1	4	33	5.99	0.124	−0.244
WinS	13.3	2	23	3.88	−0.029	−0.404
NW	16.7	4	24	5.13	−0.320	−0.869
AL	5.1	0	10	1.89	−0.009	−0.448
Total	**77.4**	**38**	**114**	**13.51**	**−0.096**	**−0.156**

the test being neither too difficult nor too easy for the majority of the participants tested. In the case of two scales, PhS and NW, the tasks proved to be a bit too easy.

The analysis of the internal correlations among the parts of the battery as well as between each part and the TUNJO total score (Table 7 below) shows that the first group of correlations, ranging from 0.17 to 0.38, displays lower values than the other group of coefficients which ranges from 0.41 to 0.80.

Table 7. TUNJO internal correlations.*

	PhA	HW	WinS	NW	AL
HW	0.34				
WinS	0.22	0.38			
NW	0.28	0.35	0.29		
AL	0.17	0.23	0.22	0.18	
TUNJO	**0.58**	**0.80**	**0.65**	**0.71**	**0.41**

* All are significant at $p < .01$.

Table 8 below gives the results of a preliminary analysis of TUNJO's predictive validity as described in Section 6.2.

Table 8. Correlations with the criterion, N=195.*

	PhA	HW	WinS	NW	AL	**TUNJO**
L2 criterion	0.16	0.15	0.26	0.27	0.12*	**0.31**

* All are significant at $p<0.05$ except for AL x L2.

Achievement tests scores were used as the L2 criterion against which TUNJO scores were correlated. The correlation of 0.31 of the TUNJO total scale with the criterion is statistically significant at $p<0.05$, which indicates that 9.61% of the variation in L2 scores can be meaningfully explained by the variation in the TUNJO total score. The only values which look promising are those for WinS and NW: 0.26 and 0.27 respectively. Still, of all the correlations reported above, only one value of *r,* for the AL task, did not reach statistical significance.

6.5. Discussion

The results presented in the previous section show two things. First, the lower values of *r* for the relationships among the sub-tests of TUNJO (cf. Table 7) can be interpreted as relative independence of the constructs tapped onto by those tasks. This was to be expected if each sub-test was meant to be an operationalisation of one of the Carroll's components of FL aptitude. Lower *r*'s among the sub-tests support the factorial structure of the model. Similarly, higher values of *r* for the relationships between each task and the total TUNJO score are indicative of the common aspects of each task's construct which is displayed by the contribution it makes, together with other constructs (sub-tests), towards the common, general construct, of which they are a part. Higher values of *r* here testify to the homogeneity of the whole battery as far as the higher-order dimension is concerned. The only somewhat disappointing result is that of the AL x TUNJO (0.41). This is probably due to the sub-test's rather low reliability (0.41), as shown in Table 5. The next lowest *r* is that for PhA x TUNJO (0.58). The reliability of the test is satisfactory (0.69) but, as the descriptive statistics reveal, it was the easiest task and as such, the least discriminating one. It is to be seen in the future validation studies whether the relatively poor discriminatory properties of PhA, are not linked to the score range restriction, characteristic of a poor representation of a wider spectrum of the abilities in a sample studied. This, it seems, was the case in the present study.

Secondly, with reference to the main aim of the current study, which was to provide a preliminary estimation of the TUNJO's predictive potential, it must be said that the results obtained are a bit disappointing. Given the fact that the comparable values reported for the MLAT in the literature (Skehan 1989, 1998; Dörnyei 2005) rarely go below 0.40, that is 16% of the explained variance, the present results are not particularly satisfactory. One of the reasons why this may be the case is the design of the study, and in particular, the characteristics of the sample population chosen for it. Secondary school learners of an L2 on which the TUNJO was tested, despite a relatively wide range of L2 proficiency abilities represented by them, are, by necessity, a limited sample of L2 learners. Consequently, the restriction of range on the scope of abilities revealed in the L2 measures will deflate the magnitude of the correlation coefficient (Cohen and Manion 1995). In this connection, it is interesting to observe that a similar study, aimed at the development and validation of an aptitude test by Kiss and Nikolov

(2005) done for 12-year-old Hungarian children, yielded the following values of *r*'s: 0.65 for "Hidden sounds"; 0.58 for "Words in sentences"; 0.70 for "Language analysis", and 0.61 for "Paired associates". Their participants displayed considerable differences in the intensity and the quality of their L2 experiences, which catered for a wider range of scores. This is a comforting thought for the present author.

7. Concluding remarks.

Generally speaking, the results of the study only partially attest to the predictive validity of the Polish version of the MLAT. The correlation of 0.31 between the battery and the L2 criterion is a bit disheartening, but not hopeless. On the other hand, it seems that despite the problems with the reliability of some of the TUNJO tasks (especially AL) which, to a certain extent, might be blamed for the battery's low predictive validity, the main problem seems to be with both the participants selected for the study and the lack of a uniform and reliable L2 criterion measure used to asses the subjects L2 proficiency. These are the factors which will have to be taken into account in subsequent TUNJO validation studies to follow this one.

REFERENCES

Anastasi, A. and S. Urbina. 1997. *Psychological testing*. Upper Saddle River: Prentice Hall.

Brzeziński, J. 2002. *Metodologia badań psychologicznych*. Warszawa: PWN.

Brown, J.D. 1988. *Understanding research in second language learning*. Cambridge: Cambridge University Press.

Carroll, J.B. 1968. "The psychology of language testing". In: Davies, A. (ed.), *Language testing symposium: A psycholinguistic approach.* London: Oxford University Press. 46–59.

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. 1981. "Twenty-five years of research on foreign language aptitude". In: Diller, K.C. (ed.), *Individual differences and universals in language learning aptitude*. Rowley, MA: Newbury House. 83–118.

Carroll, J.B. 1993. *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge: CUP.

Carroll, J.B. and S.M. Sapon. 2002 [1959]. *Modern Language Aptitude Test*. Bethesda, MD: Second Language Testing.

Cattell, R.B. and H.E.P. Cattell. 1973. *Measuring intelligence with the Culture Fair Tests.* Champaign, IL: Institute for Personality and Ability Testing.

Cohen, L. and L. Manion. 1995. *Research methods in education*. London: Routledge.

Dörnyei, Z. and P. Skehan. 2003. "Individual differences in second language learning". In: Doughty, C.J. and M.H. Long (eds.), *The handbook of second language acquisition.* Oxford: Blackwell. 589–630.

Dörnyei, Z. 2005. *The psychology of the language learner: Individual differences in second language acquisition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Ellis, R. 2004. “Individual differences in second language learning”. In: Davies, A.D. and C. Elder (eds.), *The handbook of applied linguistics*. Oxford: Blackwell. 525–551.
Figarski, W. 1984. *Wyznaczniki powodzenia w szkolnej nauce jezyka rosyjskiego*. Warszawa: Wydawnictwa Szkolne i Pedagogiczne.
Gardner, R.C. and W.E. Lambert. 1965. “Language aptitude, intelligence and second language achievement”. *Journal of Educational Psychology* 56.191–199.
Grigorenko, E.L., R.J. Sternberg and M.E. Ehrman. 2000. “A theory-based approach to the measurement of foreign language learning ability: The CANAL-F theory and test”. *Modern Language Journal* 84(3). 390–405.
Hambleton, R.K. 2006. *Adapting educational and psychological tests for cross-cultural assessment.* New York*:* Routledge.
Hambleton, R.K. and L. Pastula. 1999. “Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices”. *Journal of Applied Testing Technology* (August 1999).
Hornowska, E. 2003. *Testy psychologiczne. Teoria i praktyka*. Warszawa: Wydawnictwo Naukowe Scholar.
Hornowska, E. and W.J. Paluchowski. 2004. “Kulturowa adaptacja testów psychologicznych”. In: Brzeziński, J. (ed.), *Metodologia badań psychologicznych*. *Wybór tekstów*. Warszawa: PWN. 151–191.
Kiss, C., and M. Nikolov. 2005. “Developing, piloting, and validating an instrument to measure young learners’ aptitude”. *Language Learning* 55(1). 99–100.
Kuliniak, R. 2002. *Test Predyspozycji Językowych dla uczniów gimnazjum*. Wałbrzych: Bimart.
Lewowicki, T. 1975. *Psychologiczne różnice indywidualne a osiągnięcia uczniów*. Warszawa: Wydawnictwa Szkolne i Pedagogiczne.
Niżegorodcew, A. 1975. “Przydatność tzw. testu prognostycznego jako metody pomiaru zdolności do uczenia się języka obcego u młodzieży licealnej”. *Języki Obce w Szkole* 3/4. 299–303.
Niżegorodcew, A. 1979. “Rola specjalnych uzdolnień w nauce języków obcych w szkole”. *Przegląd Glottodydaktyczny* 3(1). 31–41.
Pearson. 1995. *Otis-Lennon School Ability Test*.
Ozga, J. and E. Tabakowska. 1973. “Egzamin wstępny z języka angielskiego dla kandydatów na filologię angielską Uniwersytetu Jagielońskiego w Krakowie”. *Języki Obce w Szkole* 1. 46–57.
Pimsleur, P. 1966 [2003]. *The Pimsleur Language Aptitude Battery*. Bethesda, MD: Second Language Testing.
Pimsleur, P. 1968. “Language aptitude testing”. In: Davies, A. (ed.), *Language testing symposium: A psycholinguistic approach.* London: Oxford University Press. 98–106.
Reber, A.S. 1985. *The Penguin dictionary of psychology.* London: Penguin Books.
Robinson, P. 2005. “Aptitude and second language acquisition”. *Annual Review of Applied Linguistics* 25. 46–73.
Robinson, P. 2007. “Aptitudes, abilities, contexts, and practice”. In: DeKeyser, R.M. (ed.), *Practice in second language: Perspectives from applied linguistics and cognitive psychology.* Cambridge: Cambridge University Press. 256–286.
Rybacka, A. 2008. A critical analysis of the validity of two Polish foreign language aptitude tests. [Unpublished MA thesis, Adam Mickiewicz University, Poznań.]
Sasaki, M. 1996. *Second language proficiency, foreign language aptitude and intelligence. Quantitative and qualitative analyses*. New York: Peter Lang.
Skehan, P. 1989. *Individual differences in language learning*. London: Arnold.

Skehan, P. 1992. "Foreign language learning ability: Cognitive or linguistic?" *Thames Valley University Working Papers in English Language Teaching* 2. 151–191.
Skehan, P. 1998. *A cognitive approach to language learning.* Oxford: Oxford University Press.
Skehan, P. 2001. "Theorizing and updating aptitude". In: Robinson, P. (ed.), *Individual differences and instructed language learning.* Amsterdam: John Benjamins. 69–94.
Sparks, R. and L. Ganschow. 1991. "Foreign language learning difficulties: Affective or native language aptitude differences?". *Modern Language Journal* 75(1). 3–16.
Sparks, R. and L. Ganschow. 2001. "Aptitude for learning a foreign language". *Annual Review of Applied Linguistics* 21(1). 90–111.
Spolsky, B. 1995. *Measured words.* Oxford: Oxford University Press.
Stansfield, C.W. and D.J. Reed. 2003. "Adaptation of the Modern Language Aptitude Test and The Pimsleur Language Aptitude Battery for examinees whose first language is not English". (Paper read at the 2nd Annual Conference of the East Coast Organization of Language Testers, 20–21 March 2003, Washington, D.C.)
Stansfield, C.W. and D.J. Reed. 2004. "The story behind the Modern Language Aptitude Test: An interview with John B. Carroll (1916–2003)". *Language Assessment Quarterly* 1(1). 43–56.
Tabakowska, E. and J. Ozga. 1975. "Przydatność testu prognostycznego przy rekrutacji kandydatów na filologię angielską". *Języki Obce w Szkole* 1. 35–37.
Terman, L.M. 1970. *Concept Mastery Test.* New York: Psychological Corporation.
Thurstone, L.L. 1938. *Primary mental abilities.* Chicago: The University of Chicago Press.
Wechsler, D. 1981. *WAIS-R manual: Wechsler Adult Intelligence Scale-Revised.* San Antonio, TX: Psychological Corporation.
Wojtowicz, M. 2006. *Test Zdolności Językowych*. Warszawa: Pracownia Testów Psychologicznych Polskiego Towarzystwa Psychologicznego.

Address correspondence to:
Jacek Rysiewicz
School of English, Adam Mickiewicz University
al. Niepodległości 4
61-874 Poznań
Poland
rjacek@ifa.amu.edu.pl