Annual Review of Applied Linguistics (2021), **41**, 25–54
doi:10.1017/S0267190520000136

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

The methodology of the research on language aptitude: A systematic review

Shaofeng Li[1]* and Huijun Zhao[2]

[1]Florida State University and [2]Yango University
*Corresponding author. Email: sli9@fsu.edu

Abstract

This article provides a comprehensive and critical synthesis of the methods utilized in studies investigating the role of language aptitude in second language acquisition (SLA). The synthesis is informed by sixty-five studies generated by a thorough search of the literature, three meta-analyses (Li, 2015, 2016, 2017), and a thematic issue of *Studies in Second Language Acquisition* (Li & DeKeyser, in press). The synthesis starts by identifying three categories of research investigating the role of aptitude in naturalistic learning, aptitude's associations with instructed learning, and the nature of aptitude pertaining to whether it increases with age and learning experience and how it is connected to other individual difference variables. The synthesis then presents an overview and critique of major measures of aptitude and discusses the construct validity of aptitude measures based on the principles of psychometric assessments. Specifically, the measures are scrutinized along the dimensions of reliability, content validity, divergent/convergent validity, and predictive validity. The content and measurement of implicit aptitude—a newly emerged construct in SLA—are highlighted. The synthesis proceeds to summarize and vet the measures of the outcome variable of aptitude research—L2 proficiency. Throughout the synthesis, methodological features are summarized, issues are identified, and remedies are proposed.

Keywords: language aptitude; second language acquisition; research methods; systematic review; explicit and implicit aptitude

The Methodology of the Research on Language Aptitude: A Systematic Review

Language aptitude (alternatively "aptitude") refers to a set of cognitive abilities that are predictive of learning rate and ultimate attainment in a second language. Knowledge about aptitude is key to an accurate understanding of the mechanism of second language acquisition (SLA) and has significant practical implications for language teaching and learning. Since its inception in the 1950s, aptitude has been examined in a large amount of research from various perspectives, including its content and characteristics as well as its relation to age, instruction type, and other important aspects of language learning, such as explicit and implicit knowledge (Li, 2018; Skehan, 2012). However, to date the methodological features of the research have not been systematically

synthesized, scrutinized, and vetted, although the value of empirical evidence relies crucially on the validity of the research methods. This article seeks to provide a comprehensive and critical synthesis of the methods utilized in the studies investigating the role of language aptitude in second language (L2) learning. The objective of this synthesis is trifold. One is to inform the field how aptitude research has been conducted, with a view to assisting aptitude researchers with their own research and enhancing their understanding of current research. The second objective is to identify issues and limitations of current research, alerting researchers to potential pitfalls in future research. The third objective is to recommend remedies to observed problems and point out future directions. In the following sections, we start by providing some background information about aptitude and then report the methods of this synthesis, followed by a detailed synthesis of the methodological parameters of the primary research including (1) research designs, foci, and analyses, (2) aptitude measures, and (3) outcome measures. A conclusion is offered at the end.

The research on language aptitude has been dominated by traditional aptitude (synonymous with explicit aptitude in this article), and implicit aptitude made its debut in recent research. Traditional aptitude, the type measured by the MLAT and similar tests, is hypothesized to consist of three components: phonetic coding, analytic ability, and rote memory, which roughly correspond to pronunciation, grammar, and vocabulary learning, respectively (Carroll & Sapon, 2002). Traditional aptitude has been found to have the following characteristics: it increases with age until a certain point (around age ten) and then starts to stabilize (Roehr-Brackin & Tellier, 2019); it is subject to learning experience (Sáfár & Kormos, 2008); it is correlated with but dissociable from intelligence, uncorrelated with motivation, and negatively correlated with anxiety (Li, 2016). In terms of its associations with learning outcomes, traditional aptitude is drawn on by both early (Abrahamsson & Hyltenstam, 2008) and late bilinguals (DeKeyser, 2000; Granena & Long, 2013) in naturalistic learning; it is a strong predictor of instructed learning, but its predictive power is stronger for beginners than for learners with higher proficiency (Li, 2015); high-aptitude learners benefit more from inductive instruction, while low-aptitude learners benefit more from deductive instruction (Hwu et al., 2014; Erlam, 2005); it is more strongly correlated with explicit instruction than implicit instruction (Li, 2017).

The finding that traditional aptitude is more likely to be involved in explicit instruction that encourages conscious learning has instigated researchers' interest in investigating implicit language aptitude—cognitive abilities important for unconscious language learning. Implicit aptitude is a new concept, and the research is still in its infancy. Initial findings demonstrate that (1) implicit aptitude is separate from explicit aptitude or cognitive abilities in the explicit paradigm (Granena, 2019; Li & Qian, in press); (2) it increases with age and does not decline as significantly as explicit aptitude (Ullman & Lovelett, 2018); (3) it is drawn on by both early and late bilinguals (Granena, 2013), similar to explicit aptitude (Abrahamsson & Hyltenstam, 2008); (4) it is predictive of naturalistic L2 attainment in learners with longer residence in the host country (Suzuki & DeKeyser, 2017); (5) it is drawn on in implicit instruction but not in explicit instruction (Yilmaz & Granena, 2019), contrary to the pattern found for explicit aptitude, which is more strongly correlated with explicit than implicit instruction (Li, 2017); (6) it is involved in later but not initial stages of learning in highly controlled incidental learning conditions (Hamrick, 2015; Morgan-Short et al., 2014); (7) it is correlated only with the learning of agreement structures such as subject-verb agreement but not structures that involve form-meaning mapping, such as the subjunctive mood (Granena, 2013).

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

The Synthesis

This methodological synthesis is informed by three meta-analyses conducted on the empirical research on traditional aptitude (Li, 2015, 2016, 2017), a special issue of *Studies in Second Language Acquisition* investigating the validity of the newly emerged construct of implicit aptitude, and a thorough search of three prestigious electronic bases in SLA and psychology: LLBA, PsycInfo, and PsycArticles. Key search words include both terms representing the constructs such as aptitude, language aptitude, explicit aptitude, implicit aptitude, cognitive aptitude, and so on, as well as names of the tests of the constructs such as MLAT, LLAMA, Hi-LAB, serial reaction time, etc. Aside from terms for overall aptitude, components of aptitude, such as phonetic coding, analytic ability, rote memory, and musical aptitude were also used as key search words. For implicit aptitude, searches were also conducted using terms for different paradigms of implicit learning, such as procedural memory, statistical learning, and syntactic priming.

Because the purpose of this synthesis is to identify trends and issues and recommend remedies rather than tally the results, and because it is impossible to include all the studies, only selected studies are included in this synthesis. The selection criteria were carefully formulated to ensure that the included studies are representative of all major foci and streams of aptitude research. Selected studies were published after 2000, and they investigated aptitude either as a predictor of learning outcomes or as an outcome variable that varies as a function of age and learning experience. Also included were studies examining aptitude's associations with other individual differences factors, such as anxiety, motivation, and learning strategies. If one or one group of authors contributed multiple studies using similar methods, only one or two representative studies were selected. Based on the selection criteria, a total of sixty-five studies were included in the synthesis. (Synthesized studies cited in this article are included in the references, and the complete list can be seen in Supplementary Information).

The synthesis has the following features. It seeks to provide a thorough review of the methodology of aptitude research rather than a fragmentary discussion of particular aspects. It provides a comprehensive coverage of the topic encompassing research design, samples, treatments, measurement of independent and dependent variables, and data analyses. The synthesis does not tally the methodological features of primary research, but it addresses all the prominent features and provides an in-depth discussion of each. It is based on transparent selection criteria and a thorough search of the literature rather than cherry-picked studies based on unknown criteria (Li & Wang, 2018). In the following sections, the synthesis starts by identifying major streams of aptitude research, providing an overview of the research foci, designs, and data analyses of each stream. The synthesis then focuses on the measurement of aptitude, which is the backbone of aptitude research, followed by a discussion of the measurement of the dependent variable of aptitude research—L2 proficiency.

Research Foci, Designs, and Analyses

Aptitude and L2 Attainment in Naturalistic Settings

A naturalistic setting is one in which the second language is the primary language of the community, such as acquiring English in an English-speaking country. Naturalistic studies fall into two categories: age-related studies involving learners who arrived in the country of the L2 at different ages and studies that involved mainly learners who

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

arrived in the setting during adulthood. Age-related studies investigate maturation effects in second language acquisition or the Critical Period Hypothesis from the perspective of language aptitude (Abrahamsson & Hystenstam, 2008; DeKeyser, 2000; Granena & Long, 2013). The primary objective of these studies is to verify the hypothesis that learners whose ages of onset (abbreviated as AO—age at which one starts to have massive exposure to a second language) are within the critical period do not draw on (explicit) language aptitude, while learners who start to learn a second language in adulthood do. Typically, these studies involve learners who lived in the country of the target language for more than ten years, and they were divided into different groups based their ages of onset. The learners are tested on their aptitude and their L2 proficiency, and then correlation analyses are conducted for each AO group to explore whether learners' aptitude and proficiency are significantly correlated. There exists some methodological variation between the studies, for example, the cutoff points for forming age groups are different. In Abrahamsson and Hyltenstam (2008), learners were split into two groups based on whether they came to the country before or after twelve years of age; in DeKeyser (2000), the cutoff point was sixteen; Granema and Long (2013) divided their participants into three AO groups: 3–6, 7–15, and 16–29, based on the assumption that there is a critical or sensitive period for phonology, lexis/collocations, and morphosyntax, respectively. The studies also used different measures of aptitude—LLAMA and the Words in Sentences subtest of the HUNLAT—and different measures of L2 attainment, and the methodological variation may be responsible for the disparity of their findings. Among the three studies, Abrahamsson and Hyltenstam's study differed from the other two studies in several ways. In their study, all learners passed for native speakers, and the outcome variable was the learners' scores on a challenging grammaticality judgment test (GJT) to prevent ceiling effects resulting from the learners' high proficiency. The researchers stated that it would have made it difficult to detect interlearner variation by using an easy test, as happened in DeKeyser's (2000) study. In data analysis, aside from correlation analysis for each AO group between learners' aptitude and GJT scores, the researchers plotted the two groups' aptitude scores to show that the late starters all had high language aptitude and the early starters' aptitude scores were spread out, suggesting that, to achieve native-like proficiency, adult learners must have exceptional aptitude, and child learners do not rely on aptitude. Note that these studies have examined explicit aptitude and they did not distinguish explicit and implicit L2 knowledge. To date, Granena (2013) is the only age-related study that investigated the associations between implicit aptitude and early and late bilinguals' implicit and explicit L2 knowledge. One original methodological feature of Granena's study is the categorization of the tested linguistic structures into agreement and nonagreement structures to explore whether aptitude has differential associations with different types of linguistic structures. Unlike the age-related studies aiming to map the relationship between aptitude and the ultimate attainment of learners who started to be exposed to the L2 at different ages, several studies only involved learners who arrived in the country of the target language during adulthood. These studies aimed to examine (1) whether implicit and explicit aptitude are associated with implicit and explicit L2 grammar knowledge (Suzuki & DeKeyser, 2015, 2017) or with knowledge about L2 collocations (Yi, 2018), or (2) whether phonetic coding abilities are predictive of adult naturalistic learners' L2 pronunciation learning (Saito et al., 2019).

Further comments are in order for naturalistic studies. First, the age-related studies all examined the associations between age of onset and ultimate attainment—the end

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

state of a second language—regardless of their learning experience. Thus, the research is revealing of how aptitude relates to the product, not the process, of SLA. The claims about how early and late bilinguals draw on aptitude differently in the learning process are retrospective in that inferences are made about what occurred in the past based on the outcome measured in the present. Although the learners with early ages of onset started to be exposed to the L2 in childhood, they were adults when participating in the research, and, to some extent, it is misleading to label these learners as child language learners and claim that the findings based on these learners represent child language acquisition. Of course, it is a valuable endeavor to examine the interface between aptitude type, age of onset, and ultimate attainment, but the results should be interpreted accordingly, rather than making inferences on how child and adult learners draw on implicit and explicit aptitude in different ways. Second, age-related studies typically include learners with early ages of onset whose L2 proficiency is native-like. Often their test scores reach ceiling and lack variability, which makes it difficult to achieve statistically significant results. One way to overcome this limitation is to control proficiency, such as in Abrahamsson and Hyltenstam (2008), where all subjects achieved native-like proficiency, which makes it possible to explore whether early and late starters' aptitude scores are distributed differently. Third, in age-related research, it seems customary to include native speakers of the target language to compare the results for L1 and L2 speakers, ascertain whether aptitude's associations with learners' L2 proficiency is unique to L2 learning, and seek explanations for aptitude effects (Abrahamsson & Hyltenstam, 2008; Yi, 2018). Fourth, studies on naturalistic learners always face the challenge of the heterogeneity of learners' background, which makes it difficult to tease out the effects of aptitude from other variables, such as learning experience, and find significant effects for the investigated variables. It is, therefore, important to include, measure, or control other factors when examining aptitude effects on naturalistic L2 learning. Fifth, the studies seem to show that length of residence mediates the relationships between aptitude and L2 attainment. For example, the studies by Suzuki and DeKeyser found that implicit aptitude (SRT) was a near-significant predictor of long-residence (> 2 years) learners' implicit knowledge, while explicit aptitude (LLAMA_F) was predictive of short-residence (< 2 years) learners' automatized explicit knowledge. Therefore, for studies on naturalistic learning, it is important to consider and investigate the impact of length of residence as a moderator variable for aptitude-attainment associations.

Aptitude and L2 Attainment in Instructed Settings

The bulk of aptitude research investigates the associations between aptitude and the outcome of instructed learning, which occurs primarily in the classroom, and these studies happen to be conducted in foreign language settings where the target language is not the language of the community. The studies can be divided into two broad categories: correlational and experimental, with the former referring to studies investigating the associations between aptitude and learning outcomes regardless of the type of instruction learners receive, and the latter to studies examining the interface between aptitude and the effects of different instructional treatments. Here the distinction between correlational and experimental research is made from the perspective of research design, not statistical analysis, and, in fact, in both types of research, correlation analyses or analyses of correlational nature such as multiple regression are the primary statistical analyses researchers employed.

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

Correlational Research

Studies adopting a correlational design aim to investigate whether aptitude is predictive of learning outcomes, and these studies recruit learners from intact classes and do not manipulate instructional intervention/treatment or seek to investigate whether aptitude is drawn on differently under different learning conditions. This is a product-oriented approach, where aptitude is a determinant of learning success. The design of correlational research is straightforward. Typically, learners from elementary, secondary, or tertiary language classes are recruited; two sets of scores are obtained, one for aptitude and one for proficiency; then correlational and regression analyses are conducted to ascertain whether aptitude is predictive of learning outcomes. In these studies, the predicting or independent variable is aptitude or aptitude components, and due to the lack of control over the instruction learners receive and other factors, it is advisable to investigate the unique and joint contributions of aptitude and other variables that potentially influence learning outcomes, such as learners' cognitive (e.g., working memory and intelligence), conative (e.g., motivation), affective (e.g., anxiety), and demographic (e.g., socioeconomic background) variation. The dependent variables of these studies include overall L2 proficiency operationalized as course grades or scores on standardized proficiency tests; aspects of L2 knowledge such as grammar, pronunciation, and vocabulary; or L2 skills such as listening, speaking, writing, and reading. Li's meta-analysis (2016) showed that, despite the seemingly large amount of correlational research examining the predictive power of aptitude for L2 attainment, when the research was divided into subcategories based on aptitude components and specific aspects of L2 learning, the number of studies in each subcategory was mostly below ten.

The need for more research on aptitude and specific aspects of learning has been responded to by recent research on pronunciation. These studies have exclusively examined the associations between cognitive abilities for pronunciation learning and L2 speech performance (Bowles et al., 2016; Saito et al., 2019; Smemoe & Haslam, 2013). In these studies, the predictor variables are phonetic coding (measured by LLAMA_D, MLAT_2, PLAB_5 and _6), musical aptitude, and implicit auditory encoding (measured via FFR and LLAMA_D). Saito et al. (2019) classified the cognitive abilities into two categories: explicit and implicit pronunciation learning aptitude, with phonetic coding and musical aptitude falling into the former category and implicit auditory encoding into the latter (see Table 1 for a list of aptitude tests). Within explicit and implicit pronunciation aptitude, Saito et al. further posited cognitive abilities for segmental (sounds) and suprasegmental (stress, rhythm, etc.) learning. The dependent variables of the studies are typically measures of L2 pronunciation such as accentedness, comprehensibility, and fluency. Correlational research is a product-oriented approach contributing evidence on the role of aptitude in influencing learning success regardless of instruction type and the process in which learning occurs. However, the results of such research need to be interpreted in consultation with the process, namely in the context of the type of instruction learners received. The participants of instructed learning research are typically classroom learners that receive traditional form-oriented instruction, which is more likely to show significant effects for explicit aptitude than implicit aptitude, because learners in such settings may possess little or no implicit knowledge. One example is Roehr-Brackin and Tellier's (2019) study, which showed an increase of aptitude measured using the MLAT, but one caveat is that the learners received form-focused instruction that matches the explicit learning abilities tested by the MLAT. Thus, it remains to be seen whether aptitude increases in other settings

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

such as immersion or naturalistic settings where little or no form-focused instruction is provided. Another example is Li and Qian (in press) who conducted a study with university ESL learners investigating the associations between aptitude and explicit and implicit L2 grammar knowledge. The study showed that learners' test scores on an elicited imitation test hypothesized to measure implicit knowledge loaded on the factor of explicit knowledge, which the researchers attributed to the heavy form-focused instruction the learners received in the local context.

The lack of instructional manipulation in correlational research alleviates the burden on time investment and other resources, making it possible to examine moderating variables for aptitude effects, such as proficiency and instructional settings and to include a larger sample. For proficiency, there is limited research, and the variable has been examined in two ways. One approach is to divide the same cohort of learners into two groups based on the median of the learners' scores on a proficiency test and determine whether aptitude is differentially predictive of the two groups' proficiency (Hummel, 2009). However, this approach reduces the variability of the resultant groups' proficiency and leads to smaller group sizes. Also, the two groups are not matched otherwise (e.g., with variables such as motivation and anxiety), which may confound the results. A more robust and theoretically interesting approach is to recruit learners who are at different stages of learning, such as in different semesters or years of study, and explore whether aptitude's predictive power varies as a function of learning stage (Artieda & Muñoz, 2016; Winke, 2005). This design has theoretical roots in Skehan's (2012) staged model, which posits different roles for different aptitude components at different stages of learning. The design also makes it possible to examine the claims of Skill Acquisition Theory and the Declarative/Procedural Memory Model that traditional/explicit aptitude is important in initial L2 learning, and implicit aptitude is involved in advanced learning. With regard to exploring the impact of instructional settings, one example is Faretta-Stutenberg and Morgan-Short's (2018) study, which examined whether declarative, procedural, and working memory were predictive of L2 Spanish learning in at-home and study abroad contexts. Another example study is Smemoe and Haslam (2013), which explored whether aptitude plays different roles in L2 pronunciation learning in EFL (China) and ESL (US) contexts. However, to date there is limited research comparing whether learners taught using different approaches, such as immersion, task-based instruction, and traditional grammar-based instruction, draw on aptitude in different ways. Finally, due to the lack of control of extraneous variables, correlational research requires a larger sample than experimental research. One example large-sample study is Kiss and Nikolov (2005), which involved more than four hundred sixth graders from ten primary schools.

Experimental Research

Experimental research examines the interaction between aptitude and the process and outcome of carefully manipulated instructional treatments. This stream of research is rooted in an interactional approach to aptitude where the role of aptitude varies between learning conditions, because of the different processing demands imposed on the learner (Robinson, 2011). In a typical experimental study, learners are divided into groups and receive different types of instructional treatments focusing on the learning of a particular linguistic target. Learners are tested before and after the treatments to measure treatment effects, and statistical analyses are conducted on learners' test scores to (1) compare the effects of different treatments, and (2) ascertain whether aptitude

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

Table 1 Major Tests of Explicit and Implicit Aptitude

Construct	Tests	Description
Explicit Aptitude	MLAT	**Part 1 Number Learning**. Measures memory and auditory alertness. Learners are asked to learn numbers in a new language. **Part 2 Phonetic Script**. Measures phonetic coding. Learners learn sound-symbol associations. **Part 3 Spelling Clues**. Measures English vocabulary and phonetic coding. Learners answer questions about English vocabulary. **Part 4 Words in Sentences**. Measures grammatical sensitivity. Learners are asked to identify linguistic functions of sentence elements. **Part 5 Paired Associates**. Measures rote memory. Learners are asked to memorize words and their meanings.
	LLAMA	**LLAMA_B**. Measures rote memory. Learners are asked to memorize the associations between shapes and sound combinations. **LLAMA_D**. Measures phonetic recognition. Learners listen to some syllables and then discriminate between old and new syllables. **LLAMA_E**. Measures sound-symbol associations. Learners memorize symbols and their pronunciations. **LLAMA-F**. Measures inductive learning ability/language analytic ability. Learners see pictures and sentences and learn grammar rules.
	PLAB	**Part 1 GPA**. Learners report their GPAs. **Part 2 Motivation**. Learners report the intensity of their interest in learning the foreign language. **Part 3 Vocabulary**. Learners are tested on their English vocabulary. **Part 4 Language Analysis**. It measures inductive learning ability (language analytic ability). Learners learn rules of an artificial language and then answer questions applying the rules. **Part 5 Sound Discrimination**. It measures the ability to distinguish sounds. Learners are asked to learn three words in another language that sound similar. They then hear short sentences and pick the word that appeared in each sentence. **Part 6 Sound-Symbol Association**. It measures phonetic coding. Learners hear nonsense words and choose the correct spelling of the word based on their knowledge about English pronunciation.
	Hi-LAB	**Executive control** **Updating**. Measured using a running memory span test where learners are presented with digit strings and recall the last X number of digits. **Inhibition**. Measured using an Antisaccade test and a Stroop test. In the Antisaccade task, learners are asked to look in the opposite or same direction of a target. The score is the difference between the reaction times for the same and opposite conditions. In the Stroop test, learners respond to words for three

		colors: red, green, and blue. For some words, the color and the word are congruent; for others, they are incongruent. Learners are asked to ignore the meaning of the word and only respond to the color. Scores are based on reaction time differences between congruent and incongruent words. **Shifting**. Measured using a switching test where learners are asked to press a button in response to odd digits and another to even digits. The score is the reaction time difference between the switching and non-switching trials. **Phonological short-term memory**. Measured through a letter span test where learners are presented with lists of letters and asked to recall, and through a non-word test in which learners see lists of nonwords, after each list, they are shown another list and asked to recognize which words are on the list just presented. **Associative memory**: this is the MLAT_5; see above. **Long-term memory**: the test is a sematic priming task which includes a prime followed by a target. In the prime, learners listen to five words, after which they are shown two topic words, one of which is synonymous with two and the other with three of the five words. They are asked to indicate which has more synonyms. In the target, they are presented with a pair of words and judge whether the two words have similar or different meanings. The target trials are of two kinds: one is called primed, where one or both words in the pair are synonymous with one of the two topic words in the preceding prime; the other is called unprimed, during which condition none of the two words is a synonym of the topic words in the prime. Scoring is based on response time and accuracy for the primed and unprimed conditions. **Sequence learning**: this is measured by an SRT task. See below. **Processing speed**: this is also measured through the SRT; it's simply the learner's response rate during the SRT. **Auditory perceptual acuity**: this is measured by using two tasks—phonemic discrimination and phonemic categorization—asking learners to discriminate and categorize sounds, respectively.
	Musical ability	These are specifically for pronunciation learning. **Comprehension**: learners hear pairs of musical notes (chords and tunes) and decide whether they are different or the same (Li & DeKeyser, 2017; Saito et al., 2018) **Production**: learners hear a tune and reproduce it in singing such as "la la…" (Li & DeKeyser, 2017)
Implicit Aptitude	Serial reaction time (SRT)	Learners respond to a dot that appears in one of four locations on the computer screen. The locations are based on a target sequence that appears more frequently and a control sequence that appears less frequently. Scores are the differences between reaction times for the target and control sequence.
	LLAMA_D	This is the LLAMA_D test explained above. However, to make it a test of implicit aptitude, the test needs to be adapted and administered differently (Suzuki, in press).

(*Continued*)

Table 1 (*Continued.*)

Construct	Tests	Description
	Statistical learning	Learners listen to nonsense syllables or see arbitrarily combined shapes repeatedly and are then tested by recognizing familiar and unfamiliar items (Godfroid & Kim, in press; McDonough & Trofimovich, 2016; Porter et al., 2017).
	Priming	Priming is of two types: semantic and syntactic. **Semantic priming**. See the above part on long-term memory in the description of the Hi-LAB. **Syntactic priming**. Learners listen to a sentence, repeat it, and then describe a picture. In the picture description, they are expected to use the linguistic structure (e.g., passive voice) in the sentence they listened to rather than the alternative (active voice).
	Process control	These refer to tests of procedural memory in the Declarative/Procedural model. These tasks ask learners to reach or maintain a certain goal by repeatedly responding to stimuli embedded with rules. Example tasks include Tower of London, Sugar Production, and Weather Prediction. In Tower of London, learners move three balls around from an initial configuration to reach a goal configuration. In Sugar Production, learners maintain the production range of a sugar factory by manipulating the number of workers hired and taking into consideration the previous production of each trial. In Weather Production, learners guess the weather condition based on combinations of tarot cards
	Frequency following response (FFR)	Learners listen to a synthesized speech syllable /da/ repeatedly wearing an electroencephalogram (EEG) device to capture their electrophysiological responses. Neural encoding of pitch and speech formants are calculated that are hypothesized to underlie the production of suprasegmental (stress, rhythm, fluency, etc.) and segmental (accuracy of pronunciation) features of L2 speech.

has differential associations with the effectiveness of different treatments. While some studies examine both the comparative effects of treatment types and aptitude-treatment interaction (Benson & DeKeyser, 2018; Kourtali & Révész, 2020), other studies focus on the results on aptitude-treatment interaction and report only the descriptive statistics for treatment effects to give the reader an overall picture of the larger project (Erlam, 2005; Li et al., 2019).

Two approaches to data analysis have emerged from the research to investigate aptitude-treatment interaction. One approach is to treat aptitude as a continuous variable and perform correlation or regression analysis for each treatment group to explore whether aptitude is predictive of learners' post-test or gain scores in different ways (Li et al., 2019; Shintani & Ellis, 2015). The other approach is to treat aptitude as a categorical variable by dividing learners into high and low aptitude based on a cutoff point such as the median (Yilmaz, 2013), two standard deviation units above or below the mean (Benson & DeKeyser, 2019), or percentile ranges (Dahlen & Caldwell-Harris, 2013). This approach allows the researcher to show whether high- and low-aptitude learners benefit from different types of instruction. However, categorical/dichotomous data are statistically less robust than continuous data, and the validity of this approach depends on whether learners falling into high- and low-aptitude groups are matched along other variables, such as previous knowledge of the target structure.

Another consideration is whether to use gain scores (calculated by subtracting pretest scores from posttest scores) or posttest scores as the dependent variable. The recommendation is to use posttest scores, because they are raw scores with unchanged variance and standard deviations. Gain scores have been criticized for having low reliability (Cronbach & Furby, 1970) and may lead to uninterpretable or unexpected findings because of the artificiality associated with them. Regardless of whether posttest or gain scores are analyzed as the dependent variable, it is important to include pretest scores or previous knowledge as a predictor or covariate, because previous knowledge has been found to be the most consistent and powerful predictor of learning gains when entered into the same prediction model with other variables (Li et al., 2019; Yalçın & Spada, 2016). Including pretest scores as a predictor makes it possible to determine whether aptitude explains a unique portion of the variance of the outcome variable after the variance explained by previous knowledge is excluded. Experimental aptitude research has examined aptitude's associations with a variety of instruction types, which are summarized in the sections that follow. The purpose of the summary is to inform the reader of the major treatment types that have been investigated or possible ways to manipulate instructional treatments.

Corrective feedback

Corrective feedback refers to responses to learners' errors in L2 comprehension and production activities. Corrective feedback has been at the forefront of SLA research over the past two decades, because of its theoretical and pedagogical significance (Mackey, 2020). Notwithstanding, most existing research concerns oral feedback in face-to-face interaction, and studies on written feedback and computerized feedback are few and far between. There has been much research on the relationship between explicit and implicit aptitude and different types of oral feedback, such as implicit and explicit feedback (Li, 2013; Yilmaz & Granena, 2019). Implicit feedback does not overtly draw learners' attention to errors and is typically operationalized as recasts, which refer to reformulation of an erroneous utterance without altering the meaning.

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

Explicit feedback alerts the learner to the presence of an error and may be provided in the form of explicit correction or metalinguistic feedback. There has also been research, albeit limited, on whether language analytic ability mediates the effects of written feedback, which can be divided into direct, indirect, and metalinguistic feedback, depending on whether an error is corrected by providing the correct form, identifying the error without correcting, or making comments on the nature of the error (Benson & DeKeyser, 2019). Finally, research has also been conducted on computerized feedback (Lado, 2017).

Deductive and inductive instruction

In deductive instruction, learners are presented with rule explanations followed by practice activities where rules are applied. In inductive instruction, learners are required to discover rules based on given materials. In aptitude research, Erlam (2005) and Hwu et al. (2014) are probably the only studies that systematically manipulated deductive and inductive instruction, despite the fact the two instruction types were implemented in different ways in the two studies. More research is warranted to demonstrate how learners with different aptitude profiles (with high and low aptitude) benefit from the two different types of L2 instruction that are frequently used in L2 classes.

Spaced versus massed practice

The distribution or spacing of practice refers to the interval between practice sessions/activities, and how the distribution of practice impacts the way learners deploy their cognitive resources is a cutting-edge topic in both psychological and SLA research. Two aptitude studies (Kasprowicz et al., 2019; Suzuki & DeKeyser, 2017) examined language analytic ability's interface with the distribution of practice activities. In spaced practice, the interval between practice sessions is longer, such as seven days, while in massed practice the interval is shorter, such as only one day. One methodological difference between the two studies is that, in Suzuki and DeKeyser's study, the only difference between the two treatment conditions is the interval between practice sessions, while in Kasprowicz et al.'s study the two treatment conditions are also different in the length of each practice session and the number of practice sessions.

Input manipulation

This may involve manipulating input distribution and input processing. Brooks et al. (2016) investigated whether implicit aptitude (called statistical learning) mediated the effect of input distribution, which was manipulated by creating two treatment conditions: balanced input and skewed input. In balanced input, L1 English speakers were exposed to three Russian structures that were equally distributed among eighteen nouns, while, in skewed input, the structures were used with the eighteen nouns with varied frequencies. VanPatten and Borst (2012) examined whether language analytic ability was correlated with the effects of input processing instruction during which L1 English and L2 German learners answered comprehension questions that forced them to adapt the "first noun" principle, which states that the first noun of a sentence is the subject. While the "first noun" principle is true of SVO sentences, it is not applicable to OVS sentences, which are typical in German. In addition to input distribution and input processing, there are other ways to manipulate the type of input learners receive, such as input enhancement and input modification. In input enhancement, certain features of textual materials are made salient, such as by highlighting or using bold type. In input modification, input material is adapted to make it more comprehensible.

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

Input enhancement and modification, as well as the variants within each type of manipulation, may pose different processing demands on learners and influence the way learners deploy their cognitive abilities.

Incidental learning

Studies conducted in this paradigm are based on the declarative/procedural memory model, according to which early stages of learning involve declarative memory (explicit aptitude), and later stages involve procedural memory (implicit aptitude) (Ullman & Lovelett, 2018). In these studies, learners engage in meaning-based comprehension and production activities, such as playing a game and describing moves during a game (Morgan-Short et al., 2014), deciding whether given sentences are difficult or easy to understand (Hamrick, 2015), or matching pictures and their descriptions (Walker et al., 2020); learners are not told to attend to any particular linguistic structures. Typically, the treatments are conducted in artificial or semiartificial languages and involve a vocabulary learning stage followed by a practice or training stage.

Task complexity

Several studies examined the relationship between aptitude and simple and complex tasks based on Robinson's (2011) model of task-based instruction—the Cognition Hypothesis. A central tenet of the Cognition Hypothesis is that the impact of cognitive abilities on task performance is more evident in complex than simple tasks, because the former impose heavier cognitive pressure on the learner. By way of illustration, Kormos and Trebits (2012) examined the relationship between aptitude and task complexity, which was operationalized as whether learners are provided with a storyline that accompanied a picture set (simple task) or given unrelated photos (complex task) when telling a narrative. Kourtali and Révész (2020) investigated the association between aptitude and the effects of corrective feedback (recasts) in simple and complex task conditions, which involved information transmission or decision-making, respectively.

Linguistic target

Several studies investigated the nature of the linguistic target as a mediator for the association between aptitude and treatment effects. Yilmaz and Granena (2019) investigated whether implicit and explicit aptitude are involved in implicit and explicit feedback in the learning of two linguistic targets: Spanish gender agreement and differential object marking, which differ in whether they involve agreement between linguistic elements. Li (2013) examined whether analytic ability and working memory have differential associations with the effects of implicit and explicit feedback in the learning of two linguistic structures—Chinese classifiers and the perfective *-le*—that differ in transparency of form-meaning mapping. Yalçın and Spada (2016) explored whether memory and analytic ability played different roles in the learning of the English passive (a difficult structure) and the past progressive (an easy structure) in form-focused instruction. It is noteworthy that grammar structures have been classified in various ways in aptitude and SLA literature. For example, DeKeyser (2000) classified L2 English structures as perceptually salient, such as sentence order and pronoun gender and nonsalient, such as plurals and determiners. Spada and Tomita (2010) categorized grammar structures as complex and simple linguistic structures, and they defined complexity as the number of steps involved in formulating a rule. Ellis (2006) proposed an overarching framework for assessing linguistic difficulty, and he argued that linguistic difficulty needs to be assessed for explicit and implicit knowledge, separately applying different

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

criteria. In any case, future studies may investigate the impact of the nature of linguistic structures on aptitude-learning associations and/or interpreting the results by considering the possible effect of the linguistic target.

The target languages of the instructional treatments of experimental studies can be divided into two categories: natural and artificial languages. Artificial languages are of two types: full artificial languages that are nonexistent (Morgan-Short et al., 2014; Walker et al., 2020) and semiartificial languages where novel structures or structures from another language are embedded in the learners' native languages (Hamrick, 2015). Studies based on natural languages have higher external validity, because the learning materials are based on real languages, and the results are more applicable to language learning and teaching in the real world. Studies based on artificial languages have higher internal validity, because it is easier to manipulate the type and amount of treatment and minimize the influence of extraneous variables such as learners' previous knowledge (Morgan-Short et al., 2014).

The Nature of Aptitude

Aptitude as a Dependent Variable

One question that is key to an accurate understanding of the nature of aptitude is whether aptitude is subject to age and the amount of language learning experience, in which case aptitude is a dependent or outcome variable. The research has investigated the impact of age in two ways. One way is to give the same aptitude test to learners who differ in age categorically, such as children and adults, and determine whether the aptitude scores of the two age groups are significantly different (Hodel et al., 2014). Another way is to recruit learners who vary in age in a continuum and explore whether the variation of learner age is a predictor of the variation of their aptitude scores (Cox et al., 2019). In both approaches, one confounding variable is experience, that is, older learners may have more language experience, which may lead to higher aptitude. One way to address this issue is to measure experience and include it as a co-variate in statistical analyses (Cox et al., 2019).

To investigate whether experience causes an increase in aptitude, one can track the same group of learners for an extended period during which they receive language instruction to see whether learners' aptitude scores increase after the instruction (Roehr-Brackin & Tellier, 2019). However, the change, if any, may be developmental or due to learners' growth of age rather than learning experience. Therefore, it is necessary to add a between-group component to the design by including a control group that does not receive the instruction the experimental group received (Potter et al., 2017; Sáfár & Kormos, 2008). Alternatively, learners with varied amounts of language learning experience may be recruited and asked to report their learning experience, and self-reported experience can be analyzed as a potential predictor of aptitude scores (Cox et al., 2019). Other forms of data reported by studies focusing on aspects other than the impact of experience on aptitude variation are also revealing about the nature of aptitude. For example, the descriptive statistics reported by Granena and Long (2013), which focused on the relation of aptitude to age effects, showed that early bilinguals' aptitude scores on the LLAMA test were substantially higher than late bilinguals' ($M = 58$ vs. 45), which is likely attributable to the larger amount of learning experience early bilinguals had. However, Dahlen and Caldwell-Harris (2013) reported a somewhat opposite pattern: the aptitude scores of bilinguals on the MLAT were significantly lower than that of monolinguals. The point here is not to discuss the disparities between the

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

findings but to show that the question of whether experience improves aptitude can be approached from different perspectives. R researchers are encouraged to investigate this question experimentally or report and discuss relevant data, even if the primary focus of the study is not on aptitude as a dependent variable.

Aptitude and Other ID Variables

The relationship between aptitude and other individual difference variables is essential to an accurate understanding of the nature of aptitude, because those variables are elements of the nomological network in which the construct is situated (Cronbach & Meehl, 1955). We would like to extend the scope of the nomological network and discuss aptitude's associations with other variables from two perspectives: how aptitude and other cognitive and affective variables are interrelated and how they jointly and independently contribute to learning outcomes. Both perspectives are important for pinpointing the nature of aptitude and assessing its validity. As will be discussed in further detail in later sections, evidence for aptitude's relation with other variables contributes to its content, divergent, and convergent validity, and knowledge about its effects on learning outcomes relative to other variables helps clarify its predictive validity. Aptitude research has rarely focused on the first perspective alone, and most of the findings on its connections to other variables are reported as part of larger projects examining aptitude's predictive validity (Li, 2016). In the research, the most frequently examined covariate of aptitude is working memory (Hummel, 2009; Lado, 2017; Suzuki & DeKeyser, 2015), and other variables examined together with aptitude include learning strategies (Smemoe & Haslam, 2013; Winke, 2013), motivation (Kiss & Nikolov, 2005; Winke, 2013), anxiety (Sparks et al., 2009), and intelligence (Bowles et al., 2016; Kaufman et al., 2010).

Measurement of Language Aptitude

In light of the trend toward distinguishing explicit and implicit aptitude in recent aptitude research, our discussion of aptitude measurement also makes the distinction. Explicit aptitude has been measured via traditional aptitude test batteries such as the MLAT, LLAMA, PLAB, and so on, and their subtests; these tests typically require learners to engage in conscious, effortful information processing. Implicit aptitude, a recent addition to aptitude research, has been measured through tasks where learners are exposed to, or repeatedly respond to, a large number of stimuli created based on rules and regularities unbeknownst to learners. Table 1 displays the details of some major tests of aptitude utilized in the research. In the following sections, we provide an overview of the major tests of explicit and implicit aptitude and discuss the principles of construct validation and their relevance to aptitude tests.

Measures of Explicit Aptitude

Among test batteries of explicit/traditional aptitude, the MLAT is the most influential, the PLAB is intended for teenagers or learners in grades seven through twelve, and the LLAMA is a free, language-neutral test modeled on the MLAT. These tests are hypothesized to measure three cognitive abilities: phonetic coding (MLAT_1, _2, _3; PLAB_5, _6; LLAMA_E), language analytic ability (MLAT_4, PLAB_4, and LLAMA_E), and

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

rote memory (MLAT_5 and LLAMA_B). Phonetic coding refers to the ability to recognize sounds and learn sound-symbol associations. While phonetic coding is a putative ability for pronunciation learning, musical aptitude, such as the ability to discriminate musical notes that differ in melody, pitch, speed, and beat, has also been examined as a predictor of learners' receptive and productive knowledge of L2 pronunciation (Bowles et al., 2016; Hu et al., 2013; Li & DeKeyser, 2017; Saito et al., 2019). Language analytic ability is operationalized as grammatical sensitivity and inductive learning ability. Grammatical sensitivity, which is measurable via MLAT_4, refers to the ability to recognize the grammatical functions of sentence elements. Inductive learning ability refers to the ability to learn rules based on examples and has been tested by using PLAB_4 and LLAMA_F. Rote memory refers to the ability to memorize the associations between words and their meanings, and it is labeled "associative memory" and "declarative memory" in information processing theories.

More information on major aptitude tests is warranted. The MLAT was validated with 5,000 foreign language learners in the US (Carroll & Sapon, 2002), and it has been found to have the highest predictive validity among all aptitude tests (Li, 2015, 2016). It consists of five parts that measure the three components and takes about 60–70 minutes to complete. The PLAB, which was validated with 6,000 language learners, consists of six parts asking test-takers to report their GPAs and their attitudes toward the foreign language they are learning (or motivation) besides measuring their analytic ability and phonetic coding ability; the test leaves out rote memory. However, GPA and motivation are not cognitive abilities for language learning and therefore should be excluded when administering the test in research on language aptitude. The whole test battery lasts about two hours. The LLAMA (Meara, 2005) is probably the most widely used test in recent aptitude research, and its popularity is partly because the MLAT is no longer available to individual researchers. The LLAMA measures similar components as the MLAT, but unlike the MLAT, which is designed for native speakers of English, the LLAMA is language neutral. The whole test takes about thirty minutes to complete, it is freely downloadable, and it is automatically scored. The test battery is comprised of four subtests: LLAMA_B, _D, _E, and _F. Among the four subtests, LLAMA_B, _E, and _F have been found to be separate from LLAMA_D, and when the four subtests load onto the same factor, LLAMA_D shows the lowest factor loading, suggesting that its correlation with the factor is weakest (Bokander & Bylund, 2020; Granena & Long, 2013; Li & Qian, in press). Furthermore, LLAMA_D has also demonstrated attributes of implicit aptitude, although its validity as a test of implicit aptitude has been questioned. Based on available evidence, it would seem justified to use LLAMA_B, _E, and _F, either separately or as a package, to measure explicit aptitude. However, Bokander and Bylund (2020) examined the internal validity of LLAMA and found that, except for the LLAMA_B, the LLAMA subtests showed relatively low internal validity.

The Hi-LAB battery (Linck et al., 2013) was developed to predict high attainment of adult L2 learners. The battery consists of twelve tests measuring both domain-general cognitive abilities, such as working memory, sequence learning, processing speed, and priming, as well as domain-specific abilities such as phonemic acuity. Working memory is operationalized as phonological short-term memory and executive control, with the former referring to the storage component of short-term memory or the phonological loop, and the latter to the central executive. The central executive comprises of three components: inhibition (the ability to suppress irrelevant information); updating (the ability to monitor an ongoing event, deleting previous information and adding new

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

information); and switching (the ability to shift between different tasks). Semantic priming measures long-term memory, which refers to the tendency to be influenced by an event that happened in a previous encounter. For example, after seeing a word, the learner would respond faster when seeing a related word than an unrelated word. Sequence learning and processing speed are measured via a serial reaction time task, a measure of implicit aptitude, which will be elaborated below. Essentially, the Hi-LAB is based on a domain-general view of language aptitude, which refers to any cognitive ability that is predictive of L2 attainment, irrespective of whether the ability has a theoretical link to the processing and learning of linguistic stimuli. However, research has shown that language aptitude is domain specific on the grounds that the cognitive abilities for language learning are distinct from domain-general cognitive abilities, such as working memory (Yalçın et al., 2016) and intelligence (Li, 2016).

Measures of Implicit Aptitude

Serial Reaction Time (SRT)

SRT measures sequence learning, which refers to learners' sensitivity to rules governing the way symbols are sequenced. SRT is probably one of the most frequently used tasks of implicit learning in major paradigms of implicit learning, such as procedural memory and statistical learning. SRT has also proven to be a consistent predictor of learning outcomes. In a typical SRT test, learners respond to a symbol, such as a black dot, a smiley face, etc., that appears at four locations. The locations where the symbol lands are based on two sequences, such as 121432413423 and 124314213234. One sequence can be called a regular sequence, which appears more frequently in the stimuli (e.g., 85%), and the other a control sequence, which is less frequent (e.g., 15%). With repeated practice, learners' responses to the target sequence will be faster than the control sequence, and their test performances are calculated by subtracting their mean reaction time for the target stimuli from the mean reaction time for the control stimuli. Regarding the presentation of the target and control sequence, there are two variants: interleaved and separated. In interleaved presentation, the target and control sequence alternate in the stimuli throughout the test in all blocks (Li & Qian, in press; Suzuki & DeKeyser, 2015). In separated presentation, the two sequences are presented in separate blocks: only the final block is based on the control sequence and all preceding blocks are based on the target sequence (Hamrick, 2015; Walker et al., 2020). The scoring of the two forms of SRT is also different: in interleaved presentation, the scores are based on all stimuli, whereas in separated presentation, scoring is based on the final (control) and the penultimate (target) blocks. In addition to the distribution of target and control sequences, SRT tests vary in other aspects, such as the number of items, the choice between the median and mean score for reaction time calculations, etc. It is unclear whether the two forms of SRT and methodological variation in other aspects of the test cause any difference in the validity of the test.

LLAMA_D

In this test, learners listen to some nonexistent sound sequences in the exposure phase and are asked to recognize old and new sequences during the testing phase. Evidence for LLAMA_D as a measure of implicit aptitude comes from Granena (2015), which shows that LLAMA_D and SRT underlay one factor labeled as implicit aptitude and

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

the remaining three subtests of the LLAMA test battery loaded under the same factor, named explicit aptitude. The implicit and explicit factors were correlated with an implicit-intuitive learning style and an explicit-analytic learning style, respectively. Other studies also showed that LLAMA_D measures a cognitive ability that is separate from the three other subtests of the LLAMA test battery, because it either loads onto a separate factor than the three subtests (Artieda & Muñoz, 2016; Bokander & Bylund, 2020; Suzuki, in press) or has a weaker factor loading than other subtests of the LLAMA when loading onto the same factor (Li & DeKeyser, in press). LLAMA_D is also predictive of L2 oral production (Granena, 2019; Saito et al., 2019), which likely implicates implicit knowledge rather than explicit knowledge (R. Ellis, 2005). However, LLAMA_D has been found to be an inconsistent measure of implicit aptitude because there has been evidence that it loaded with explicit aptitude rather than implicit aptitude (Li & Qian, in press), and it is predictive of explicit knowledge instead of implicit knowledge (Granena, 2013). Therefore, caution should be exercised when using LLAMA_D as a measure of implicit aptitude, and steps may be taken to increase its likelihood of measuring implicit aptitude, such as by asking learners to check the sound volume instead of focus on the auditory stimuli to encourage incidental learning (Saito et al., 2019; Suzuki, in press).

Priming

Priming refers to the tendency to be influenced by, or to use, a linguistic structure or item exposed to in a previous encounter in a similar event or situation. Priming is of two types: semantic and syntactic. An example semantic test is the test of long-term memory in the Hi-LAB (Linck et al., 2013), where each trial consists of two items: a prime followed by a target. In the prime, learners see a pair of words and five other words on a separate list; they judge which word in the pair has more synonyms among the five words. In the target, learners are presented with a pair of words and judge whether they are synonymous. There are two types of trials: primed and unprimed. In primed trials, one of the words in the target trial appeared in the prime; in unprimed trials, the prime and target items are unrelated. Learners will respond faster in primed trials than unprimed trials. Similarly, in syntactic priming, each item consists of a prime followed by a target. For the prime, leaners listen to and repeat a sentence consisting of a certain structure, and for the target, they describe a picture showing a similar event. It is hypothesized that, in the picture description, learners are more likely to use the structure contained in the prime rather than an alternative. For example, after hearing the sentence "The father gave a present to his daughter," one is more likely to say, "The teacher gave a book to the student," even though "The teacher gave the student a book" is equally correct. As with other paradigms of implicit learning, priming has been used to examine the process or outcome of learning, not as a predictor of learning (but see Li & Qian, in press).

Tests of Process Control

These tests are used to measure procedural memory in the Procedural/Declarative model (Ullman & Lovelett, 2018). The tests require learners to maintain a certain target or output by manipulating an input variable, and there are rules behind the relationships between the input and output variables. The most frequently used tests include Weather Prediction, Tower of London, and Sugar Production. In Weather Prediction,

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

learners make predictions about weather conditions (rain or sunshine) based on different configurations of "tarot cards," created following certain regularities. In Tower of London, learners move three balls around among three pegs based on certain rules. In Sugar Production, learners maintain the levels of the sugar production of a factory by varying the number of workers and considering the sugar production of the previous trial; unbeknownst to learners is that the variation of sugar production is based on a certain formula.

The Validation of Aptitude Tests

Aptitude is a cluster of cognitive abilities represented by learners' scores on the tests used to measure aptitude, and, to some extent, aptitude is defined by what is measured by aptitude tests. The validity of research findings and justifiability of the practical uses of aptitude scores relies crucially on the validity of aptitude tests. Validity refers to whether a test accurately measures what it is intended to measure and whether "evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA, & NCME, 2014, p. 11). Applied to language aptitude, the notion of validity refers to the meaningfulness and accuracy of all the interpretations of the test scores, and the validity of the construct is assessed by whether there is evidence for the theoretical claims about aptitude's attributes, components, and relationships with covariates (e.g., other cognitive and affective variables) and with learning outcomes. The validity of psychometric tests can be examined along dimensions of reliability, content validity, predictive validity, and convergent and divergent validity.

Reliability

Reliability refers to whether candidates' performance on a test is consistent. Reliability is a prerequisite for validity, because a valid test must be reliable, so reliability constitutes one aspect of validity. Reliability is indexed by test-retest reliability, which refers to whether candidates' scores on different administrations are correlated, or by internal consistency, which refers to whether candidates' answers to different test items in one administration are correlated. Test-retest reliability requires the same learners to take the same test twice, which is logistically challenging, and, because of the challenge test-retest reliability poses, aptitude researchers typically report internal reliability. Aptitude research has reported high or at least acceptable reliabilities for tests of explicit aptitude, in the range of .7 and .9 (Carroll & Sapon, 2002; Yi, 2018), but measures of implicit aptitude are typically low, falling between .4 and .6 (Kaufman et al., 2010). Thus, it would seem necessary to adopt different criteria for the evaluation of reliability indices for the two types of aptitude: while the recommended threshold of > .70 (Field, 2013) may be applied for tests of explicit aptitude, a lower threshold such as > .4 may be set for tests of implicit aptitude (Granena, 2013). Strategies to increase reliability include adding more test items, deleting items with low item discrimination (correlations between scores for individual items and total test scores), ensuring appropriate item difficulty (percentage of candidates who answered a test item correctly), removing irrelevant and/or misfitting items, using items requiring controlled responses, providing clear and specific instructions, excluding items that are subject to multiple interpretations and/or multiple answers, targeting one (not multiple) point/structure/skill in each item, applying the same scoring criteria consistently, and so on (Bokander & Bylund, 2020). It must be pointed out that high reliability is necessary but insufficient for a

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

valid test, and that a reliable test can be invalid if it does not accurately measure what it purports to measure, for example, if the test does not measure the right content—topic of the next section.

Content Validity

Content validity refers to the relevance and scope of the knowledge and skills measured by a test, with relevance referring to whether the content is important and scope to whether all relevant content is covered. The content of a test of language aptitude should be based on two sources: (1) the theoretical conceptualization of what aptitude entails, and (2) the observation of the mechanism of L2 learning and the role aptitude plays in the mechanism. Tests of traditional aptitude are primarily based on observations of "language courses that were then [1950s] being conducted in schools, colleges, and governmental or military organizations" (Carroll, 1981, p. 89). These courses, according to Carroll, were based on the "audio-lingual methods" (p. 87) characterized by rote learning and mechanical practice. Carroll's observations of how language learning happens in audiolingual classes led him to hypothesize a three-component structure for language aptitude consisting of phonetic coding ability, analytic ability, and rote memory, which roughly, albeit not perfectly, correspond with pronunciation, grammar, and vocabulary, respectively. The observations also influenced the format of the items of the MLAT, where three out of the five subtests—I, II, and V—are learning tasks or work samples of language learning. Thus, the content of traditional aptitude tests, which are represented by the MLAT, is primarily based on observations of how languages are learned in the classroom, not theoretical conceptualization. Also, the content validity of traditional aptitude has been criticized for its pure focus on the formal (as opposed to meaning) aspects of L2 learning and its failure to include pragmatic competence—the ability to learn how language use varies between contexts—and the ability for L2 production skills (Li, 2016). Therefore, traditional aptitude tests suffer from construct underrepresentation, that is, the measured content is insufficient, despite the relevance of the existing content to language learning. The construct underrepresentation of traditional aptitude tests is exacerbated by the lack of a measure of implicit aptitude, given that traditional aptitude concerns primarily explicit learning.

The content of implicit language aptitude has received little theoretical and empirical scrutiny in the literature. The tests are borrowed intact from psychological research, and the extent to which the measured content is relevant to language learning needs clarification. Tests used to measure implicit aptitude, such as SRT, Weather Prediction, Tower of London, and so on, gauge learners' sensitivity to the distributional and transitional probabilities of available stimuli. Distributional probability refers to the frequency of an event, that is, the number of times an event occurs, while transitional probability refers to contingency or mutual dependency, namely, how likely the occurrence of one event can be predicted by another event. These two cognitive abilities can be collectively called sequence learning, which, according to N. Ellis (2005), is characteristic of language learning in that language learning is fundamentally a process of learning sequences of sounds, words, and phrases. However, language learning not only involves learning the rules and regularities governing how linguistic elements are configured and sequenced but also how linguistic forms are connected to meaning. Form-meaning mapping is missing in current measurement of implicit aptitude, although it is unclear whether sequence learning is sufficient and whether form-meaning mapping is necessary. Aside from sequence learning, one other essential

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

component of implicit aptitude is selective attention, which refers to the ability to register and detect relevant input (N. Ellis & Wulff, 2015; Gass, 2018; Kaufman et al., 2010; Long, 2015). As the term suggests, the function of selective attention is to select, not to store or process, input materials, and, therefore, it involves minimal levels of awareness and does not require effortful, conscious processing. Once input is selected, it will be processed implicitly thereafter. Regarding test stimuli used to measure the content of aptitude, tests of implicit aptitude also involve learning tasks, but these tasks are not authentic learning tasks that are typical of tests of explicit aptitude such as the MLAT and LLAMA.

Predictive Validity

Predictive validity refers to whether aptitude is predictive of, or correlated with, learning outcomes represented by course grades or scores on tests of L2 proficiency or achievements. In this sense, all studies examining whether aptitude is correlated with learning gains contribute evidence on aptitude's predictive validity, which may take the form of its associations with ultimate attainment regardless of context or with the effects of different instructional treatments, as discussed in earlier sections. Predictive validity is crucial for the validity of an aptitude test, because one core attribute of this cognitive trait is its importance for learning success. The importance of predictive validity is evident in the validation studies examining the validity of newly developed aptitude tests. For example, in the manual of the MLAT, Carroll and Sapon (2002), who developed and validated the test battery, stated that "the primary purpose of the MLAT is prediction" (p. 23), and in the section on "validity" (p. 12), they only reported the correlations between MLAT scores and course grades without addressing other types of validity. Kiss and Nikolov (2005) adapted the MLAT and developed a version of the test for young L1 Hungarian learners of English. The primary objective of their validation study was to examine aptitude's associations with learners' L2 proficiency. Similarly, in the validation studies for the Hi-LAB, only predictive validity was examined (Doughty, 2019; Linck et al., 2013). Despite the importance of predictive validity in construction validation, we would like to argue that predictive validity is insufficient for construct validity and that simply finding significant correlations between aptitude scores and learning outcomes is not enough justification for the validity of the construct. For example, the PLAB includes motivation and GPA as components of aptitude, but motivation is an affective variable and GPA is a proxy of general academic achievements. Therefore, although PLAB scores may show significant correlations with learning outcomes, the correlations do not constitute robust evidence for the validity of language aptitude, because they do not fully and accurately represent the nature of aptitude, which is conceptualized as a cognitive rather than affective variable and as a predictor of achievements rather than achievements per se (Carroll, 1981). Thus, the PLAB may have predictive validity, but it lacks content validity.

Convergent and Divergent Validity

Convergent validity refers to whether measures of similar or related traits are correlated with each other, and evidence for the convergent validity of a construct falls into two types: internal and external. An internal perspective is based on the notion that the components of the same construct should be correlated with each other, because they share common variance, and they comprise a higher-order overarching latent

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

variable. However, the components should also be independent of each other, that is, the correlations should not be strong to the extent that they are not separable. For example, the three components of traditional aptitude—phonetic coding, analytic ability, and rote memory—should be correlated and yet separable. An external perspective can be understood in two ways, depending on whether the point of comparison is another test of the same construct or a test of a different construct in the same paradigm. To illustrate the former standpoint, both the LLAMA and the MLAT are claimed to test language aptitude, and therefore learners' scores on the two tests should be highly correlated. In the case of a different construct in a similar paradigm, the standpoint goes beyond the construct in question and predicts that tests of the construct should be correlated with other constructs hypothesized to belong to the same paradigm. For example, traditional aptitude taps into explicit learning and therefore should be correlated with other cognitive abilities for explicit learning, such as working memory and intelligence. While convergent validity relates to the relationships between a construct's components and between a construct and its "kinship" constructs, divergent (alternatively named "discriminant") validity means that a construct should be different from, or uncorrelated with, constructs that are purportedly different from the construct in point or that belong to another domain based on theoretical conceptualization. For example, if we hypothesize that explicit and implicit aptitude are fundamentally different, then test scores for the two should be uncorrelated, or even negatively correlated, if there is theoretical ground for such a prediction. Convergent and divergent validity are important because the nature of a construct is defined and represented by its relationships with variables that it is similar to—evidence for what it is—as well as variables that it is different from—evidence for what it is not.

Do aptitude tests have convergent and divergent validity? Tests of explicit aptitude have demonstrated consistent convergent and divergent validity. In terms of convergent validity from an internal perspective, subtests of batteries of traditional aptitude such as the MLAT (Carroll & Sapon, 2002) and the LLAMA (except LLAMA_D) (Granena, 2019; Li & Qian, in press) are typically significantly correlated with each other, but the correlations (r's < .60) are not high enough to consider the tests as measures of the same component. The subtests also tend to load onto the same factor, suggesting that they share common variance (Granena, 2019; Hummel, 2009). However, there is a lack of cross-validation research investigating the extent to which different test batteries, such as the MLAT, the LLAMA, and the PLAB, measure the same construct—aptitude. In terms of convergent validity from an external perspective, explicit aptitude has been found to be significantly correlated with, but different from, other cognitive abilities for explicit learning, such as working memory and intelligence (Li, 2016; Yalçın et al., 2016; Yi, 2018). For divergent validity, explicit aptitude has been found to be uncorrelated with motivation and negatively correlated with anxiety, suggesting that motivation and anxiety are two affective variables that are fundamentally different from language aptitude (Li, 2016). With regard to implicit aptitude, one striking finding is its lack of convergent validity, that is, measures purporting to measure implicit aptitude, such as SRT, Weather Prediction, Tower of London, syntactic priming, and so on are uncorrelated or even negatively correlated (Godfroid & Kim, in press; Li & Qian, in press; Buffington et al., in press). The findings suggest that implicit aptitude is likely a multicomponential construct, which consists of abilities that are different from each other but that may all contribute to L2 learning. Tests of implicit aptitude, however, have demonstrated divergent validity, the evidence being that (1) they are uncorrelated with, or load under different factors from, tests of explicit aptitude (Granena, 2019; Li & Qian, in press; Yi, 2018), and (2) learners

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

with cognitive impairments in one domain (implicit or explicit) have normal cognitive abilities in the other (Li & DeKeyser, in press).

Measurement of Outcome Variables

In aptitude research, the dependent variable is typically L2 proficiency or learning outcomes, although aptitude itself can also be a dependent variable, as discussed in earlier sections. Proficiency measures can be unfocused or focused. Unfocused measures do not target a particular structure and are typically employed in correctional research that aims to examine aptitude's predictive power for general L2 proficiency; domains of L2 knowledge such as grammar, vocabulary, and pronunciation; or L2 skills such as listening, reading, writing, and speaking. Focused measures are typically used in experimental research and aim to gauge learning gains resulting from instructional treatments targeting a particular linguistic feature, be it grammatical, phonological, or lexical (see Supplementary Information for example outcome measures). Among L2 outcomes, grammar is probably the most researched in aptitude research, and one emerging trend is the distinction between explicit and implicit grammar knowledge. Explicit knowledge is conscious, analytic, verbalizable, and accessible through effortful retrieval; implicit knowledge is unconscious, intuitive, unverbalizable, and automatic. Explicit knowledge has been measured by means of untimed grammaticality judgment, error correction, metalinguistic knowledge, multiple choice, and so on, which encourage learners to consciously access, process, and apply grammar rules (Kasprowicz et al., 2019; Lado, 2017; Sheen, 2008; Yalçın & Spada, 2016). Implicit knowledge has been tested via elicited imitation, word monitoring, self-paced reading, timed grammaticality judgment, and free oral production (Granena, 2013; Yalçın & Spada, 2016). These tests are typically administered under time pressure to limit conscious processing and to encourage automatic retrieval of linguistic knowledge. To minimize attention to forms, some of the tests include a meaning component, such as asking learners to answer comprehension questions or make plausibility judgments. Tests of implicit knowledge can be classified into time-based and accuracy-based, depending on whether better performance is represented by faster reaction time or by higher accuracy in the comprehension or production of L2 structures. Based on this taxonomy, word monitoring and self-paced reading, and are time-based, and elicited imitation, oral production, and timed grammaticality judgment are accuracy-based. Performance on time-based measures is typically calculated by subtracting learners' average reaction time for grammatical sentences from their average reaction time for ungrammatical sentences. Another index that has appeared in recent aptitude research is coefficient of variance (CV), which is calculated as the ratio between the standard deviation of reaction time and mean reaction time for correct responses (Suzuki, 2018). A decrease in CV represents faster reaction time as well as lower variability of learning outcomes. CV is typically used as a measure of automatization and restructuring as a result of practice. A list of representative measures of explicit and implicit knowledge is provided in Supplementary Information, with the caveat that, whether certain measures tap explicit and implicit knowledge, needs further empirical verification. For example, whereas R. Ellis (2005) identified elicited imitation as a measure of implicit knowledge, Suzuki and DeKeyser (2015) found it to be a measure of automatized explicit knowledge. However, the disparity is likely attributable to the different ways the test was administered in the two studies.

Similar to grammar, other forms of outcome have also been measured in various ways. However, unlike grammar, for which the measures have been validated

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

theoretically and empirically (Godfroid & Kim, in press; R. Ellis, 2005), assessments of some other outcome variables need to be more fine-tuned. Vocabulary measures can be classified according to whether receptive or productive knowledge is assessed or whether lexical knowledge is measured in isolation or in context, and the selection of measured words should be based on the purpose of the test or linguistic corpuses. However, in aptitude literature, vocabulary is typically measured using discrete items that tap receptive knowledge, and the selective criteria are often not reported. One type of lexical knowledge that has been tested in aptitude research and that has theoretical basis is knowledge about collocations, which, according to Granena and Long (2013), is representative of L2 lexical competence and subject to maturation effects. Similar to vocabulary, the measures of pronunciation that have emerged in the research can be divided into two categories: receptive and productive knowledge, depending on whether learners are required to understand or discriminate phonetic or phonological features in the L2 (M. Li & DeKeyser, 2017) or to provide speech samples (Saito et al., 2019). Tasks used to elicit speech samples can be categorized as controlled (e.g., reading aloud) and free production (e.g., telling a narrative) tasks based on whether learners are provided with the script of the speaking material or asked to compose a text. Speech samples were rated in terms of accentedness, which refers to whether the L2 speech is nativelike; comprehensibility, which refers to whether it is easy or difficult to understand the speech; and fluency—whether the speech is smooth. Accuracy of pronunciation was judged along segmental/phonetic (sounds) and suprasegmental/phonological (syllables, word or sentence stress, rhythm, etc.) dimensions. While pronunciation concerns the phonetic and phonological aspects of L2 speech, speaking relates to overall speech performance, including both linguistic (vocabulary, grammar, and pronunciation) and content aspects. In aptitude literature, speaking was assessed either subjectively (and holistically) based on human ratings (Smemoe & Haslam, 2013) or objectively based on script analysis where transcripts of speech samples were coded, and the data were calculated for complexity, accuracy, and fluency (Kormos & Trebits, 2012). Reading and listening, which are two comprehension skills, have been typically measured using multiple choice questions based on written or audio texts, but other, atypical assessments, such as translating L2 sentences into L1 and recalling main ideas after listening to a text have also been used. One important distinction that can be made in measuring comprehension skills is between questions that require contextual information and schematic knowledge and questions that rely on local, literal information. The former question type involves primarily top-down processing and the latter bottom-up processing. Certainly, there are other ways to divide comprehension skills, but the point here is to take a more nuanced approach and examine how aptitude relates to specific aspects of L2 comprehension. The recommendation is also applicable to writing, which has been assessed based on holistic ratings, but which can be assessed otherwise, such as by conducting a script analysis in terms of complexity, accuracy, and fluency, similar to the assessment for oral production. Finally, in aptitude literature, overall proficiency has been measured via course grades (Sparks et al., 2012) or standardized tests purporting to measure all skills and aspects of the L2 (Hummel, 2009).

Recommendations

Aptitude research relies heavily on testing, and the value and robustness of research findings depend on whether the tests of aptitude and outcome measures are reliable

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

and valid. Therefore, researchers need to examine or present evidence for the validity of the tests they utilize rather than assuming their validity. Even if a test has been previously validated, the psychometric attributes and measured process/outcome may change due to the changed sample and other idiosyncrasies, such as providing instructions that are (even slightly) different from previous research to test-takers (Suzuki, in press). For the measurement of implicit aptitude, researchers are encouraged to prioritize SRT given its content and predictive validity found in previous research, but caution should be exercised when using tasks of process control, such as Weather Prediction, Tower of London, and Sugar Production as measures of implicit aptitude (Buffington et al., in press; Godfroid & Kim, in press). When using the SRT, researchers should be aware of the two variants of SRT: interleaved and separated (see the section on aptitude measures), and the different scoring schemes. The methodological variation may impact the results. Any measure of implicit aptitude must show characteristics of implicit learning, for example, the process and outcome of the learning task in the test must be unconscious. Furthermore, given the importance of selective attention in implicit learning, a test of implicit aptitude should include a measure of selective attention, such as by counting the number of high-pitched tolls while performing an SRT or building two types of stimuli (e.g., shapes and colors) in a statistical learning task.

For explicit aptitude, the most practical test is the LLAMA, which is free, efficient, and automatically scored. Based on the Bokander and Bylund's review of studies using the LLAMA, the overall reliability of the whole test battery and that of LLAMA_B are above .70, the reliability of LLAMA_E and _F are between .6 and .7, and the reliability of LLAMA_D is normally below .6. LLAMA_B, _E, and _F should be used to measure explicit aptitude, either as a combined package, which may increase the predictive validity, or as discrete measures of rote memory, phonetic coding, and language analytic ability, respectively for the purpose of examining the roles of aptitude components and learning outcomes. The psychometric attributes of LLAMA_D are inconsistent across studies, and therefore it should be either excluded or used with caution. For studies investigating young learners (grades 7–12), the PLAB test is an ideal choice, but Parts 1 (GPA), 2 (motivation), and 3 (English vocabulary) should be excluded, as they do not fit the conceptualization of aptitude as cognitive abilities. Parts 4, 5, and 6 measure inductive learning (analytic ability), sound discrimination, and sound-symbol association, and, except for Part 6, which may favor test-takers familiar with English, the other two subtests are language neutral and can therefore be used for learners with any linguistic background.

Next, we would like to propose a nuanced approach to aptitude, which can be interpreted in two ways. One is a dynamic perspective, that is, the role of aptitude in SLA is not fixed, and therefore it is important to examine variables that mediate the effects of aptitude, such as learning stage, the nature of the linguistic target, age, instruction type, and so on. Researchers may examine these variables as independent variables, but, even in studies where the variables are not of primary interest, it is worth conducting post hoc analyses to probe whether they are significant moderators of aptitude effects. For example, in a study investigating the role of aptitude in grammar learning that is operationalized as overall scores on a grammar test, one may explore whether aptitude has differential associations with individual structures if multiple structures are included in the grammar test. Aptitude may not show significant correlations with total grammar scores, but it may be correlated with certain structures (DeKeyser, 2000; Granena, 2013).

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

Another perspective of a nuanced approach is to view the predictor and outcome variable as componential. From this perspective, aptitude is componential, and so is learning. Thus, it is necessary to examine the associations between specific cognitive abilities and specific aspects of L2 learning, Traditional or explicit aptitude has been conceptualized as a global or unitary construct and examined as "a lump sum" (Carroll & Sapon, 2002), although it is posited to consist of three components. A componential view applies equally to implicit aptitude. It is possible that there exist a set of implicit cognitive abilities that are counterparts of the three components of explicit aptitude: phonetic coding, analytic ability, and memory; the implicit counterparts may correspond to the unconscious learning of pronunciation, grammar, and vocabulary, respectively. To date, there is no systematic theorization of the content or components of implicit aptitude that has been primarily operationalized as sequence learning, tested via SRT, which is posited to be essential for L2 grammar. A good example study of a componential approach to aptitude is Saito et al. (2019), which focused on the relation between pronunciation aptitude and pronunciation learning, distinguished explicit and implicit pronunciation learning abilities, and established links between different pronunciation learning abilities, on the one hand, and segmental and suprasegmental pronunciation learning, on the other.

Finally, we would like to make some recommendations for statistical analysis and reporting. Aptitude's associations with learning outcomes are typically investigated by performing correlation-type analyses, such as simple correlation analysis, factor analysis, multiple regression, path analysis, and structural equation modeling. When multiple constructs (e.g., implicit and explicit aptitude) are investigated and multiple measures for each construct are utilized, it is advisable to conduct a factor analysis to reduce the number of variables to be entered into the subsequent regression model and to map the relationship between the cognitive variables for the purpose of construct validation. This step is especially important for implicit aptitude, a new construct whose conceptualization and measures are under development. Latent variables can then be entered into the regression model using factor scores (Granena, 2019) or composite scores (Morgan-Short et al., 2014). However, it would be inappropriate to use composite scores without evidence for convergent validity, that is, the combined score represents the same latent variable. In multiple regression analysis with multiple predictors, it is advisable to perform a hierarchical analysis and enter variables at different steps to determine the unique variance explained by each variable (Li et al., 2019; Roehr-Brackin & Tellier, 2019). Regarding the reporting of results, it is important to report descriptive statistics (sample size, means, and standard deviations), statistical assumptions, and reliability. For inferential statistics, it is necessary to report results of both null hypothesis statistical testing (p values) and effect sizes such as *d, r,* or *odds ratio*. p values are sensitive to sample size, and a nonsignificant p value may be associated with a large effect size, while a significant p value may be associated with a small effect size. For example, in Abrahamsson and Hystensten's study (2008), the correlation between late bilinguals' aptitude scores and grammaticality judgment scores was .53, which is a strong effect based on Cohen's criteria, but the p value was .09. However, the finding was based on only eleven participants, and increasing the sample size could have made the result statistically significant. The power based on the sample and the effect size is only .55 (the threshold is .80 [Field, 2013]), which means that the probability of finding significant correlations is only 55%. Therefore, interpretations of inferential statistics should be based on at least three indices: p, effect size, and power. A post hoc

https://doi.org/10.1017/S0267190520000136 Published online by Cambridge University Press

power analysis can be conducted to ascertain whether the sample size is large enough to achieve significance in NHST given the obtained effect size.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0267190520000136

References

*Included in the synthesis; see Supplementary Information for a full list of synthesized studies

*Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition*, *30*(4), 481–509.

AREA, APA, & NCME. (2014). Standards for educational and psychological testing. Washington, DC: AREA.

*Artieda, G., & Muñoz, C. (2016). The LLAMA tests and the underlying structure of language aptitude at two levels of foreign language proficiency. *Learning and Individual Differences*, *50*, 42–8.

*Benson, S., & DeKeyser, R. (2019). Effects of written corrective feedback and language aptitude on verb tense accuracy. *Language Teaching Research*, *23*, 702–26.

*Bokander, L., & Bylund, E. (2020). Probing the internal validity of the LLAMA language aptitude tests. *Language Learning*, *70*, 11–47.

*Bowles, A., Chang, C., & Karuzis, V. (2016). Pitch ability as an aptitude for tone learning. *Language Learning*, *66*, 774–808.

*Brooks, P., Kwoka, N., & Kempe, V. (2016). Distributional effects and individual differences in L2 morphology learning. *Language Learning*, *67*, 171–207.

*Buffington, J., Demos, A., & Morgan-Short, K. (in press). The reliability and validity of procedural memory assessments in second language acquisition research. *Studies in Second Language Acquisition*.

Carroll, J. (1981). Twenty-five years of research on foreign language aptitude. In K. Diller, (Ed.), *Individual differences and universals in language learning aptitude* (pp. 83–118). Newbury House.

Carroll, J., & Sapon, S. (2002). *Manual for the MLAT*. Second Language Testing, Inc.

*Cox, J., Lynch, J., Mendes, N., & Zhai, C. (2019). On bilingual aptitude for learning new languages: The roles of linguistic and nonlinguistic individual differences. *Language Learning*, *69*, 478–514.

Cronbach, J., & Furby, L. (1970) How should we measure 'change' or should we? *Psychological Bulletin*, *74*, 68–80.

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

*Dahlen, K., Caldwell-Harris, C. (2013). Rehearsal and aptitude in foreign vocabulary learning. *Modern Language Journal*, *97*, 902–16.

*DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, *22*(4), 499–533.

*Doughty, C. (2019). Cognitive language aptitude. *Language Learning*, *69*, 101–26.

Ellis, N. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition*, *27*, 305–52.

Ellis, N., & Wulff, S. (2015). Usage-based approaches to SLA. In B. VanPatten & J. Williams (Eds.), *Theories of second language acquisition* (pp. 75–93). Routledge.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*, 141–72.

*Erlam, R. (2005). Language aptitude and its relationship to instructional effectiveness in second language acquisition. *Language Teaching Research*, *9*(2), 147–71.

*Faretta-Stutenberg, M., & Morgan-Short, K. (2018). The interplay of individual differences and context of learning in behavioural and neurocognitive second language development. *Second Language Research*, *34*, 67–101.

Field, A. (2013). *Discovering statistics using SPSS*. Sage.

Gass, S. (2018). *Input, interaction, and the second language learner*. Routledge.

*Godfroid, A., & Kim, K. (in press). On the contributions of implicit-statistical learning aptitude to implicit grammatical knowledge. *Studies in Second Language Acquisition*.

*Granena, G. (2013). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning*, *63*(4), 665–704.

——— (2015). Cognitive aptitudes for implicit and explicit learning and information-processing styles: An individual differences study. *Applied Psycholinguistics*, *37*, 577–600.

——— (2019). Cognitive aptitudes and L2 speaking proficiency. *Studies in Second Language Acquisition*, *41*, 313–36.

*Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research*, *29*(3), 311–43.

*Hamrick, P. (2015). Declarative and procedural memory abilities as individual differences in incidental language learning. *Learning and Individual Differences*, *44*, 9–15.

*Hodel, A., Markant, J., Van Den Heuvel, S., Cirilli-Raether, J., & Thomas, K. (2014). Developmental differences in effects of task pacing on implicit sequence learning. *Frontiers in Psychology*, *5*, 1–10.

*Hu, X., Ackermann, H., Kartin, J., Erb, M., Winkler, S., & Reiterer, S. (2013). Language aptitude for pronunciation in advanced second language (L2) learners: Behavioural predictors and neural substrates. *Brain & Language*, *127*, 366–76.

*Hummel, K. (2009). Aptitude, phonological memory, and second language proficiency in nonnovice adult learners. *Applied Psycholinguistics*, *30*, 225–49.

*Hwu, F., Wei, P., & Sun, S. (2014). Aptitude-treatment interaction effects on explicit rule learning: A latent growth curve analysis. *Language Teaching Research*, *18*, 294–319.

*Kasprowicz, R., Marsden, E., & Sephton, N. (2019). Investigating distribution of practice effects for the learning of foreign language verb morphology in the young learner classroom. *Modern Language Journal*, *103*, 580–606.

*Kaufman, S., DeYoung, C., Gray, J., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition*, *116*, 321–40.

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

*Kormos, J. & Trebits, A. (2012). The role of task complexity, modality, and aptitude in narrative task performance. *Language Learning*, *62*, 439–72.

*Kourtali, N., & Révész, A. (2020). The roles of recasts, task complexity, and aptitude in child second language development. *Language Learning*, *70*, 179–218.

*Lado, B. (2017). Aptitude and pedagogical conditions in the early development of a nonprimary language. *Applied Psycholinguistics*, *38*, 679–701.

*Li, S. (2013). The interactions between the effects of implicit and explicit feedback and individual differences in language analytic ability and working memory. *Modern Language Journal*, *97*, 634–54.

——— (2015). The associations between language aptitude and second language grammar acquisition: A meta-analytic review of five decades of research. *Applied Linguistics*, *36*, 385–408.

——— (2016). The construct validity of language aptitude. *Studies in Second Language Acquisition*, *38*, 801–42.

——— (2017). The effects of cognitive aptitudes on the process and product of L2 interaction: A synthetic review. In L. Gurzynski-Weiss (Ed.), *Expanding individual difference research in the interaction approach: Investigating learners, instructors and researchers* (pp. 41–70). John Benjamins.

——— (2018). Language aptitude. In A. Burns & J. Richards (Eds.), *Cambridge guide to learning English as a second language* (pp. 63–72). Cambridge: Cambridge University Press.

*Li, M., & DeKeyser, R. (2017). Perception practice, production practice, and musical ability in L2 Mandarin tone-word learning. *Studies in Second Language Acquisition*, *39*, 593–620.

Li, S., & DeKeyser, R. (in press). Implicit language aptitude: Conceptualizing the construct, validating the measures, and examining the evidence. *Studies in Second Language Acquisition*.

*Li, S., & Qian, J. (in press). Exploring syntactic priming as a measure of implicit aptitude. *Studies in Second Language Acquisition*.

*Li, S., Ellis, R., and Zhu, Y. (2019). The associations between cognitive ability and L2 development under five different instructional conditions. *Applied Psycholinguistics*, *40*, 693–722.

Li, S., & Wang, H. (2018). Traditional literature review and research synthesis. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.), *Palgrave handbook of applied linguistics research methodology* (pp. 123–44). London: Palgrave.

*Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., Smith, B. K., Bunting, M. F., & Doughty, C. J. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning*, *63*(3), 530–66.

Long, M. (2015). *Second language acquisition and task-based language teaching*. Blackwell.
Mackey, A. (2020). *Interaction, feedback, and task research in second language learning: Methods and design*. Cambridge: Cambridge University Press.
Meara, P. (2005). *LLAMA Language Aptitude Tests*. Lognostics.
*Morgan-Short, K., Faretta-Stutenberg, M., Brill-Schuetz, K., Carpenter, H., & Wong, P. (2014). Declarative and procedural memory as individual differences in second language acquisition. *Bilingualism: Language and Cognition*, *17*, 56–72.
*Potter, C., Wang, T., & Saffran, J. (2017). Second language experience facilitates statistical learning of novel linguistic materials. *Cognitive Science*, *41*, 913–27.
Robinson, P. (Ed.) (2011). *Second language task complexity*. Amsterdam: John Benjamins.
*Roehr-Brackin, K., & Tellier, A. (2019). The role of language-analytic ability in children's instructed second language learning. *Studies in Second Language Acquisition*, *41*, 1111–31.
*Sáfár, A., & Kormos, J. (2008). Revisiting problems with foreign language aptitude. *International Review of Applied Linguistics in Language Teaching*, *46*(2), 113–36.
*Saito, K., Sun, H., & Tierney, A. (2019). Explicit and implicit aptitude effects on second language speech learning: Scrutinizing segmental and suprasegmental sensitivity and performance via behavioural and neurophysiological measures. *Bilingualism: Language and Cognition*, *22*(5) 1123–40.
*Sheen, Y. (2008). The effect of focused written corrective feedback and language aptitude on ESL learners' acquisition of articles. *TESOL Quarterly*, *41*, 255–83.
*Shintani, N., & Ellis, R. (2015). Does language analytical ability mediate the effect of written feedback on grammatical accuracy in second language writing? *System*, *49*, 110–19.
Skehan, P. (2012). Language aptitude. In Gass, S. & A. Mackey (Eds.), *Handbook of second language acquisition* (pp. 381–95). Routledge.
*Smemoe, W., & Haslam, N. (2013). The effect of language learning aptitude, strategy use, and learning context on L2 pronunciation learning. *Applied Linguistics*, *34*, 435–56.
Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis". *Language Learning*, *60*, 263–308.
*Sparks, R. L., Patton, J., Ganschow, L., & Humbach, N. (2012). Relationships among L1 print exposure and early L1 literacy skills, L2 aptitude, and L2 proficiency. *Reading and Writing*, *25*, 1599–1634.
*Suzuki, Y. (2018). The role of procedural learning ability in automatization of L2 morphology under different learning schedules: *An exploratory study. Studies in Second Language Acquisition*, *40*, 923–37.
——— (in press). Probing the construct validity of LLAMA_D as a measure of implicit learning aptitude: Incidental instructions, confidence ratings, and reaction time measure. *Studies in Second Language Acquisition*.
*Suzuki, Y., & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language Learning*, *65*, 860–95.
——— (2017). The interface of explicit and implicit knowledge in a second language: Insights from individual differences in cognitive aptitudes. *Language Learning*, *65*, 860–95.
Ullman, M., & Lovelett, J. (2018). Implications of the declarative/procedural model for improving second language learning: The role of memory enhancement techniques. *Second Language Research*, *34*, 39–65.
*VanPatten, B., & Borst, S. (2012). The roles of explicit information and grammatical sensitivity in processing instruction: Nominative-accusative case marking and word order in German L2. *Foreign Language Annals*, *45*(1), 92–109.
*Walker, N., Monaghan, P., Schoetensack, C., & Rebuschat, P. (2020). Distinctions in the acquisition of vocabulary and grammar: An individual differences approach. *Language Learning*, *70*, 221–54
*Winke, P. (2005). *Individual differences in adult Chinese second language acquisition: The relationships between aptitude, memory, and strategies for learning* [Unpublished doctoral dissertation]. Georgetown University.
——— (2013). An investigation into second language aptitude for advanced Chinese language learning. *Modern Language Journal*, *97*, 109–30.
*Yalçın, Ş., Çeçen, S., & Erçetin, G. (2016). The relationship between aptitude and working memory: an instructed SLA context. *Language Awareness*, *25*, 144–58.
*Yalçın, Ş., & Spada, N. (2016). Language aptitude and grammatical difficulty: An EFL classroom-based study. *Studies in Second Language Acquisition*, *38*, 239–63.

*Yi, W. (2018). Statistical sensitivity, cognitive aptitudes, and processing of collocations. *Studies in Second Language Acquisition*, *40*(4), 831–56.
*Yilmaz, Y. (2013). Relative effects of explicit and implicit feedback: The role of working memory and language analytic ability. *Applied Linguistics*, *34*, 344–68.
*Yilmaz, Y., & Granena, G. (2019). Cognitive individual differences as predictors of improvement and awareness under implicit and explicit feedback conditions. *Modern Language Journal*, *103*, 686–702.

Cite this article: Li, S., & Zhao, H. (2021). The methodology of the research on language aptitude: A systematic review. *Annual Review of Applied Linguistics*, *41*, 25–54. https://doi.org/10.1017/S0267190520000136