Chapter 10

The stability of language aptitude: Insights from a longitudinal study on young learners' language analytic abilities

Isabelle Udry[a,b] & Jan Vanhove[a]
[a]University of Fribourg, Institut de Plurilinguisme [b]Zurich University of Teacher Education

An enduring question in aptitude research is the extent to which aptitude is a stable trait or a time-varying attribute. If aptitude were a perfectly stable trait, interindividual differences in aptitude at one point in time should be perfectly correlated with interindividual differences at a later point in time. However, raw test scores are affected by measurement error, a result of which is that correlations between raw test scores at different points in time underestimate the correlations between the actual skills measured by these tests at different points in time. The analyses of the longitudinal LAPS II aptitude data ($n = 636$; translated and adapted versions of MLAT and PLAB subtests) take into account measurement error and indicate that the children's ability to solve the MLAT and PLAB tests at the first data collection (autumn 2017, mean age: 10;5 years) and their ability at the third data collection (spring 2019, mean age: 12;1 years) are correlated at $\rho = 0.74$ (95% CrI: [0.69, 0.79]). This suggests that the ability to solve the two aptitude tests is not a perfectly stable interindividual trait, but that, by and large, interindividual differences are nonetheless maintained over the course of one-and-a-half years of cognitive development.

1 Introduction

Language aptitude as defined by Carroll (1958) consists of four basic components, i.e. phonetic coding ability, grammatical sensitivity, inductive ability, and rote

Isabelle Udry & Jan Vanhove. 2021. The stability of language aptitude: Insights from a longitudinal study on young learners' language analytic abilities. In Raphael Berthele & Isabelle Udry (eds.), *Individual differences in early instructed language learning: The role of language aptitude, cognition, and motivation*, 197–209. Berlin: Language Science Press. DOI: 10.5281/zenodo.5464763

memory (see Chapter 1 for a discussion). One of the key debates in aptitude research is whether the construct is a stable characteristic or an ability that can be developed. Addressing this issue contributes to establishing a conceptual aptitude framework. It could also clarify whether fostering aptitude components enhances language learning. Despite educational and theoretical relevance, studies dealing with the stability of language aptitude in general, and particularly in children, remain scarce.

Aptitude stability can be explored in the data in different ways: 1. Researchers consider average aptitude test scores achieved by one or several groups of participants (we refer to this as the *group averages* approach) or 2. they determine an individual's relative ranking within the group (which we call the *relative ranking* approach). In both approaches, researchers then look for patterns, either cross-sectionally (by comparing groups at different developmental stages) or longitudinally (by comparing data obtained at different times from the same individuals or groups). How these patterns are explained depends amongst other things on the researchers' conceptualization of construct stability: Developmental changes (expressed as age-related gain scores obtained in an aptitude test) can be interpreted as construct malleability. Alternatively, such changes can be seen as indicating construct stability if an individual's ranking within the population or group remains largely constant, despite increased aptitude scores at different times of testing.

In our view, describing changes in average scores only is scarcely insightful in our context, as children are expected to score higher on aptitude tests as they mature, namely due to developmental changes in cognition. It would be more revealing to find out if they progress at the same rate (indicating a general developmental pattern and a stable trait) or differentially (indicating individual developmental patterns and therefore a malleable ability). Depending on whether the *group averages* and *relative ranking* change, different conclusions may be drawn:

1. Group averages and the relative ranking both change: possible indication of malleability
2. Group averages change, the relative ranking remains the same: possible indication of stability
3. Group averages do not change, but the relative ranking does (due to some individuals progressing while others' ability to solve the test decreases): possible indication of malleability
4. Neither group averages nor relative rankings change: possible indication of stability

Several authors assume that language aptitude is subject to change (see for instance Grigorenko et al. 2000, Singleton 2017). Their view runs contrary to earlier conceptions, as expressed by Carroll (1981: 86) who described language aptitude as "relatively hard to modify in any significant way." In the same article, the author (1981: 84) slightly qualified his statement by adding that the initial aptitude components could be modelled as "more or less enduring characteristics" and as a "current state".

The view of a stable characteristic has been substantiated in particular with evidence from a long-term study by Skehan (1986) and Skehan & Ducroquet (1988) which found that some measures of L1 development, namely L1 vocabulary and mean length of utterance, were predictive of L2 aptitude measures assessed 13 years later in the same participants.

It is worth remembering that Carroll, who first conceptualized language aptitude, was mainly concerned with capturing a snapshot of people's potential before they started learning a language in order to predict their later L2 achievement. Whether this potential was innate or malleable was not an explicit question in these early stages of aptitude research. However, as early as 1964, careful readers could detect Carroll's awareness of the issue tucked away in a footnote, stating that the extent to which "the behaviour measured on the aptitude tests … can be modified by training" would still "need to be investigated" (Carroll 1964: 89). Later, he expressed doubts about the feasibility of such a training, stating that "there are some general grounds for pessimism regarding the teaching of aptitudinal skills" (1973: 8), contributing to his view was the fact that not enough research had been conducted on the matter.

2 Review of the literature

Studies investigating the stability of language aptitude are rather scarce and usually underpinned by the hypothesis that aptitude is shaped by language experience. Roehr-Brackin & Tellier (2019) examined the relationship and development of language aptitude and metalinguistic awareness with 111 anglophone beginning learners of L2 French aged 8 to 9 years. In phase 1 (16 weeks) of the project participants were divided into four groups that were taught either German, Italian, Esperanto, or Esperanto with an added focus-on-form element. In phase 2 (16 weeks), all children learnt French with an element of focus-on-form. The authors administered tests for aptitude (an adaptation of the MLAT-E by Carroll & Sapon (1976) for British English speakers), metalinguistic awareness, and L2 French proficiency (listening, reading, writing, grammar) in a pretest–posttest

design which included immediate and delayed posttests for L2 French. The authors detected increases in aptitude test scores with a medium effect size. According to the "*group averages*" definition of stability given in the introduction, they concluded that language aptitude was dynamic in the sample. Moreover, the authors found that children who performed well on the aptitude pretest also did well on the aptitude posttest, and vice versa, suggesting a largely stable ranking. This finding was interpreted in favour of the aptitude test's predictive value for young learners' L2 performance. However, based on the "*relative ranking*" definition of stability, we argue that this finding could also point to the stability of language aptitude.

An increase in aptitude test scores with age was also detected by Suárez Vilagran (2010) who administered Spanish and Catalan translations of the MLAT-E (see §2.2) to 629 Spanish-Catalan bilingual learners of English aged 8 to 15. She observed a considerable increase in aptitude scores between ages 8 and 9. After the age of ten, gain scores weren't as large, and the author therefore suggests that aptitude may stabilize around age 11.

Kiss & Nikolov (2005) tested aptitude (with a Hungarian MLAT based test), motivation, and English proficiency (listening, reading, writing) in 419 12-year-old L2 English learners. The authors also recorded time of exposure to English at school and in private tuition which ranged considerably from 100 to 1085 hours ($M = 343; \mathrm{SD} = 131$). Kiss and Nikolov explored the effects of language experience on aptitude by establishing correlations between time spent on learning and aptitude test scores. This correlation being weak, they concluded that language aptitude in the Carrollian sense did not improve with "the amount of time used for practice and exposure" (Kiss & Nikolov 2005: 134).

In a subsequent study, Kiss (2009) reconsidered the interplay between language experience and aptitude. The author administered a Hungarian aptitude test for young learners to 52 Hungarian children. The aim was to select 26 8-year-olds for a newly established bilingual English-Hungarian teaching program. The author compared the results from all 8-year-old children to those from 12-year-olds from a previous study. She found that the 12-year-olds performed better on the vocabulary learning subtest than their younger counterparts. Kiss (2009: 268f) speculates that these differences are owed to the older children's greater language learning experience and knowledge of strategy use. Referring to the idea that increased group averages reflect aptitude malleability, the author argues that language aptitude is dynamic, at least up to the age of 12.

A frequently cited study by Sáfár & Kormos (2008) addressed the stability of the construct with 61 Hungarian learners of English aged 15 to 16 years. The authors assessed language aptitude and short-term memory at the beginning

and end of the academic year. 41 participants followed an English-Hungarian bilingual program (with sixteen 45 min English lessons per week + 4 × 45 min. CLIL per week) and 21 participants were from a regular Hungarian secondary school (with 4 × 45 English lessons a week predominantly communicative with some focus-on-form instruction). Aptitude test scores increased significantly in both groups between the two data collections, independent of the intensity of instruction. Learners in the bilingual program, however, improved more than their counterparts in the regular setting. Based on these findings, the authors concluded that language aptitude is dynamic and changes with language experience. The authors also stated that language aptitude appears to be less relevant in communicative teaching with a focus-on-form element. It seems worth noting that some information on other potential influences, such as general learning abilities or admission criteria for the bilingual program, may have contributed to explaining the results.

In summary, current empirical findings suggest that language aptitude is dynamic in children younger than 12. Primary school children score higher on aptitude tests as they get older, with important increases in test scores being observed between 8 and 9 years (Milton & Alexiou 2006, Kiss 2009, Suárez Vilagran & Muñoz 2011). These findings are based on the *group averages* definition of stability, i.e. a gradual improvement in the average performance of groups of children.

3 Method

We investigated the stability of language aptitude in primary school children over a period of 1.5 years. To this aim, we defined language aptitude as language analysis (Skehan 1998), i.e. the grammatical sensitivity and inductive ability components from Carroll's (1958) construct definition (see also Chapter 1 for a discussion). As outlined earlier, primary school children are still maturing cognitively, and their aptitude scores are expected to improve with age. Therefore, we were more interested in the extent to which individual differences in language aptitude remain stable over time, i.e. whether participants who perform well relative to other participants on an aptitude test at one point T1 will still do so at a later time of testing T2 (or even T3, as in our data). This can be inferred from the correlation between the participants' test results at T1 and T2 (and T3): The stronger this correlation is, the more stable interindividual differences in language aptitude are.

4 Participants and procedure

The study design is fully described in Chapter 2 and will be summarised briefly: To assess language analytic ability (i.e. grammatical sensitivity and inductive ability), we adapted the following tests for German-speaking young learners[1]: MLAT-E subpart "Matching Words" on grammatical sensitivity (Carroll & Sapon 1976) and PLAB subpart 4 on inductive ability (Pimsleur et al. 2004). The same participants from LAPS II (4th and 5th graders at T1), completed these tests at three different times T1–T3: T1 = Autumn 2017 (mean age 10;5); T2 = Spring 2018 (mean age 11); T3 = Spring 2019 (mean age 12;1). A total of 636 participants completed the tests between T1 and T3. Figures 1 and 2 show the correlations between the test scores from T1 to T3. Table 1 summarizes the test results.

Table 1: MLAT and PLAB test results at T1–T3. The total number of participants completing the tests at different times between T1–T3 is 636. MLAT has 30 items and PLAB 15.

Test	Time	Participants	Mean[a]	SD
MLAT	T1	590	0.48	0.23
	T2	575	0.58	0.22
	T3	565	0.66	0.19
PLAB	T1	596	0.37	0.19
	T2	577	0.44	0.22
	T3	564	0.56	0.25

[a]Proportion of correct answers

Clearly, the results from the three testing times do not correlate perfectly (this would be indicated by a correlation coefficient of +1). Even if the ability to solve the MLAT or the PLAB were interindividually stable, we would still not expect to see a perfect correlation. This is due to measurement error: If two latent variables that correlate perfectly with each other are measured imperfectly, these measurements will still not yield a perfect correlation.

If we knew the measurement error or the reliability coefficient of the tests, we could solve this problem by calculating disattenuated correlations. While we do not know the actual reliability of the instruments, we can estimate them. The reliability coefficients (ω_{RT}, McNeish 2018, Revelle 2019) for the MLAT variable are

[1]We would like to thank Charles W. Stansfield for permission to translate and adapt parts of the MLAT-E and PLAB for this study.

0.90 (T1), 0.89 (T2) and 0.88 (T3); for PLAB: 0.74 (T1), 0.80 (T2) and 0.84 (T3). If we use these values to disattenuate the correlations (using the correct.cor() function in the psych package for R, William 2018), we obtain the values in Table 2.

Table 2: Disattenuated correlations MLAT and PLAB

	MLAT		PLAB	
	T1	T2	T1	T2
T2	0.77		0.60	
T3	0.62	0.78	0.61	0.65

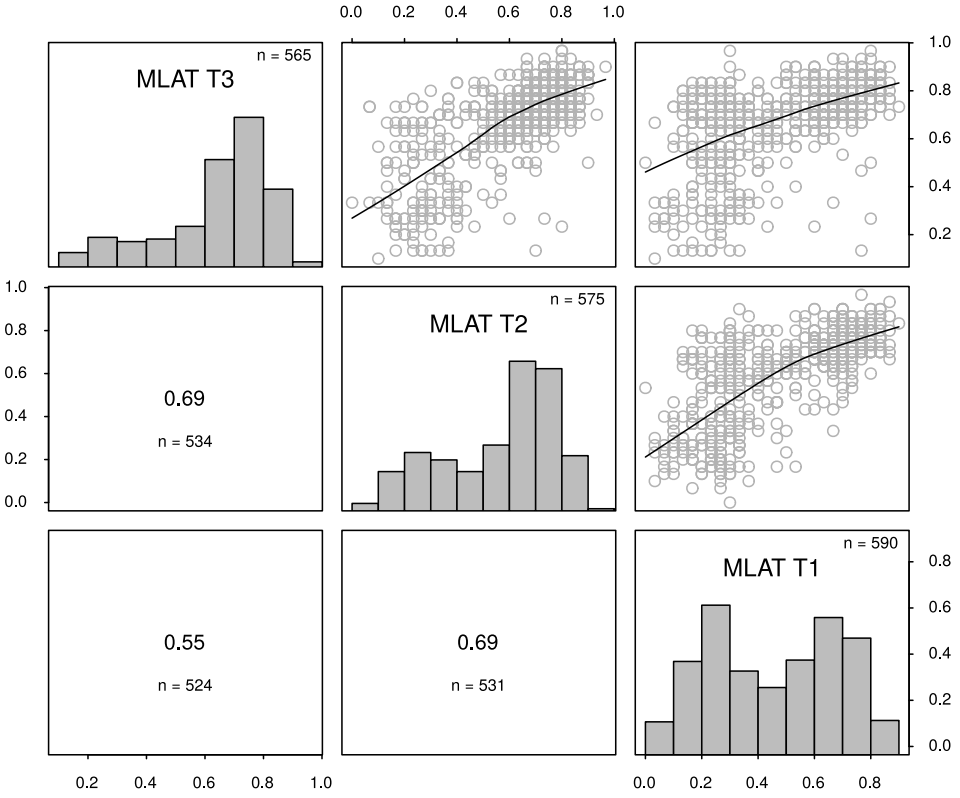

Figure 1: Scatterplot matrices with the MLAT results at the three data collections. Upper triangle: Scatterplots with scatterplot smoothers. Main diagonal: Histograms. Lower triangle: Pearson correlation coefficients as well as the number of data points on which these were based.

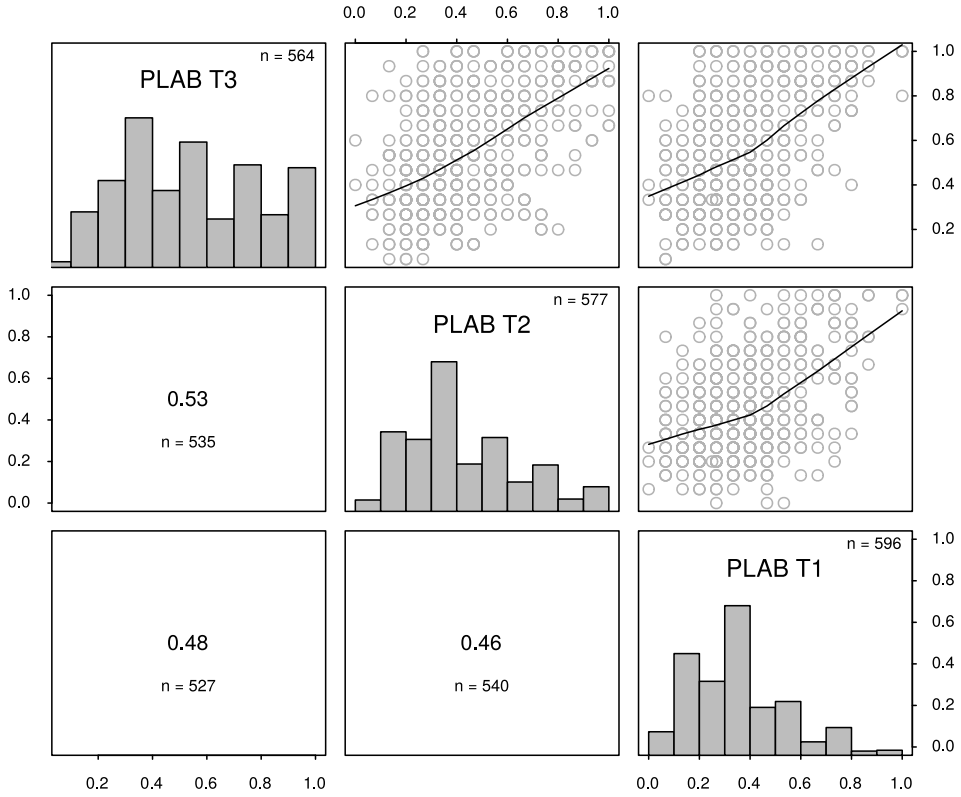

Figure 2: PLAB results and correlations at different times of testing T1–T3. Scatterplot matrices with the MLAT results at the three data collections. Upper triangle: Scatterplots with scatterplot smoothers. Main diagonal: Histograms. Lower triangle: Pearson correlation coefficients as well as the number of data points on which these were based.

While computing disattenuated correlation coefficients is fairly straightforward, this analysis does not consider the dependence that exists between the tests: The MLAT scores for T1, T2 and T3 are based on the same test items, and the same goes for the PLAB scores. To take these dependencies into account, we ran an alternative, if more involved, analysis in which the participants' item-level responses were modelled. This analysis was run using generalized (logistic) mixed-effects models, which are capable of estimating the participants' latent abilities as well as the items' latent difficulties.

In language research, analyses based on mixed-effects models usually focus on the fixed effects, but in our case, it is the random effects that are of particular interest. For each participant, the latent ability to solve the MLAT and PLAB can

be estimated for T1, T2 and T3. Also, the correlations between the participants' estimated latent abilities at T1, T2 and T3 can be estimated. In doing so, the analysis can also take into account the fact that items vary in their difficulty and that the relative difficulty of the test items may vary between T1, T2 and T3.

We fitted three models: one on the MLAT responses, one on the PLAB responses and one on all responses combined. The models we fitted were Bayesian generalized (logistic) mixed-effects models; for our purposes, Bayesian models have the advantage that they can not only estimate the correlation between the participants' latent abilities at T1, T2 and T3, but also quantify the uncertainty about this estimation. The models were fitted using the brm() function in the brms package for R (Bürkner 2017). "Result" is a binary variable that indicates for each individual response whether it was correct (1) or not (0). "Time" is a categorical variable with three levels (T1, T2, T3), "Item" is a categorical variable specifying the test item, and "StudentID" is a categorical variable specifying the participant. The three models were specified as follows (in brms notation):

```
m <- brm(result ~ 0 + Time + (0 + Time | Item) + (0 + Time |
    StudentID), data = d, family = bernoulli(link = "logit"), cores
    = 4, iter = 4000, warmup = 1000)
```

This model estimates (a) the probability (in logits) of a response being correct at Times 1, 2 and 3, (b) between-item differences in this probability, (c) between-participant differences in this probability, (d) the correlations among the between-item differences at Times 1, 2 and 3, and (e) the correlations among the between-participant difference at Times 1, 2 and 3. For our purposes, (e) is what is important, viz., the extent to which differences among the participants' abilities at Time 1 are maintained at Times 2 and 3.

5 Results

The MLAT model was fitted on 51,875 responses (30 items × 636 participants × 3 data collections, with some missing data); the PLAB model was fitted on 26,044 responses (15 items × 636 participants × 3 data collections, with some missing data); the combined model was fitted on 77,919 responses (45 items × 636 participants × 3 data collections, with some missing data).

Tables 3–5 summarise the main results pertinent to our research question. The full model output can be inspected at https://osf.io/pf5g8/. Posterior predictive checks indicated that the models reported here can generate the key characteristics of the dataset; these checks are also reported in the online appendix.

Overall, correlations range from moderate to strong (0.63–0.83). The MLAT correlations are stronger (0.65–0.79) than the PLAB correlations (0.63–0.68) and even stronger correlations are obtained when the two tests were considered together (0.74–0.83). Also, correlations are stronger for short intervals (T1–T2; T2–T3) than for the longest period T1–T3.

The first model estimates that the differences in solving the MLAT (expressed in logits) is correlated between T1 and T2 at 0.78 (95% uncertainty interval: [0.74; 0.83], between T2 and T3 at 0.79 ([0.74; 0.84]), and between T1 and T3 at 0.64 ([0.58; 0.71]).

For the PLAB, this ability was estimated to correlate with 0.63 (95% uncertainty interval: [0.54; 0.72]) from T1 to T2, 0.68 from T2 to T3 and 0.66 between T1 and T3.

In a third model, the overall ability to solve both MLAT and PLAB was analysed in the same way. Results show that language analytic abilities as measured by both tests were correlated at 0.83 ([0.78 0.86]) between T1 and T2, 0.82 ([0.78 0.86]) between T2 and T3 and 0.74 ([0.69 0.79]) from T1 to T3.

6 Discussion

We were interested in the stability or relative development of language-analytic ability as a subcomponent of language aptitude. Language-analytic ability was assessed in 636 primary school children aged 10–12 years at three times over 1.5 years with adaptations of the MLAT-E Matching Words (grammatical sensitivity) and the PLAB subtest for inductive ability. Our findings indicate that overall results improved over time, with test scores being highest at T3 for both MLAT and PLAB. A gradual increase in test scores was expected due to the children's cognitive maturation. As one reviewer pointed out, higher scores may also be linked to test familiarity. Even though the MLAT and PLAB were administered at intervals of 6 months (T1–T2) and 12 months (T2–T3), practice effects cannot be ruled out entirely and it is possible that maturation and test familiarity are intertwined to some degree.

We also adopted an interindividual perspective, assessing whether our participants' *relative* ability to solve the aptitude tests remained stable with increasing age. Recall that by latent ability or relative ability to solve the aptitude tests, we mean the correlation of the *relative difference* of the test scores between testing times. In other words, this correlation indicates how strongly the ranking among participants based on their scores has changed over time.

We will discuss the longest interval T1 to T3 which most adequately mirrors long-term changes in our data. Correlations between T1 and T3 are strong

Table 3: The estimated correlations between the relative differences of participants' abilities to solve the MLAT test (expressed in logits) and their 95% credible intervals.

Parameter	Est.	95% credible interval	
		lower bound	upper bound
$\hat{P}$ T1–T2	0.78	0.74	0.83
$\hat{P}$ T1–T3	0.65	0.58	0.71
$\hat{P}$ T2–T3	0.79	0.74	0.84

Table 4: The estimated correlations between the relative differences of participants' abilities to solve the PLAB test (expressed in logits) and their 95% credible intervals.

Parameter	Est.	95% credible interval	
		lower bound	upper bound
$\hat{P}$ T1–T2	0.63	0.54	0.72
$\hat{P}$ T1–T3	0.66	0.57	0.74
$\hat{P}$ T2–T3	0.68	0.61	0.75

Table 5: The estimated correlations between the relative differences of participants' abilities to solve the MLAT and PLAB tests (expressed in logits) and their 95% credible intervals.

Parameter	Est.	95% credible interval	
		lower bound	upper bound
$\hat{P}$ T1–T2	0.83	0.78	0.86
$\hat{P}$ T1–T3	0.74	0.69	0.79
$\hat{P}$ T2–T3	0.82	0.78	0.86

($\rho = 0.74$) when both tests are considered together, and moderate when the tests are considered separately, i.e. $\rho = 0.65$ for the MLAT and $\rho = 0.66$ for the PLAB. These results suggest that language-analytic ability, as operationalised by these tests, is not entirely stable, as this would have been evidenced by correlations in an even higher range. At the same time, moderate to strong correlations indicate that a relationship between language-analytic ability at T1 and T3 is still present in the data. In conclusion, the ability to solve the two aptitude tests is not a perfectly stable interindividual trait, but, by and large, interindividual differences were nonetheless maintained over the course of 1.5 years of cognitive development.

References

Bürkner, Paul-Christian. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80(1). 1–28.

Carroll, John B. 1958. A factor analysis of two foreign language aptitude batteries. *The Journal of General Psychology* 58(1). 3–19. DOI: 10.1080/00221309.1958.9710168.

Carroll, John B. 1964. The prediction of success in intensive foreign language training. In *Training research and education*, 87–136. Pittsburgh: University of Pittsburgh Press.

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

Carroll, John B. & Stanley M. Sapon. 1976. *Modern language aptitude test – Elementary*.

Grigorenko, Elena L., Robert J. Sternberg & Madeline E. Ehrman. 2000. A theory-based approach to the measurement of foreign language learning ability: The Canal-F theory and test. *The Modern Language Journal* 84(3). 390–405.

Kiss, Csilla. 2009. The role of aptitude in young learners' foreign language learning. In Marianne Nikolov (ed.), *The age factor and early language learning*, 253–276. Berlin: De Gruyter Mouton.

Kiss, Csilla & Marianne Nikolov. 2005. Developing, piloting, and validating an instrument to measure young learners' aptitude. *Language Learning* 55(1). 99–150.

McNeish, Daniel. 2018. Thanks coefficient alpha, we'll take it from here. *Psychological Methods* 23(3). 412–433. DOI: 10.1037/met0000144.

Milton, James & Thomaï Alexiou. 2006. Language aptitude development in young learners. In C. Abello-Contesse, R. Chacón-Beltrán & M. D. López-Chiménez (eds.), *Age in L2: Acquisition and teaching*, 177–192. New York: Peter Lang.

Pimsleur, Paul, Daniel J. Reed & Charles W. Stansfield. 2004. *Pimsleur language aptitude battery PLAB*. Rockville: Second Language Testing.

Revelle, William. 2019. *Using R and the psych package to find ω*. http://personality-project.org/r/psych/HowTo/omega.pdf.

Roehr-Brackin, Karen & Angela Tellier. 2019. The role of language-analytic ability in children's instructed second language learning. *Studies in Second Language Acquisition* 41. 1111–1131.

Sáfár, Anna & Judit Kormos. 2008. Revisiting problems with foreign language aptitude. *International Review of Applied Linguistics in Language Teaching* 46(2). 113–136.

Singleton, David. 2017. Language aptitude: Desirable trait or acquirable attribute? *Studies in Second Language Learning and Teaching* 7(1). 89–103.

Skehan, P. & L. Ducroquet. 1988. *A comparison of first & foreign language learning ability* (ESOL Working Document). London: University of London. https://books.google.ch/books?id=XCJfcgAACAAJ.

Skehan, Peter. 1986. Cluster analysis and the identification of learner types. In Vivian J. Cook (ed.), *Experimental approaches to second language acquisition*, 81–94. Oxford: Pergamon Press.

Skehan, Peter. 1998. *A cognitive approach to language learning*. Oxford: Oxford University Press.

Suárez Vilagran, Maria Del Mar. 2010. *Language aptitude in young learners: The elementary modern language aptitude test in Spanish and Catalan*. Barcelona: Universitat de Barcelona. (Doctoral Dissertation).

Suárez Vilagran, Maria Del Mar & Carmen Muñoz. 2011. Aptitude, age and cognitive development: The MLAT-E in Spanish and Catalan. *EUROSLA Yearbook* 11(1). 5–29. DOI: 10.1075/eurosla.11.03sua.

William, Revelle. 2018. *psych: Procedures for personality and psychological research*. https://www.scholars.northwestern.edu/en/publications/psych-procedures-for-personality-and-psychological-research.