CrossMark

Studies in Second Language Acquisition, 2016, **38**, 801–842.
doi:10.1017/S027226311500042X

THE CONSTRUCT VALIDITY OF LANGUAGE APTITUDE

A Meta-Analysis

Shaofeng Li
University of Auckland

A meta-analysis was conducted to examine the construct validity of language aptitude by synthesizing the existing research that has been accumulated over the past five decades. The study aimed to provide a thorough understanding of the construct by aggregating the data reported in the primary research on its correlations with other individual difference variables and with second language (L2) achievement. A total of 66 studies were retrieved that contributed effect sizes based on 109 unique samples and 13,035 foreign language learners. The results revealed that (1) aptitude was independent of other cognitive and affective factors: it was distinct from motivation, had a negative correlation with anxiety, and overlapped with, but was distinguishable from, intelligence; (2) executive working memory was more strongly associated with aptitude and aptitude components than phonological short-term memory; (3) aptitude measured using full-length tests was a strong predictor of general L2 proficiency, but it had low predictive validity for vocabulary learning and L2 writing; and (4) different aptitude components demonstrated differential predictive validity for different aspects of learning. The findings are useful for tackling a number of conundrums surrounding language aptitude and shed light on how to reconceptualize the construct and reorient the research.

The project this article is based on was presented as part of the colloquium "Language Aptitude: Reconceptualising the Construct and Reorienting the Research" at the 2014 AILA World Congress held in Brisbane, Australia. I would like to extend my sincere gratitude to the five anonymous reviewers for their thorough, insightful comments that significantly enhanced the quality of this article. I would also like to thank Rod Ellis for his support and insights and John Read for his specialized knowledge about test validity. I am solely responsible for any errors.

© Cambridge University Press 2015

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

INTRODUCTION

Language aptitude has been one of the most fertile grounds for research on the impact of learners' cognitive abilities on their second language development or attainment. The proliferation of aptitude studies is attributable to the explanatory power of cognitive aptitudes for second language acquisition (SLA), the availability of measures with strong predictive validity (although they have been criticized for lack of construct validity), and the relevance of the research findings for second language (L2) pedagogy. Carroll (1990) stated that aptitude as measured by the MLAT (Modern Language Aptitude Test) is distinct from other individual difference variables, not subject to training effects, and predictive of L2 achievement. The construct of language aptitude, therefore, entails its content, characteristics, and associations with covariates and with the criterion variable—L2 achievement. This study aims to examine the validity of the construct of language aptitude by conducting a meta-analysis of the empirical data reported in the primary research that is couched in Carroll's framework.

Such a study is warranted on two accounts. First, over the past five decades there has been a considerable volume of research on language aptitude, but to date there has been no attempt to systematically synthesize the quantitative results that have been accumulated and verify the theoretical claims about the construct. Carroll, Sapon, Reed, and Stansfield (2010) commented that the results of the large amount of aptitude research "*seem* [emphasis added] to offer strong support for the major claims about the nature of language aptitude" (p. 19). However, the theoretical claims need to be tested through aggregation of the quantitative results rather than impressionistic judgments. Also, there is much heterogeneity in the findings of the primary studies that has not been accounted for in previous reviews. For example, a preview of the data of this meta-analysis shows that the correlations between motivation and aptitude ranged from r = -.20 to r = .32, and the MLAT Manual (Carroll & Sapon 2002) reports that the correlations between foreign language learners' MLAT scores and their course grades ranged from r = .13 to r = .83. A meta-analysis can not only provide precise numeric estimates about the construct but also identify the variables that account for the variation of the findings across the sampled populations. Second, the traditional conceptualization of aptitude and the tests used to measure it have received much criticism, and there has therefore been a call to revamp the construct and update the tests (Robinson, 2001; Skehan, 2002). However, the weaknesses of existing theoretical claims and aptitude measures cannot be assumed, and changes and revisions must be based on what has been demonstrated about the construct in empirical research. It is hoped that the meta-analytic findings of this study contribute to the

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

accurate understanding of traditional aptitude and afford empirical justification for future developments. It must be further pointed out that the objective of this meta-analysis is to provide an inductive review of what has been found about traditional aptitude by aggregating the quantitative data contributed by the primary studies, rather than prove or disprove any existing theories or claims about the construct.

THE CONSTRUCT OF LANGUAGE APTITUDE

Cronbach and Meehl defined a construct as "some postulated attribute of people, assumed to be reflected in test performance" (1955, p. 283). A construct is a concept and is operationalized as a specific measure that may consist of items gauging the multiple dimensions of what is termed a *latent* variable, which is not directly observable. The validity of a construct involves the "theoretical knowledge of the trait and ability being measured, knowledge of other related variables, hypothesis testing, and statements regarding the relationship of the test variable to a network of other variables that have been investigated" (Groth-Marnat, 2003, p. 22). In the following, an overview of the construct of aptitude is provided including its nature and components and how it has been measured and validated.

The Nature, Content, and Measurement of the Construct

Perhaps the best way to start this section is to point out that aptitude has been defined and conceptualized in different ways in different disciplines. For example, Snow (1991), a pioneer in aptitude research in the field of educational psychology who proposed the ATI (aptitude-treatment interaction) approach, argued that aptitude refers to any person characteristic that is important to achieving a learning goal including affective (feelings and emotions) and conative (goal setting and determination) processes such as anxiety and motivation as well as cognitive abilities such as analytic ability and memory. In SLA, aptitude is generally considered as a cognitive variable, but the way aptitude is conceptualized and operationalized in the research is far from homogeneous, which is evidenced by the existence of a variety of aptitude tests developed for various purposes (see the following text for a discussion of different aptitude measures).

Characteristics. Carroll (1981) claimed that language aptitude is (1) distinct from other individual difference variables such as motivation, anxiety, and intelligence; (2) impervious to external influence;

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

(3) predictive of L2 learning rate; (4) drawn upon in formal instruction where learners make "a deliberate effort to learn a foreign language" (p. 83); and (5) the "initial state of readiness" (p. 86) for learning a foreign language. These claims have the following implications. First, aptitude, as measured by Carroll's seminal test—the MLAT—and other tests based on Carroll's views, is a cognitive variable distinguished from the variables in the affective and conative domains. Second, language aptitude is domain specific in the sense that it relates only to language learning and is different from abilities for other academic subjects such as intelligence. Third, aptitude is not affected by previous language learning experience, not subject to training, and remains stable for a relatively long period of time during one's life span. Fourth, it concerns the ease and speed with which one learns a foreign language in comparison with peers during a certain period, and therefore it is not intended to predict ultimate L2 success. Fifth, it is hypothesized to be implicated in instructed SLA, and it was uncertain to Carroll whether it is important in nontutored or naturalistic settings, where learning often happens unconsciously through exposure to a large amount of linguistic input. Sixth, it is a set of abilities that are of central importance in the preliminary stages of L2 development, and the extent to which it affects advanced levels of learning is not known. To be clear, these are Carroll's views on the characteristics of aptitude, some of which have not been followed in the research. For example, although the aptitude that Carroll conceptualized is intended to be predictive of learning rate in instructed contexts, there has been a body of research examining its role in affecting ultimate learning success in naturalistic settings (e.g., Abrahamsson & Hyltenstam, 2008; DeKeyser, 2000; Granena & Long, 2013).

Components. Carroll (1981) posited four fundamental abilities for language learning: phonetic coding, grammatical sensitivity, inductive learning, and rote learning, which jointly underlie a latent, second-order variable labeled as language aptitude. Grammatical sensitivity and inductive learning relate to the analytic aspects of aptitude and can therefore be unified as *language analytic ability* (Skehan, 2002), which led to a more parsimonious, tripartite structure. Although Carroll stated that aptitude was componential, he saw it as a unitary concept represented by a composite score that is predictive of course grades in his validation research and in his comments on the validity of the construct (Carroll et al., 2010). Related to this issue is a need for transparency over what is meant by the term *aptitude*, and in fact the confusion may have been a source of misinterpretation of research findings and theoretical claims (Granena & Long, 2013). For example, Krashen (1981), in his criticism of the concept of language aptitude, referred to aptitude as grammatical sensitivity, considered phonetic coding as irrelevant, and ignored rote learning. Both Abrahamsson and Hyltenstam (2008) and DeKeyser (2000) examined the relationship

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

between aptitude and age, but in the former a composite aptitude score representing all three components was used while in the latter aptitude was operationalized as grammatical sensitivity. This discrepancy may have been partly responsible for the different results between the two studies regarding whether aptitude is correlated with child SLA.

Despite its strong predictive power for L2 learning rate, traditional aptitude has been challenged on the grounds that it consists of a set of abilities that are only important for preliminary L2 learning but not advanced SLA, for learning language as discrete items rather than learning that happens in context, for explicit learning instead of implicit learning, and for learning the formal aspects as opposed to the pragmatic aspects of a language (Li, 2015; Robinson, 2001; Skehan, 2012). However, it must be recognized that some of the criticisms are only justifiable on theoretical grounds and have not been empirically verified. For example, although traditional aptitude components have been assumed to be only relevant for preliminary L2 learning, primary studies seeking to identify abilities relevant for advanced learning are few and far between. Recently a research unit led by Catherine Doughty at the University of Maryland (see Linck et al., 2013) developed an aptitude battery called the Hi-LAB which is composed of 12 measures of seven cognitive abilities. The tests were administered to 476 government employees and military personnel in the United States, with a view to discovering abilities predictive of high attainment operationalized as either learners' outstanding performance on a standardized test (Defence Language Proficiency Test) or high-level job performance in one or more foreign languages. Logistic regression analyses showed that measures of phonological short-term memory (PSTM), implicit learning, and rote memory were significant predictors of high attainment.

One feature that stands out about the Hi-LAB is the inclusion of six measures of the executive and storage functions of working memory, which demonstrates the researchers' emphasis of its importance for advanced learning. Working memory is a cognitive device that performs the dual function of information storage and processing. It has been argued that working memory is where all three components of aptitude (phonetic coding, language analytic ability, and memory) converge (Miyake & Friedman, 1998) and that the traditional rote memory component does not fit the kind of learning that happens in instructional contexts where more emphasis is placed on exposure to authentic language use and incidental learning. Although some state-of-the-art reviews (DeKeyser & Koeth, 2011; Skehan, 2012) have included working memory as a component of aptitude, there needs to be more theoretical elaboration and empirical exploration of how working memory relates to traditional aptitude and aptitude components. One objective of this meta-analysis is to draw on existing data to map the relationship between traditional aptitude and working memory.

Another breakthrough of Linck et al.'s validation study is the inclusion of a measure of implicit aptitude, an initiative that constitutes a departure

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

from, and a significant complement to, the paradigm of traditional aptitude, which comprises of a set of abilities that are likely more correlated with conscious L2 learning. This initiative heralds a vibrant stream of aptitude research in search of implicit learning abilities and validating the measures of these abilities, echoing the currently live debate on the mechanism and measurement of implicit learning in SLA research (e.g., Ellis, 2005; Suzuki & DeKeyser, 2015; Williams, 2009). A good example study is Granena (2015), who found a separation between learners' performance on the LLAMA test—a measure of explicit aptitude—and their performance on a serial reaction time test—a measure of implicit aptitude. The two types of aptitude were found to be significantly, but distinctly, correlated with two different cognitive styles: rational-analytical and experiential-intuitive, with the former representing a propensity for explicit learning and the later implicit learning.

Grigorenko, Sternberg, and Ehrman (2000) sought to identify abilities important for contextualized learning and proposed a model based on a cognitive theory of knowledge acquisition, which posits a central role for abilities to cope with novelty and ambiguity. The test battery (CANAL-F) gauging the abilities includes tasks involving the learning of the vocabulary and morphosyntax of an artificial language called Ursulu. Vocabulary is learned through tasks requiring learners to infer the meanings of Ursulu words embedded in English texts as well as paired-associate tasks that tap rote memory. Morphosyntax is learned by means of an inductive task (sentence inference) where learners are provided with Ursulu sentences and their English translations, encouraged to make inferences about the morphosyntactic aspects of the language, and then tested through multiple-choice items. The aptitude battery also includes a deductive task where learners are presented with rules about how Ursulu works and later apply the knowledge when answering multiple-choice questions. In addition to tasks relating to the lexical and syntactic aspects of language learning, the test also consists of comprehension tasks that engage learners in semantic processing.

Measurement. The most frequently used and comprehensive test batteries of traditional aptitude are the MLAT and the PLAB (Pimsleur, 1966), and the MLAT "seems to be the best overall instrument for predicting language learning success" (Parry & Child, 1990, p. 52). The popularity of the MLAT and PLAB is largely attributable to their strong predictive validity that is established based on the data contributed by large samples (5,000 learners for the MLAT and 6,000 for the PLAB) from a large number of foreign language classes at different levels of L2 development. Although Carroll (1981) claimed that the MLAT and the PLAB measure somewhat similar abilities and are comparable in validity, there are appreciable differences between them in that the PLAB, which is intended for high school students, includes sections for GPA and

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

motivation, and leaves out rote learning. Other subsequently developed tests include the DLAB, which targets high-aptitude learners (Petersen & Al-Haik, 1976); the VORD, which is intended for the learning of challenging languages (Child, 1998); and the LLAMA (Merea, 2005), a free computerized test modeled on the MLAT. Notwithstanding slight differences between these tests, they are based on Carroll's views about the nature and characteristics of aptitude. These are developed for L1 English speakers (except for the LLAMA, which is language neutral), used for predictive purposes, and validated following the tradition of psychometric testing.

As previously discussed, more recent developments in aptitude testing seeking to address the limitations of the MLAT and similar tests include the CANAL-F (Grigorenko et al., 2000) and the Hi-LAB (Linck et al., 2013). The CANAL-F is theory based and consists of a number of inductive and deductive tasks that replicate classroom tasks and mirror the process of instructed language learning that involves the processing of linguistic materials from the word level to the sentence level and from form to meaning. However, the learning tasks are heavily form oriented and seem to tap primarily into explicit aptitudes. The test has not been widely used in aptitude research, and more information is in order about its validity and reliability (but see Thompson, 2013).The Hi-LAB is innovative in a number of aspects, but the generalizability of the findings of the validation study is subject to caveats relating to the way proficiency was operationalized (job performance or scores of listening and reading on a proficiency test) and the fact that only rote memory was included in the battery, leaving one wonder whether including other components of traditional aptitude such as language analytic ability and phonetic coding would have changed the results. Furthermore, despite the originality of the concept of implicit aptitude, which is a central component of the Hi-LAB, there has been only weak empirical evidence for the validity of the construct. For example, Granena (2015) only observed a weak, marginally significant correlation between a measure of implicit aptitude and a measure of implicit cognitive learning style. Also, even if one predictor variable is found to overlap or covary with another predictor variable with a similar mechanism, which constitutes evidence for convergent validity, evidence needs to be collated for its predictive validity, that is, whether it is correlated with the criterion variable, which is implicit learning in this case.

Validation of the Construct

A Tripartite versus Unitary View of Validity. Traditionally, validity in educational assessment is a tripartite concept comprising content validity,

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

criterion-oriented validity, and construct validity (Cronbach & Meehl, 1955). Content validity, usually evaluated through expert opinions, refers to the extent to which test items represent the skills and knowledge of the substantive domain or whether the tasks included in a test mirror what happens in the real world. Criterion-oriented validity concerns how the test scores correlate with a criterion measure. There are two types of criterion-oriented validity: one is *concurrent validity*, in which case the test for the predictor variable and that for the criterion variable are administered at the same time and the purpose is to see whether one can replace the other; the other is called *predictive validity*, which refers to whether one's test performance forecasts one's performance on another test that is administered later. In aptitude research, predictive validity has been of primary interest, as demonstrated by Carroll and Sapon's (1959) validation research in which aptitude scores were obtained at the beginning of the training (e.g., a semester) and correlated with scores for the criterion variable (e.g., course grades) obtained at the end of the training. However, in many subsequent aptitude studies, aptitude and criterion tests are administered at the same time, in which case the term *predictive* needs to be understood in a logical sense, that is, aptitude has an influence on L2 learning outcome, but not the other way round. Finally, in traditional testing theories, construct validity indicates whether the theoretical claims relating to the concept are warranted and whether the hypothesized/postulated relationships are corroborated by empirical evidence.

Notwithstanding the theoretically plausible distinction between the three types of validity, they are interrelated and difficult to disentangle (Popham, 2000). Consequently, different from the traditional tripartite view, current testing theories hold that validity is a unitary concept in which construct validity is an overarching term for all three types of validity; under this view, content and criterion-related validity are considered aspects of construct validity. Validation of a construct involves the collation of different types of empirical evidence to evaluate the appropriateness of the inferences associated with the test performance (Messick, 1989). In the case of language aptitude, validation of the construct entails drawing on existing research findings to examine the concept in terms of its relationships with other individual difference variables (such as motivation, anxiety, and intelligence) and with the criterion variable—L2 attainment.

Aptitude and L2 Learning. Most aptitude research has investigated the relationship between language aptitude and L2 achievement. The research falls into two broad categories informed by different theoretical approaches, the fundamental difference lying in whether aptitude is perceived as a trait that affects L2 *outcome*—a predictive approach, or L2 *process*—an interactionist approach. Carroll (1981) argued that the

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

primary function of aptitude is predictive, and the purpose is to investigate the association between aptitude and L2 attainment. In a typical predictive study, the researcher obtains two sets of scores: one for aptitude or an aptitude component and one for L2 attainment in the form of course grades or scores on a proficiency test, and learners take the tests in one sitting. Analyses of a correlational nature (including multiple regression analysis) are then conducted to determine whether aptitude is a significant predictor. Some predictive studies also investigated the predictive power of aptitude in comparison to other variables, such as motivation (Bialystok & Fröhlich, 1978; Gordon, 1980), anxiety (Bell & McCallum, 2012; Sparks, Patton, Ganschow, & Humbach, 2009), and intelligence (Bond, 2011; Sparks, Patton, Ganschow, Humbach, & Javorsky, 2006).

Within the interactionist paradigm, researchers (Robinson, 2002, 2005) contended that instead of viewing aptitude as a determinant of L2 success, it is better to treat it as a dynamic construct, that is, its role varies as a function of the processing demands of different learning conditions. In a typical interactionist study (e.g., de Graaff, 1997; Sheen, 2007), the learners take an aptitude test, receive some type of instructional treatment (implicit vs. explicit, inductive vs. deductive, etc.) that focuses on the learning of a particular linguistic structure, and then take a posttest (and often also a pretest prior to the treatment) as a measure of their learning gains after the treatment. The data are then analyzed to explore whether aptitude or aptitude components are drawn on differently in different treatment conditions. A meta-analysis of the interactionist studies showed that aptitude was more likely to be involved in explicit treatments than implicit treatments (Li, 2015). The studies on how aptitude interfaces with inductive versus deductive instruction seem to show that aptitude was more likely to be implicated in inductive approaches that favored high-aptitude learners, and that it was either unimportant or less important in deductive approaches (Erlam, 2005; Hauptman, 1971; Hwu & Sun, 2012). However, these findings must be treated circumspectively because of the different ways the inductive-deductive distinction is operationalized and the different aptitude tests used in the studies, and there is an obvious need for more studies in this line of research before firm conclusions can be reached.

The Nature and Characteristics of Aptitude. While there has been copious research on how aptitude is correlated with L2 learning outcomes, there have been fewer studies investigating its nature and characteristics such as whether it is malleable or affected by learning experience and how it relates to other affective and cognitive variables such as motivation, anxiety, and intelligence. The question of whether experience leads to improvement in aptitude can be examined in three ways. One is to use a between-group design where learners with less

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

language learning experience are compared with those that have more experience in terms of their aptitude scores. The second option is to employ a within-group design where the same learners are given an aptitude test before and after exposure to a foreign language to see if their aptitude scores change over time. The third and best way is to investigate both between- and within-group differences—a mixed design that overcomes the weaknesses of the first two approaches by showing that (1) the higher scores of those with more experience are not due to their higher aptitude (as in a between-group study), and (2) the chronological improvement in the same learners' aptitude scores is only attributable to study experience instead of maturation effects (as in a within-group study). Previous studies using either a between-group (Eisenstein, 1980; Harley & Hart, 1997) or within-group (Ganschow & Sparks, 1995) design have produced mixed results. The only mixed design study is by Sáfár and Kormos (2008) who found that (1) bilingual students that received intensive language instruction scored higher on an aptitude test than those at a regular school who received less language instruction and (2) both groups' aptitude scores were also significantly higher after one academic year than their initial scores. The findings of Sáfár and Kormos appear to suggest that aptitude is malleable, but there is a lack of experimental research investigating the impact of carefully designed training/treatment on subjects' aptitude scores. Also, it remains to be seen whether improved aptitude is transferrable into learning efficiency, as Carroll and Sapon (2002) suspected.

The relationship between aptitude and other individual difference variables entails two dimensions: one is the predictive power of aptitude for L2 achievements in comparison with other variables and the other is the interrelationship between aptitude and other predictor variables. With regard to the former, by and large aptitude has been found to be the most consistent predictor of the criterion variable when examined together with other cognitive and affective variables (Bell & McCallum, 2012; Cochran, McCallum, & Bell, 2010; Granena, 2012; Sparks et al., 2009). The other dimension, which is the focus of this meta-analysis, concerns the associations between aptitude, on one hand, and motivation, anxiety, and intelligence, on the other. Although how aptitude relates to other individual difference variables is critical to the validity of the construct, it has not been the focus of the primary research. However, when examining the comparative effects of aptitude and other variables such as motivation in affecting L2 learning (e.g., Bialystok & Fohlich, 1978; Hsieh, 2004), some researchers also reported the correlations among the predictors. The additional data from these studies can therefore be utilized to zero in on the link between aptitude and other individual difference factors.

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

This Study

This study is part of a large-scale meta-analytic project that examines various aspects of language aptitude by synthesizing the quantitative data available in primary research. The data relating to the relationship between aptitude and L2 grammar acquisition were reported in another study (Li, 2015).[1] It was found that aptitude showed an overall moderate association with grammar acquisition, r = .31, 95% CI = .25 /.36. A subsequent moderator analysis demonstrated that aptitude was more correlated with explicit treatments than implicit treatments and that age was a negative predictor in predictive studies but a positive predictor in interactionist studies. This meta-analysis will include studies taking a predictive approach, and the purpose is to examine other aspects of the construct by testing the claims made by Carroll (1973, 1981, 1990; Carroll & Sapon, 1959; Carroll et al., 2010), who pioneered aptitude research and developed the most influential aptitude test that has dominated this research domain. The claims to be subjected to empirical investigation, which are represented in the three research questions in the following text, relate to its associations with other cognitive and affective variables and with the criterion variable including general L2 proficiency, L2 knowledge (grammar and vocabulary), and L2 skills (listening, speaking, reading, and writing).

The significance of this study is as follows. First, this meta-analysis constitutes the first attempt to comprehensively synthesize the empirical research on language aptitude to examine the attributes of this construct and how it relates to learning as well as other individual difference variables. The retrospective synthesis of past research also benefits future initiatives of reexamining the concept and developing new measures for the construct. Second, the meta-analysis goes beyond the focus of the primary research and examines not only how aptitude relates to L2 proficiency in general but also its associations with specific aspects of L2 learning and the factors that may moderate such associations; it not only shows how aptitude predicts L2 attainment but also how it relates to other individual difference variables. Third, it includes both published articles and unpublished PhD dissertations and MA theses reporting data relating to the research questions and constitutes the most comprehensive synthesis of aptitude research.

This meta-analysis was guided by the following research questions:

1. What is the relationship between language aptitude and other individual difference variables such as motivation, anxiety, and intelligence?[2]
2. What is the relationship between aptitude and working memory, a possible component of aptitude?

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

3. What is the relationship between aptitude and L2 achievement?
   3.1 How do aptitude and aptitude components relate to general L2 proficiency?
   3.2 How do aptitude and aptitude components relate to specific aspects of L2 learning such as L2 knowledge (grammar and vocabulary) and L2 skills (listening, speaking, reading, and writing)?

The first question seeks to verify Carroll's claim that aptitude is distinct from other individual difference variables. The second question aims to map the relationship between working memory[3] and traditional aptitude and explore the viability of including working memory as an aptitude component, as suggested by experts in aptitude research (DeKeyser & Koeth, 2011; Robinson, 2001; Skehan, 2012). The third question subsumes two subquestions relating to the predictive power of aptitude and aptitude components for general L2 proficiency and for different aspects of L2 learning.

METHOD

Study Retrieval

The methods used to search for related primary studies include (1) ancestry chasing—tracing the references in primary research, (2) computer search—searching electronic databases and academic journals, (3) browsing—checking bookshelves in the library, and (4) consultation—consulting experts in the academic field. The searched databases, which were selected based on the findings of synthetic reviews of SLA meta-analyses (In'nami & Koizumi, 2010; Li, Shintani, & Ellis, 2012; Oswald & Plonsky, 2010), include *LLBA*, *ERIC*, *PsycINFO*, *PsycARTICLES*, *MLA International Bibliography*, and *ProQuest Dissertations and Theses*; *Google Scholar* was searched as a complement to the domain-related databases. The primary studies were identified by using keywords relating to (1) the concept and components of aptitude such as language aptitude, phonetic coding, language analytic ability, and rote learning (memory), and (2) the names of the test batteries such as MLAT, PLAB, DLAB, VORD, CANAL-F, LLAMA, and so forth and their subtests such as Number Learning (MLAT 1) and Sound-Symbol Association (PLAB 6). In addition to the predictor variable—language aptitude, keywords relating to other individual difference variables (motivation, anxiety, and intelligence), working memory, and the outcome variable (second language learning) were also used during the search for related primary studies. Upon the completion of the study search, the retrieved articles were sieved and selected based on the research questions and defined research domain.

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

Inclusion/Exclusion Criteria

This meta-analysis included any study that reported data that can be synthesized to address any or all of the research questions, including studies reporting (1) correlations between aptitude and other individual difference variables such as motivation, anxiety, and intelligence (to address the first research question); (2) correlations between aptitude working memory (RQ 2); and (3) correlations between aptitude/aptitude components and L2 achievements (RQ 3). In addition to research articles, another important source of data is the MLAT Manual (Carroll & Sapon, 2002), from which the extracted data included correlation coefficients between foreign language learners' aptitude scores and course grades. The data reported in the manual were collected by the test developers to establish the predictive validity and percentile norms of the test batteries but have not been aggregated using meta-analytic procedures. To take full advantage of the data reported in the primary studies and seek alternative explanations for the results (see Discussion section), descriptive statistics (means and standard deviations) of the learners' aptitude scores were also extracted. To minimize publication bias, that is, studies that obtained significant results are more likely to be published, this meta-analysis included both published studies and unpublished PhD dissertations and MA theses. Furthermore, to be included in the meta-analysis, a study had to be published prior to May 2014, the closing time for data collection.

The following exclusion criteria were formulated. First, this meta-analysis is a synthesis of predictive studies which are motivated by Carroll's (1990) claims about the characteristics of aptitude; therefore, interactionist studies (the results for which were reported in a different study) that examined the contingent relationships between aptitude and different types of instructional treatment were excluded from the data in answering the third research question, which concerns the predictive validity of aptitude. This is because predictive and interactionist studies are based on different theoretical assumptions and are conducted in different ways; mixing the two types of studies would have led to the classical *apples and oranges* problem of meta-analysis (Lipsey & Wilson, 2001; Plonsky & Brown, 2015). However, a few studies in the interactionist paradigm (e.g., Carpenter, 2008; Lado, 2008; Robinson, 2002; Sachs, 2010) reported data that could be used to address the research questions (e.g., the correlations between aptitude and working memory) and were therefore included for that reason. Second, a study was excluded if the aptitude test deviates from traditional aptitude measures, such as Ranta (2002) who used an L1 error correction test as a measure of language analytic ability. Third, if two studies by the same author reported data based on the same sample (e.g., Robinson, 2002, 2005; Sparks et al.,

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

2006, 2009), only one was kept in the dataset, to prevent effect size inflation and Type I error. Fourth, due to logistic constraints, no attempt was made to find studies published in other languages or unpublished manuscripts "tucked away in the researchers' file drawers."

Coding

The coding protocol of this meta-analysis involves a description of how information from the primary studies was recorded and categorized pertaining to the measures of (1) aptitude—the predictor variable; (2) the covariates or other individual difference variables—motivation, anxiety, and intelligence; (3) working memory—a possible aptitude component; and (4) L2 achievement or the criterion variable. Table 1 provides an overview of the coding scheme, including the constructs and the instruments used to measure them. For the convenience of analysis and for the purpose of obtaining a clear picture about the results for different aptitude measures and the underlying constructs, the measures of aptitude were coded in two ways: in terms of the names of the tests (MLAT, PLAB, etc.) and their subtests (MLAT 1 Number Learning, PLAB 4 Language Analysis, etc.), and in terms of the underlying construct, that is, aptitude and aptitude components, including phonetic coding, language analytic ability, and rote memory. Information about the matching between constructs and the frequently used tests is presented in Table 1. Specifically, whole test batteries were referred to as *aptitude*; MLAT 2, PLAB 5, PLAB 6, LLAMA-D, and LLAMA-E as *phonetic coding*; MLAT 4, PLAB 4, and LLAMA-F as *language analytic ability*; and MLAT 5 as *rote memory*. Where there is no clear correspondence between a subtest and an aptitude component, such as in the case of MLAT 1 Number Learning and MLAT 3 Spelling Clues, the name of the subtest was retained and no further coding was pursued. The preceding labels for the MLAT and the PLAB also apply to the adapted versions of the two tests in other languages such as the HUNLAT, LABJ, and so forth.

With respect to the coding of the covariates (Table 1), motivation, anxiety, and intelligence were coded as unitary constructs although they may consist of multiple dimensions, because most of the primary studies only reported the correlations between aptitude and those variables as holistic constructs. In the coding of working memory, a distinction was made between PSTM and executive working memory (EWM) (Wen, 2015), the former referring to the storage component of short-term memory and the latter to one's ability to store and process information simultaneously. Measures such as word/nonword repetition and digit span where learners are presented with, and required to recall,

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

Table 1. Coding protocol for the variables

Variables	*Constructs*		*Example Instruments*
Predictor	Aptitude		MLAT[a]/QMLAT, [b] PLAB/QPLAB, [b] LLAMA, VORD
	Phonetic Coding		MLAT/QMLAT 2, PLAB/QPLAB 5, PLAB/QPLAB 6, LLAMA-D, LLAMA-E
	Language Analytic Ability		MLAT/QMLAT 4, PLAB/QPLAB 4, LLAMA-F, Language Analysis Test
	Rote Memory		MLAT/QMLAT 5, LLAMA-B
	Miscellaneous[c]		Number Learning, Spelling Clues, GPA, English Vocabulary, Interest in Foreign Language Learning
Covariate	Motivation		National Test Battery, AMQ
	Anxiety		FLCAS
	Intelligence		WAIS, Test of Cognitive Skills, PMA
	Working memory	PSTM	Word/Nonword Repetition
		EWM	Reading/Listening Span, Backward Digit
Criterion	Proficiency		Course/Exam Grades, TOEFL
	Knowledge	Grammar	Error Correction, Grammaticality Judgment Test, Metalinguistic Knowledge
		Vocabulary	Translation, Vocabulary Recognition, Vocabulary Production
	Skills	Listening	TOEFL Listening, Sentence/Passage Comprehension
		Reading	Reading Comprehension
		Speaking	Oral Narrative, Elicited Imitation,[d] Interview
		Writing	Composition, Letter Writing

[a] Whole test batteries of aptitude.
[b] Quasi-MLAT/PLAB: versions of the MLAT and the PLAB in other languages (e.g., the HUNLAT is the Hungarian version of the MLAT).
[c] GPA, English Vocabulary, and Interest in Foreign Language Learning were not included in the analyses.
[d] Elicited imitation was a validated measure of implicit knowledge that is drawn on in oral production (Ellis, 2005; Erlam, 2006).

sets of unrelated items were coded as PSTM, while measures such as listening and reading span tests where learners performed the dual task of information storage and processing were coded as EWM. The distinction between PSTM and EWM is important given the current call to include working memory as an aptitude component, which necessitates a fine-grained examination of the connections between aptitude and the two distinct types of short-term memory.

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

The coding of L2 attainment, the criterion variable, is based on three sets of measures tapping into (1) general proficiency, (2) knowledge of specific aspects of the L2 such as grammar and vocabulary, and (3) language skills such as listening, speaking, reading, and writing. The criterion variable was coded as general proficiency if it was assessed through course or exam grades (e.g., Bain, McCallum, Bell, Cochran, & Sawyer, 2010; Sparks et al., 2009), standardized proficiency tests such as TOEFL (Ross, Yoshinaga, & Sasaki, 2002), and cloze tests (Harley & Hart, 1997).[4] As with general proficiency, L2 knowledge and skills were gauged through a variety of measures in the included studies, and the overarching principle followed in coding these measures was to retain the original labels for the constructs reported in the included studies. For example, a test was labeled vocabulary if it was claimed to measure learners' vocabulary knowledge. A hybrid test that measured more than one construct, such as a test of both grammar and vocabulary (Sáfár & Kormos, 2008), was coded as such and subsequently excluded from the analysis for the predictive validity of aptitude for either grammar or vocabulary.

Analysis

Three sets of analyses were performed to address the research questions respectively. The first set involved aggregation of the correlation coefficients for the associations between aptitude and other cognitive and affective variables, including motivation, anxiety, and intelligence. The second set aimed to examine how aptitude relates to working memory (EWM and PSTM). The third set of analyses pertains to the predictive validity of aptitude, that is, whether learners' aptitude scores could be used to forecast their L2 attainment. The analyses include effect size aggregation for the predictive validities of aptitude and aptitude components for general L2 proficiency and the specific aspects of L2 learning including the learning of L2 knowledge (grammar and vocabulary) and L2 skills (listening, speaking, reading, and writing). A moderator analysis was performed using a Q_b test to probe whether aptitude had differential predictive validities for high school and university language learners—the two groups of learners that were most investigated in the included studies.

A confidence interval was calculated for each mean effect size. A narrow interval indicates a robust effect, and an interval that includes zero means that the effect is statistically nonsignificant ($p > .05$) or that the possibility that there is a null effect in the population cannot be ruled out. Publication/availability bias was assessed for the results on aptitude-learning associations through a funnel plot that displays the distribution

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

of effect sizes, in combination with a trim-and-fill analysis that shows the missing studies and adjusted mean effect size as well as a Fail-Safe N analysis that identifies the total number of studies needed to nullify the observed effect or to increase the p value above .05. Publication bias refers to the fact that studies that report significant effects and large effect sizes are more likely to be published (due to the space limit, the results of the analysis are placed in supplementary material).

A meta-analysis can be performed using a random-effects model, which allows the population effect to vary across studies, or a fixed-effects model, which is premised on the assumption that the included studies share a common effect size. Following Borenstein, Hedges, Higgins, and Rothstein's (2009) recommendation, this meta-analysis used a mixed-effects model where within-group effects were computed using a random-effects model because of the methodological heterogeneity of the included studies. Between-group effects were examined through a fixed-effects model to test whether the effect size variations could be explained by the systematic influence of substantive features of the primary research (also see Fischer & Boer, 2011). Results of heterogeneity tests (Q_{within}) may serve as basis for decisions regarding model selection and moderator analysis, but such tests typically have low statistical power (Cross, Copping, & Campbell, 2011; Oswald & Plonsky, 2010) and were therefore not conducted in this meta-analysis.

The following steps were taken to ensure the robustness of the results and minimize the potential limitations of meta-analysis. First, each effect size was weighted proportional to its sample size (Norris & Ortega, 2000) so that those relating to smaller samples carried less weight. Second, an analysis was not performed if the number of studies was lower than three, although theoretically meta-analysis can be applied to as few as two studies because other forms of synthesis "are less transparent and/or are less likely to be valid" (Valentine, Pigott, & Rothstein, 2010, p. 215). Second, in order to prevent the *oranges and apples* problem, that is, mixing effect sizes relating to different constructs, hybrid measures of aptitude (e.g., MLAT 1 + MLAT 5) and of L2 achievements (vocabulary + grammar) were not amalgamated with measures with transparent underlying constructs. Third, with regard to the interpretations of effect sizes, most meta-analyses have followed Cohen's (1988) criteria for small (r = .10), medium (r = .30), and large (r = .50) effects, but Plonsky and Oswald (2014) pointed out that effect sizes in SLA tend to be inflated, and they recommended more conservative benchmarks (.25, .40, and .60 for small, medium, and large effects, respectively). In this meta-analysis, both criteria were considered, and the interpretations of the results were also discussed in comparison with the mean effect sizes of other meta-analyses in SLA and cognitive psychology. Finally, in the analysis for aptitude-learning associations, Fischer Z was used as the unit of analysis due

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

to its better statistical properties such as normal distribution and stable variance, but it was transformed back to *r* when the results were reported.

RESULTS

Methodological Features of the Primary Research

The whole dataset consisted of 66 studies that reported data contributed by 109 unique samples (including 34 samples from the MLAT Manual), and the sampled population consisted of 13,035 L2 learners. Table 2 provides a summary of the methodological information of the included studies, and the details of the studies, which were coded for 10 methodological features, are available in the online supplementary material that accompanies this article. As Table 2 shows, among the 66 retrieved studies, 47 (including the MLAT Manual) are published, and 19 are PhD and MA dissertations. The MLAT and its adapted versions in other languages (QMLAT) were used in 57 of the included studies, and the PLAB appeared in nine studies; other aptitude measures used in the primary research include VORD, LLAMA, and the Language Analysis Test (a measure of language analytic ability). The number of studies examining the relationships between aptitude, on one hand, and other individual difference variables and working memory, on the other, ranges from 7 to 12. As regards the criterion variable, 27 studies investigated how general L2 proficiency was correlated with aptitude and aptitude components, 22 explored the learning of L2 knowledge (n = 14 for grammar and n = 8 for vocabulary), and 46 related to the learning of different L2 skills (n = 17, 12, 9, and 8 for speaking, reading, listening, and writing, respectively). In terms of the instructional context, most of the studies were carried out in high school (n = 22) or university (n = 26) language programs, and only a small number of studies were conducted in other contexts. Finally, it can be seen that the majority of the studies (n = 49) were conducted with L1 English speakers who were learners of French (n = 24), Spanish (n = 20), and German (n = 12), and the studies were primarily undertaken in foreign language settings (n = 53) where the target language is not the language spoken in the community, rather than in second language settings (n = 10) in which the target language is also the language of the community.

Meta-Analytic Results

The Relationship between Aptitude and the Covariates. The first research question concerns the validity of the construct of language aptitude as

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

Table 2. Characteristics of the retrieved studies

Methodological Features			*N*
Publication type			
	Published studies[a]		51
	Dissertations		15
	Total		66
Aptitude measures			
	MLAT		57
	QMLAT[b]		
		GTT (Japanese)	1
		HUNLAT (Hungarian)	3
		LABJ (Japanese)	4
		MLAT in Dutch	1
		TALV (French)	1
	PLAB[c]		9
	LLAMA		2
	VORD		1
Covariates			
	Motivation		8
	Anxiety		7
	Intelligence		8
Working memory			
	PSTM		7
	EWM		12
Criterion			
	Proficiency		27
	L2 knowledge		
		Grammar	14
		Vocabulary	8
	L2 skills		
		Listening	9
		Reading	12
		Speaking	17
		Writing	8
Instructional context			
	High school		22
	University		26
	Elementary		2
	Middle school		3
	Intensive training		6
	Naturalistic		3
	Others		4

Continued

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

Table 2. continued

Methodological Features		*N*
L1		
	English	49
	Other languages	16
L2		
	English	11
	French	24
	German	12
	Spanish	20
	Others (NR,[d] mixed, and other languages)	15
Research setting		
	Foreign language	54
	Second language	12
	Not reported	1

[a] The MLAT Manual was coded as one study although it contributed data based on multiple samples.
[b] Adapted MLAT and subtests.
[c] Adapted PLAB and subtests.
[d] Not reported.

manifested through its associations with other individual difference variables such as motivation, anxiety, and intelligence. The results are presented in Table 3, including the mean effect sizes, the number of effect sizes for each analysis (*k*), and the 95% confidence interval for each mean effect size. For each covariate, the analysis was conducted for all aptitude measures, including whole test batteries, subtests, and different combinations of subtests (e.g., MLAT 1 + MLAT 5 in Winke [2013]Winke [2013]W []), and for the MLAT (total scores) only, to determine whether similar results and patterns were obtainable for different aptitude measures.

The analyses showed the following results. First, the associations between aptitude and motivation were weak: $r = .14$, 95% CI = .07/.21, for all aptitude measures including whole aptitude tests and subtests; $r = .16$, 95% CI = .05/.27, for the MLAT. Second, there was a negative correlation

Table 3. The correlations between aptitude and other variables

	All Aptitude Measures			*MLAT Total*		
Variables	*r*	*k*	*95% CI*	*r*	*k*	*95% CI*
Motivation	.14	17	.07, .21	.16	4	.05, .27
Anxiety	-.17	16	-.26, -.08	-.35	3	-.53, -.13
Intelligence	.50	24	.43, .55	.64	6	.55, .71

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

between aptitude and anxiety, and the MLAT showed a larger mean effect size, r = -.35, 95% CI = -.53/-.13, than mixed aptitude measures, r = -.17, 95% CI = -.26/-.08. Third, aptitude was strongly correlated with intelligence, and the result was found for both mixed aptitude measures, r = .50, 95% CI = .43/.55, and the MLAT, which showed a larger effect size, r = .64, 95% CI = .55/.71.

The Relationship between Aptitude and Working Memory. The second research question investigates the links between language aptitude and working memory. Based on the available data, analyses were conducted for aptitude as a holistic construct measured by whole test batteries and for the three components of aptitude (phonetic coding, language analytic ability, and rote memory) measured using the related subtests. The results, which are reported in Table 4, show that EWM was moderately correlated with aptitude as a whole, r = .37, 95% CI = .26/.46, and with phonetic coding ability, r = .33, 95% CI = .21/.44; it had weaker associations with language analytic ability, r = .22, 95% CI = .09/.33, and with rote memory, r = .25,95% CI = .09/.40. Overall, PSTM showed either weak (r = .16, 95% CI = .04/.27 for aptitude as a whole, and r = .24,95% CI = .02/.61 for rote memory) or nonsignificant correlations (r = .33, 95% CI = -.01/.61 for phonetic coding, and r = .02, 95% CI = -.22/.26 for language analytic ability) with language aptitude and its components.

Aptitude and L2 Learning. The final research question seeks to explore the predictive validity of aptitude, that is, how L2 learners' aptitude scores can be used to predict their learning achievement. The results are reported according to whether the studies examined the relationships between aptitude and general L2 proficiency or between aptitude and specific aspects of L2 learning, which were further categorized as L2 knowledge (grammar and vocabulary) and L2 skills (listening, speaking, reading, and writing). With respect to general L2 proficiency, effect sizes were aggregated for all measures of a construct as well as particular measures (MLAT, QMLAT, PLAB), if there were more than three effect sizes for a particular measure, for the purpose of examining whether

Table 4. The relationship between aptitude and working memory

	Aptitude			*Phonetic Coding*			*Language Analytic Ability*			*Rote Memory*		
	k	*r*	*95% CI*	*k*	*r*	*95% CI*	*k*	*r*	*95% CI*	*k*	*r*	*95% CI*
EWM	7	.37	.26, .46	4	.33	.21, .44	9	.22	.09, .33	5	.25	.09, .40
PSTM	4	.16	.04, .27	3	.33[a]	-.01, .61	7	.02[a]	-.22, .26	5	.24	.02, .61

[a] The confidence interval for the effect size crosses zero and is therefore statistically nonsignificant.

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

different aptitude measures had similar or different predictive validity. However, due to the limited amount of research on the associations between aptitude/aptitude components and the learning of L2 knowledge and skills, effect size aggregation was only conducted for the constructs (aptitude and its components), not for different measures of the constructs.

Table 5 displays the results for the relationship between aptitude/aptitude components and general L2 proficiency. As can be seen, overall aptitude as measured by whole test batteries was a strong predictor for L2 learning, r = .49, 95% CI = .45/.54, and this constitutes a robust finding given the large number of effect sizes (n = 53) involved in the analysis. Among the three aptitude components, phonetic coding ability appeared to be a stronger predictor than the other two components, and rote memory was the weakest predictor. The two subtests of the MLAT and Quasi-MLATs—Number Learning and Spelling Clues—were also significant predictors of L2 proficiency, although theoretically there is no transparent matchup between these two subtests and the three aptitude components. The results also showed that aptitude measures other than the MLAT, such as the QMLATs, showed lower predictive validities, and the disparity was more evident for overall aptitude and rote memory. Furthermore, in light of the fact that most of the studies were conducted with either high school or university foreign language learners, a post

Table 5. Aptitude and L2 proficiency

Construct	*Measure*	*k*[b]	*r*	*95% CI*
Aptitude	All Measures[a]	53	.49	.45, .54
	MLAT	46	.51	.47, .55
	QMLAT[c]	3	.22	.09, .34
Phonetic Coding	All Measures	9	.45	.21, .64
	MLAT 2	5	.53	.21, .75
Language Analytic Ability	All Measures	16	.38	.28, .48
	MLAT 4	9	.38	.29, .46
	QMLAT 4	3	.30	.17, .41
	PLAB 4	3	.26	.15, .35
Rote Memory	All Measures	6	.24	.33, .54
	MLAT 5	9	.30	.24, .37
	QMLAT 5	3	.04[ns]	-.19, .27
Number Learning	All Measures	6	.44	.33, .54
	MLAT 1	5	.40	.31, .49
Spelling Clues	All Measures	8	.38	.10, .61
	MLAT 3	6	.35[ns]	-.02, .63

Note. ns: The effect size is statistically nonsignificant.
[a] All aptitude measures.
[b] Analysis was not performed for variables with less than three effect sizes.
[c] Quasi-MLAT: adapted MLAT in other languages.

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

hoc analysis was performed to compare the mean effect sizes associated with the two instructional settings. The *Q* test showed a significant difference, $Q_b = 10.05$, $p < .001$; the mean effect size for high school learners ($k = 24$, $r = .54$, 95% CI = .47/.60) was larger than that for university learners ($k = 21$, $r = .40$, 95% CI = .35/.45).

The results for the role of aptitude and different aptitude components in the learning of L2 knowledge and skills are presented in Table 6, which can be inspected in two ways. One is to examine the effect sizes horizontally, that is, to compare the effect sizes in each row to identify whether a construct or predictor variable (aptitude and aptitude components) is correlated with each outcome variable (L2 knowledge and skills) in different ways. Another way is to view the effect sizes vertically, that is, to compare the effect sizes in each column to determine whether an outcome/criterion variable is associated with each predictor in different ways. A horizontal inspection of the results shows that (1) aptitude measured through whole test batteries showed moderate associations with the criterion variables (with *r* values above .30) except for a rather weak association with vocabulary learning ($r = .15$) and a nonsignificant correlation with L2 writing; (2) phonetic coding showed a stronger correlation with vocabulary ($r = .38$) than with other outcome variables, and it had a weak correlation with listening comprehension ($r = .12$); (3) language analytic ability was a stronger predictor for grammar learning ($r = .39$) and reading comprehension ($r = .35$) than for other forms of achievement; (4) in comparison with other constructs, rote memory was not a strong (with effect sizes around or below .20) or consistent (two of the effect sizes were nonsignificant) predictor for L2 knowledge and skills; and (5) with regard to the two miscellaneous aptitude components, number learning demonstrated a higher correlation with L2

Table 6. Aptitude and the learning of L2 knowledge and skills

	L2 Knowledge		*L2 Skills*			
Construct	Grammar	Vocabulary	Listening	Speaking	Reading	Writing
Aptitude	.32[a] (9 [b])	.15 (4)	.30 (4)	.37 (12)	.39 (10)	.34[ns] (5)
Phonetic Coding	.22 (4)	.38 (6)	.12 (8)	.26 (11)	.28 (8)	.26[ns] (5)
Language Analytic Ability	.39 (12)	.29 (9)	.25 (10)	.24 (15)	.35 (11)	.20[ns] (9)
Rote Memory	.19 (8)	.20[ns] (7)	.21 (8)	.19 (11)	.20 (10)	.12[ns] (5)
Number Learning	N/A[c]	.34 (4)	N/A	.32 (6)	.29 (4)	.42 (3)
Spelling Clues	.29 (7)	.36 (5)	.29 (3)	.34 (7)	.44 (7)	.42 (3)

Note. ns: The effect size is nonsignificant, that is, the confidence interval includes zero or $p > .05$.
[a] Correlation coefficient (*r*).
[b] Number of effect sizes (*k*).
[c] N/A: analysis not performed because the number of effect sizes is below three ($k < 3$).

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

writing than with other outcome variables, and spelling clues was more strongly correlated with reading comprehension and writing.

Examining the results in Table 6 vertically from the perspective of the outcome variable reveals the following patterns: (1) grammar learning was better predicted by language analytic ability than by other aptitude components; (2) for vocabulary learning, phonetic coding, number learning, and spelling clues were stronger predictors and overall aptitude was the weakest predictor; (3) the weakest predictor for listening comprehension was phonetic coding; (4) speaking showed the strongest correlation with overall aptitude among all the predictor variables, followed by the two miscellaneous components; (5) reading comprehension was more correlated with overall aptitude, language analytic ability, and spelling clues than with other components; and (6) the effect sizes for writing were statistically nonsignificant and mostly small except for those relating to number learning and spelling clues—the two miscellaneous aptitude components.

DISCUSSION

This study was undertaken to provide a comprehensive picture of the construct of language aptitude in Carroll's conceptualization by synthesizing all empirical studies reporting data about its associations with other cognitive and affective variables and with different aspects of L2 learning; also under investigation is the associations between aptitude and a possible aptitude component—working memory. Language aptitude was found to be largely separate from motivation, negatively correlated with anxiety, and strongly associated with intelligence. EWM, which performs the dual task of information storage and processing, had stronger links with aptitude than PSTM, which is only responsible for the temporary storage of information. Regarding its predictive validity, aptitude measured by whole test batteries was a strong predictor of L2 learning, but its predictive power varied across instructional contexts and outcome variables. This section discusses how the results facilitate our understanding about the construct of aptitude.

How Does Aptitude Relate to Other Individual Difference Variables?

Motivation. The correlation between aptitude and motivation is weak and negligible (r = .14, .16, for all aptitude measures and the MLAT, respectively), which suggests that they are relatively separate constructs. The result corroborates the finding of several factor analytic studies (Clément & Kruidenier, 1985; Gardner, Tremblay, & Masgoret, 1997;

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

Gordon, 1980; Sparks, Patton, Ganschow, & Humbach, 2011), namely that L2 learners' aptitude scores and their responses to motivation surveys cluster around different factors. The finding undermines the speculation that aptitude may have a positive influence on motivation, that is, learners with higher aptitude may be more motivated in learning a second language. The fact that aptitude and motivation are relatively separate constructs also suggests that the two variables may account for unique variance in L2 learning. This hypothesis seems to have been confirmed in several studies, which showed that motivation was a significant predictor of L2 learning (Garcia, 1996; Sparks et al., 2009), although it was less predictive than aptitude. Masgoret and Gardner (2003) conducted a meta-analysis of the empirical studies that used the attitude/motivation test battery and found that motivation was significantly correlated with course grades, $r = .37$. However, the meta-analysis did not take into account the variance explained by other variables, so the unique amount of variance attributable to motivation in these studies is unclear. Notwithstanding the need for more research on the comparative effects of the two variables, it would seem justified to claim, based on existing evidence, that motivation is distinct from aptitude, and that it explains a significant amount of variance of L2 learning. However, it remains to be seen whether ultimate L2 success is still achievable with more motivation in the event of a deficiency of aptitude.

One caveat is that in most of the studies motivation is measured by means of Gardner's (1985) Attitude/Motivation Test Battery (AMTB), which emphasizes the general orientation to language learning such as integrative and instrumental motives and attaches less importance to the motives relating to the process of learning. The field of motivation has moved forward, new theories have been proposed, and new instruments have been developed. It would be interesting to ascertain how aptitude relates to the more recently proposed L2 motivational systems such as Dörnyei's (2005) L2 Self system.

Anxiety. The negative correlation between aptitude and anxiety suggests that learners with high aptitude have low anxiety, and vice versa. The finding was consistent across the related studies: all three effect sizes for the MLAT total and 14 out of the 16 for hybrid aptitude measures were in the negative direction, with the only two positive effect sizes being close to zero (leaning toward the negative side). The negative correlation is consistent with what previous research has shown about their opposite relationships with L2 learning: while aptitude has been found to be facilitative, anxiety has consistently been found to be negatively correlated with achievement (Dörnyei, 2005; Horwitz, 2001). A plausible explanation for the different roles of anxiety and aptitude in affecting L2 development is that the debilitating effect of anxiety occurs through cognitive interference, in that learners with high anxiety allocate

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

more attentional resources to task-irrelevant demands associated with apprehension about negative evaluation, tension, and so forth, whereas those with low anxiety are able to appropriate more resources to task-relevant demands (Tobias, 1985). The functioning of language aptitude, however, requires the use of attentional resources, which are depleted by anxiety. It is also possible, as Sparks and Patton (2013) argued, that anxiety is a consequence, rather than cause, of low aptitude. Sparks and associates (e.g., Sparks et al., 2009, 2011) found that language learning anxiety as measured by the FLCAS questionnaire (Horwitz, 2001) was predictive of L2 aptitude, L1 skills, and L2 achievements. They speculated that (1) the FLCAS is a measure of learners' perceptions of their language learning skills and an indirect measure of L2 proficiency, and (2) anxiety is a consequence of the deficiency in language aptitude rather than a determinant of L2 learning success. A compromise between these two different perspectives would lead to the hypothesis that there may exist a reciprocal relationship between aptitude and anxiety, that is, anxiety disrupts the functioning of aptitude and low aptitude causes anxiety.

One possibility, which is related to the "cognitive interference" interpretation, is that the negative correlation between aptitude and anxiety is partly due to test anxiety, that is, those who were more anxious about test taking scored lower on aptitude tests. Horwitz (2001) argued that language learning anxiety is separate from test anxiety because there was only a moderate correlation between them (r = .53, p = .001). However, the correlation coefficient referred to as moderate would be considered large by Cohen's standards (1988). Also, at least two of the items in the FLCAS, which has been the predominant measure of foreign language anxiety in the research, are directly related to test anxiety. Therefore, whether the negative correlation between aptitude and anxiety has to do with test anxiety is yet to be clarified. Finally, noteworthy is the fact that to date most studies have only investigated the correlations between aptitude and anxiety, which prevents us from making claims about the causal relationships between the two variables.

Intelligence. The strong, positive correlations (r = .64, .50 for the MLAT and hybrid aptitude measures, respectively) between aptitude and intelligence suggest a large overlap between the two constructs. Factor analytic studies also failed to find a clear separation between the two variables. For example, Granena (2012, 2013) found that L2 Spanish learners' scores on a test of general intelligence called GAMA loaded on the same factor as their scores on three subtests (sections B, E, F) of the LLAMA aptitude battery; intelligence and the three LLAMA components were therefore collectively called "explicit aptitude." Sparks et al. (2011) reported that MLAT 5 (Rote Learning) clustered with intelligence measured through the Test of Cognitive Skills. In Wesche, Edwards, and Well's (1982) study, a factor was generated that included MLAT 3 (Spelling Clues)

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

and the Verbal Meaning section of the PMA (Primary Mental Abilities) intelligence test. Sasaki (1996) performed a structural equation modeling analysis showing that the model with the best fit with the data was one where aptitude measured through the LABJ and intelligence measured using the Kyoto University New NX-15 underlay one second-order general cognitive factor, which in turn significantly predicted second language proficiency. These studies, together with this meta-analysis, show that aptitude and intelligence are not entirely separate, and one of the main reasons for this overlap is likely due to some similar components between intelligence and aptitude measures. For example, most intelligence tests consist of subtests of vocabulary and memory, which are also included in traditional aptitude measures.

Despite the large overlap, aptitude and intelligence do not seem to be identical constructs, as the correlations between them, albeit strong, did not approach 1.0—a perfect correlation suggesting an isomorphic relationship (Ackermen, Beier, & Boyle, 2005)—or exceed the .75 threshold that would make it questionable to consider two related variables as distinct (Zikmund, Babin, Carr, & Griffin, 2013). Another reason for aptitude to be considered as a variable independent of intelligence or abilities for general academic learning is that the former is likely a stronger predictor of L2 learning, and there has been some empirical evidence for this argument. For example, Sparks et al. (2009) found that intelligence was only a weak predictor for L2 word decoding, whereas aptitude was a strong predictor for all L2 skills and for general L2 proficiency. Theoretically, what distinguishes aptitude from intelligence, according to Carroll and Sapon (2002), are the "special abilities" (p. 24) for language learning: phonetic coding measured by MLAT 2 and language analytic ability by MLAT 4. However, as Carroll and Sapon agreed, given the overlap between the two constructs, it might be necessary to consider both aptitude and intelligence when making pedagogical decisions related to selection, guidance, and placement.

One issue of this stream of research is the diverse measures of intelligence that have been used, which partly reflects the uncertainty over what intelligence entails. As Wasserman (2012) pointed out in his account of the historical development of the research on intelligence, after nearly a century of research, a consensus has yet to be reached on the definition of the construct. To exacerbate the problem, most of the related studies in aptitude research treated intelligence as a single construct, lumping together a cluster of cognitive abilities including verbal and nonverbal intelligence as well as fluid (reasoning) and crystallized (knowledge) intelligence, some of which may have little to do with language learning. There is therefore a need for fine-tuned, in-depth investigation of the correlations between specific aspects of intelligence and aptitude.

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

How Does Aptitude Relate to Working Memory?

The extent to which working memory should be included as an aptitude component can be explored by identifying the relationship between the two constructs and the contribution of working memory to L2 achievement. As for the relationship between aptitude and working memory, this meta-analysis showed significant, consistent correlations between aptitude and EWM, but weak and/or nonsignificant correlations between aptitude and phonological working memory. In terms of the predictive power of the two types of working memory, in a meta-analysis by Linck, Osthus, Koeth, and Bunting (2014), complex tasks that were used to measure EWM were found to be better predictors of L2 proficiency than simple tasks that were used to measure PSTM. The stronger correlations between EWM and traditional aptitude and its stronger associations with L2 achievement seem to suggest that it is a more promising aptitude component than PSTM.

However, it seems premature and simplistic to jump to conclusions about which of the two types of short-term memory is more predictive of L2 achievement. In Linck et al.'s meta-analysis (2014), proficiency is a hybrid criterion variable consisting of both composite measures of general proficiency and discrete measures of specific aspects of learning such as L2 knowledge and L2 skills. It is possible that the two types of short-term memory may have differential predictive validity for different aspects of L2 learning. For example, PSTM has been postulated and found to be an especially useful device for vocabulary learning (Baddeley, Gathercole, & Papagno, 1998), and EWM is more useful in reading and listening comprehension where the ability to simultaneously store and process incoming stimuli is essential (Daneman & Merikle, 1996). There is also evidence that PSTM may be critical to the acquisition of implicit knowledge while EWM is more likely to be called upon in the learning of explicit knowledge (Révész, 2012).Therefore, there seems to be a need for more fine-grained empirical and/or meta-analytic research that affords a detailed picture of the links between different memory types and specific aspects of L2 learning as well as general proficiency.[5]

How Predictive Is Aptitude?

Aptitude and General L2 Proficiency. There was a strong association (r = .49) between aptitude as a whole construct measured through full-length tests and general L2 proficiency operationalized as learners' course/exam grades or scores on standardized tests such as TOEFL.

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

A correlation of .49 means approximately 25% of the variance of L2 learning is accounted for by language aptitude. The correlation is considered strong based on Cohen's (1988) benchmarks (close to strong by Plonsky and Oswald's [2014] standards) and in comparison with the predictive power of other individual difference variables in SLA and other fields. For example, L2 achievement was found to be correlated with motivation at .37 (Masgoret & Gardner, 2003) and with working memory at .25 (Linck et al., 2014); the correlation between intelligence and academic performance was approximately .5 (Neisser et al., 1996). However, the strong predictive validity of aptitude should be interpreted in relation to the instructional contexts of the related studies. Among the 53 samples that contributed data for the correlations between overall aptitude and general L2 proficiency, 34 are from the MLAT Manual where the reported data were based on audiolingual classes popular in the 1950s, and 39 are based on studies conducted before the 1990s when communicative/task-based instruction was not widely implemented; 50 are from foreign language classes in North America (48 in the United States) and only three from contexts outside of North America (Gordon, 1980; Keigtes, 1986; Sáfár & Kormos, 2008). Furthermore, all data are from foreign language settings where the target language is not the primary language outside of class and none from second language settings where the target language is the primary language of the community. It remains to be seen whether the same results can be obtained in other contexts and with learners receiving more communicative instruction. Unfortunately, due to a lack of research conducted in different instructional settings, it was impossible to perform a moderator analysis that investigates how aptitude relates to achievement in contexts other than audiolingual classes that fit the mechanism of traditional aptitude.

The results also showed that aptitude was more strongly correlated with high school learners' general L2 proficiency than university learners' general L2 proficiency. This may indicate that traditional aptitude is more sensitive to initial stages than later stages of L2 learning as learners usually have had some foreign language instruction when they start to study at a university. However, it is also possible that the correlations for university learners were attenuated by their more homogeneous aptitude scores. An analysis of the learners' MLAT scores reported in the primary studies indicated that this was indeed the case: university foreign language learners' aptitude scores showed lower within-group variance, $Q_w = 98.56$, compared with high school learners' scores, $Q_w = 145.78$. The homogeneity of university learners' aptitude scores is perhaps related to the fact that they were likely a more select group than high school learners. It must be pointed out that although aptitude had a weaker association with university learners' L2 proficiency, the effect size was significant and moderate in magnitude ($r = .40$). Therefore, more primary research is needed to

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

test the hypothesis that aptitude is implicated differently at different stages of learning.

As to the predictive power of individual aptitude components, several findings are noteworthy. First, phonetic coding as measured by MLAT 2 Phonetic Script and its equivalents seems to show strong associations with general L2 proficiency and with vocabulary learning. Phonetic coding, the ability to identify sounds and to form and retain sound-symbol associations (Carroll, 1981), therefore, appears to be important in beginning foreign language classes from which most of the data of this study are derived. This hypothesis receives some indirect support from Natelson's (1975) large-scale study (N = 819) on the interaction between aptitude and learner experience, where MLAT 2 Phonetic Script was found to be the strongest predictor for the speaking scores of learners with little language learning experience; for learners with substantial experience, MLAT 1 Number Learning was the strongest predictor. Second, language analytic ability, for which there were 16 effect sizes, was the most studied, probably because of the assumption that the ability to "recognize the grammatical functions of words" (Carroll, 1981, p. 105) and to induce grammar rules is of special importance in learning a foreign language. However, in comparison with other subtests of the MLAT, MLAT 4 Words in Sentences, which is the predominant measure of language analytic ability, did not show substantially superior predictive validity for general L2 proficiency. Third, among the three aptitude components, rote memory was least predictive of L2 proficiency. However, we cannot lose sight of the fact that although less predictive than other components, rote memory was a significant predictor with an effect size of moderate magnitude. As reported earlier, even in Linck et al.'s (2013) validation study for the Hi-LAB, an aptitude battery developed for advanced L2 proficiency, rote memory was found to be a significant predictor. Fourth, the two remaining subtests of the MLAT, MLAT 1 Number Learning, and MLAT 3 Spelling Clues, which do not correspond with any of the three aptitude components and were therefore coded by retaining the names of the subtests, also proved to be significant predictors of L2 proficiency. Although Carroll (1981) was not sure what the two subtests measured, he speculated that they each tapped into several abilities: MLAT 1 may measure sound alertness, memory, and inductive learning ability, and MLAT 3 gauges phonetic coding and L1 vocabulary knowledge. The *versatility* of these subtests probably contributed to their predictive power, although this constitutes a weakness of traditional aptitude tests because of their emphasis on predictive validity at the cost of meaningful interpretations for the related findings.

Aptitude and Specific Aspects of L2 Learning. In terms of the associations between aptitude and the learning of L2 knowledge, the following findings stood out. First, language analytic ability was the strongest

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

predictor of grammar learning, which confirms the hypothesized link between this aptitude component and the learning of the morphosyntactic aspects of a second language. However, it is not clear whether language analytic ability has differential associations with the learning of explicit and implicit grammatical knowledge. Although none of the primary studies made a distinction between the two knowledge types, it appeared that the item-based written grammar tests in most of the studies encouraged or allowed the use of conscious knowledge, which traditional aptitude measures were sensitive to (Granena & Long, 2013). Second, aptitude as a whole showed the weakest correlation with learners' vocabulary scores (r = .15) among all criterion variables. The result invites the speculation that compared with other aspects of L2 learning, aptitude is likely not critical for vocabulary learning. The speculation remains to be confirmed in further research because of the following two caveats: (1) among the four effect sizes, two are based on QMLATs, which had consistently lower validity than the MLAT, and (2) despite the weak predictive validities of overall aptitude and rote memory for vocabulary learning, phonetic coding, language analytic ability, and the two miscellaneous subtests of the MLAT all showed moderate correlations with vocabulary learning. Furthermore, it is also possible that vocabulary learning requires abilities other than those measured in traditional aptitude tests, such as PSTM, which has been found to be a consistent predictor of vocabulary learning (e.g., Engel de Abreu & Gathercole, 2012; Hummel, 2009; Speciale, Ellis, & Bywater, 2004).

Equally challenging is to interpret the results for the role of aptitude in learning the four L2 skills. To begin with, among the three aptitude components, phonetic coding was the weakest predictor for listening comprehension. This is somewhat unexpected given the importance of the ability to recognize sounds and establish sound-symbol associations in processing auditory stimuli. One possible explanation is that in listening comprehension, learners tend to prioritize meaning over form, and therefore the ability for top-down processing that involves the use of schematic knowledge and contextual clues is perhaps more important than phonetic coding, which plays a central role in bottom-up processing that relies more on the ability to decode sounds or segments of sounds. As pointed out by Juffs and Harrington (2012), compared with native speakers, nonnative speakers are more likely to resort to top-down processing to compensate for their linguistic deficiency.

Moving on to speaking, the finding that overall aptitude was a stronger predictor has two implications: (1) although aptitude tests do not include measures of production skills, they are predictive of oral production; and (2) language aptitude is likely also important in the acquisition of implicit knowledge that is capitalized on in learners' performance of oral tasks (Ellis, 2005; Erlam, 2006), although this claim needs to be further investigated in carefully designed studies where implicit and

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

explicit knowledge are properly conceptualized/measured. Another interesting finding is that phonetic coding ability did not show superior predictive power for speaking, which is somewhat surprising given its hypothesized link with the ability for mimicry (Carroll & Sapon, 2002), which is in turn a predictor of foreign accent (Piske, MacKay, & Flege, 2001). The finding may be interpreted as proof for the dissociation between phonetic coding and mimicry because the former involves sound comprehension while the latter relates to sound production. It is also possible that the ability for mimicry is not important for oral performance, which is often evaluated holistically instead of merely on the basis of accurate articulation of sounds.

Next, overall aptitude, language analytic ability, and spelling clues appeared to be better predictors of reading comprehension than phonetic coding, rote memory, and number learning. What is common about the less predictive variables is that the cognitive skills they measure (auditory alertness, phonetic sensitivity, and rote memory) do not implicate the functions of the central executive that are responsible for semantic processing during reading comprehension; in contrast, these functions have stronger presence in overall aptitude and language analytic ability. The predictive power of spelling clues might be attributable to its importance in vocabulary learning, which in turn facilitates reading comprehension.

One conspicuous finding is that aptitude and its three components were nonsignificant and weak predictors for L2 writing, indicating that writing may require a different set of skills from those measured in traditional aptitude tests. The importance of content and organization in evaluating learners' written performance may account for the low predictive validity of traditional aptitude tests for this aspect of L2 learning. However, because most studies did not report how L2 writing was evaluated, this explanation is speculative. Also, in light of the important role played by aptitude in learning the linguistic aspects of L2 competence, which are also important in the evaluation of written performance, the claim that language aptitude is dissociated from L2 writing competence cannot be theoretically justified. This, along with the fact that the number learning and spelling clues subtests were significant predictors of L2 writing, is justification for more research into the unique and comparative effects of traditional aptitude and skill-specific abilities on L2 written performance.

Finally, two common patterns that emerged across all outcome variables need to be discussed. One is that rote memory was almost invariably the weakest predictor among all aptitude components/subtests. This contributes to the rationale for considering working memory an alternative for rote memory as an aptitude component. However, there needs to be more theoretical elaboration of the relationship between working memory and traditional aptitude components and more empirical

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

research into the predictive power of working memory in comparison with rote memory and other aptitude components. If both working memory (EWM and/or PSTM) and rote memory explain a unique portion of the variance of SLA or they contribute to the learning of different aspects of SLA, it is advisable to include both as aptitude components. The other noteworthy pattern is the heterogeneity of aptitude measures (MLAT, QMLATs, PLAB, LLAMA, VORD, etc.) used in the primary research to examine how aptitude relates to the learning of L2 knowledge and L2 skills. Compared with the findings for general L2 proficiency, which are predominantly based on the MLAT, those for the specific aspects of L2 learning are likely less robust. Notwithstanding, a summative, synthetic review of all related empirical research on the same topic affords a more objective, holistic picture of the population effect of the construct under investigation.

CONCLUSION

Using meta-analytic techniques, this study has investigated the theoretical claims about traditional language aptitude by synthesizing the empirical findings of the research that has been conducted over the past five decades since the publication of the most influential aptitude test, the MLAT (Carroll & Sapon, 1959). This study is apposite given that interest in cognitive aptitudes in SLA research is growing and there is a need to reexamine the construct and reorient the research. The study obtained the following findings. First, aptitude was found to be a relatively distinct construct: it is different from motivation that would appear to make a unique contribution to achievement (Bialystok & Fröhlich, 1978; Garcia, 1996; Keitges, 1986); it has a negative correlation with anxiety, which may be attributable to the cognitive interference of anxiety and the possibility that low aptitude causes high anxiety (MacIntyre, 1995); and it has a large overlap and yet is not isomorphic with intelligence, which has proven less predictive for language learning (Granena, 2012; Sparks et al., 2009). Second, EWM that taps both the storage and processing functions of short-term memory seems to be a more viable component of aptitude than PSTM because the former is more beneficial for L2 attainment (Linck et al., 2014) and bears stronger associations with aptitude and its components. However, this statement is subject to whether PSTM has superior predictive validity for certain aspects of L2 learning such as the learning of implicit knowledge, which has not received sufficient attention in the research. Third, aptitude as a whole is strongly predictive of L2 proficiency and is more predictive of high school students' L2 achievements than university students' L2 achievements. Overall aptitude is also a consistent predictor for the learning of L2 knowledge and L2 skills, except for vocabulary learning

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

and L2 writing. Fourth, as regards the predictiveness of different aptitude components, language analytic ability is a good predictor for grammar learning; phonetic coding is unimportant for listening comprehension and speaking; and rote memory is consistently less predictive than other aptitude components.

This research synthesis also indicates that there are issues to be addressed in future empirical research. First, as mentioned earlier, the majority of the primary studies are conducted in the United States with beginning foreign language learners who received form-oriented instruction, and there is a lack of research on the relevance of aptitude for advanced learners and for the learning that happens in more communicative classes. Second, traditional aptitudes are relevant in instructed L2 learning which favors explicit learning, and there is a need to develop measures or include components that are sensitive to implicit learning that tends to occur in nontutored settings (Granena, 2015). Third, existing aptitude measures demonstrated differential predictive validities, and it is necessary to cross-validate the tests to determine to what extent they tap the same construct. Fourth, previous research prioritized the predictive validity of aptitude, and there is a relative lack of more fine-grained, theory-guided research on how aptitude and clusters of cognitive abilities are implicated at different stages of SLA (Skehan, 2012) and under different learning conditions (Robinson, 2005). Fifth, although there appears to be a large amount of research on aptitude as a research domain, there is insufficient research on the relationships between aptitude and specific aspects of learning. The scattering of studies among subareas of a research domain may partly be driven by the desire to be different from previous research or to be original, but from the perspective of research synthesis, the value of replication research cannot be exaggerated in order for robust conclusions to be reached about a certain construct. To the extent that there has been insufficient research and that there is theoretical and empirical basis for the differential predictive power of aptitude and aptitude components for different aspects of L2 learning, the researcher makes a call for more research in this regard. Sixth, similar to aptitude—the predictor variable, the criterion variables, and covariates are also measured or operationalized in different ways in the primary research. For example, the measures used to assess general L2 proficiency include course grades, exam scores, multiple-choice tests, and standardized tests (such as TOEFL, the Michigan Test, etc.). The extent to which the results on these measures are comparable is unknown. It is also uncertain whether some of the measures are accurate reflections of learners' language proficiency. Course grades, for example, may include students' attendance and in-class participation, and the percentage they account for in the final grade varies considerably across classes and schools.

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

Finally, it must be pointed out that aptitude is purported to predict learning rate, which refers to the amount of learning one achieves within a certain period of time in comparison with peers under the same learning condition while holding previous knowledge constant (i.e., gains from time A to time B). However, in most predictive studies, the criterion variable is learners' attainment measured at fixed time points without considering previous knowledge and the duration of study (i.e., gains observed at time B). There are several ways to address this issue. In Carroll's validation research (Carroll & Sapon, 2002), learners with previous L2 knowledge/experience were excluded, so mid-term or final exam grades were an indicator of learning rate. However, given that it is not always possible to recruit learners who have not received any language instruction or have no prior exposure to a foreign language, a more realistic approach is to account for the effect of previous knowledge by using gain scores as the dependent variable, or by including pretest scores as a predictor or covariate to determine the amount of variance explained by aptitude over and above that explained by learners' prior knowledge.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit http://dx.doi.org/S027226311500042X.

Received 8 August 2015
Accepted 23 October 2015
Final Version Received 24 November 2015

NOTES

1. The mean effect sizes for predictive grammar studies were also included in this study to provide a clear picture of the differential predictive power of aptitude for the learning of different aspects of the L2 system and different L2 skills (i.e., aptitude's correlations with grammar vis-à-vis vocabulary and the four skills: listening, speaking, reading, and writing). However, the effect sizes were not subjected to moderator analysis, nor were the results for grammar focused on in the discussion.

2. These variables were included in this meta-analysis because they are cognitive and affective individual difference variables that, together with aptitude, are hypothesized to jointly and uniquely account for the variance of SLA (see Dörnyei, 2005). These are also the variables investigated in the primary studies. Meta-analysis is "inductive" in that it aggregates results and findings that are available in the primary research, and therefore in a sense the selection of variables was based on availability rather than preference.

3. One reviewer argued that it is not justifiable to include working memory as an aptitude component based on Carroll's definition (e.g., 1990) because it is involved in many higher order cognitive skills, thus failing to meet the domain-specific criterion of language aptitude. However, as explained in the article, Carroll's conceptualization of aptitude has been challenged, and given the undeniable importance of working memory

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

in SLA, leading aptitude researchers (DeKeyser & Koeth, 2011; Robinson, 2002; Skehan, 2012) have made repeated calls for including this cognitive ability as an aptitude component. It is beyond the scope of this article to make a decision for inclusion or exclusion; rather, the purpose is to draw on the data available in the primary research to clarify the relationship between working memory and traditional aptitude.

4. Although course grades and standardized language proficiency tests are theoretically different, a Q_b test showed no significant differences between them in terms of their associations with global aptitude, r = .47, .45 for course grades and proficiency tests, respectively. Therefore they were conflated and coded as measures of general proficiency.

5. One reviewer pointed out that another approach to revamping the construct of aptitude is to consider measures that highly correlate with both traditional aptitude measures and L2 learning (e.g., EWM) as alternative measures of the same constructs, and include those that do not highly correlate with traditional measures but do correlate with L2 learning (e.g., PSTM) as new measures of different constructs that could be included in an expanded conceptualization of aptitude. This perspective, however, does not seem to fit the case of working memory, because the correlations between measures of EWM and traditional aptitude measures, albeit significant, were not high (r's < .40; this study) enough for the former to replace the latter. In the case of PSTM, it did show weaker correlations with traditional aptitude than EWM, but it also showed weaker predictive validity for L2 learning (Linck et al., 2014). The solution seems to be to examine the joint and unique contributions of traditional aptitude components and the two types of short-term memory and ascertain whether they account for a unique portion of L2 variance over and above the variance explained by traditional aptitude.

REFERENCES (studies included in the meta-analysis are marked with asterisks)

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

Ackermen, P., Beier, M., & Boyle, M. (2005). Working memory and intelligence. *Psychological Bulletin*, *131*, 30–60.

*Alderson, J. C., Clapham, C., & Steel, D. (1997). Metalinguistic knowledge, language aptitude and language proficiency. *Language Teaching Research*, *1*, 93–121.

Baddeley, A., Gathercole, S., & Papagno, C. (1998). Phonological loop as a language learning device. *Psychological Review*, *105*, 158–73.

*Bain, S. K., McCallum, R. S., Bell, S. M., Cochran, J. L., & Sawyer, S. C. (2010). Foreign language learning aptitudes, attitudes, attributions, and achievement of postsecondary students identified as gifted. *Journal of Advanced Academics*, *22*, 130.

*Bell, P. (2009). *Le cadeau* or *la cadeau*? The role of aptitude in learner awareness of gender distinctions in French. *The Canadian Modern Language Review*, *65*, 615–643.

*Bell, S. M., & McCallum, R. S. (2012). Do foreign language learning, cognitive, and affective variables differ as a function of exceptionality status and gender? *International Education*, *42*, 86–106.

*Bialystok, E., & Fröhlich, M. (1978). Variables of classroom achievement in second language learning. *The Modern Language Journal*, *62*, 327–336.

*Biedron, A., & Szczepaniak, A. (2012). Working memory and short-term memory abilities in accomplished multilinguals. *Modern Language Journal*, *96*, 290–306.

*Bond, K. (2011). *Examining morphological sensitivity and individual differences in the acquisition of agreement in a second language* (Unpublished MA thesis). University of Kansas.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* Chichester, UK: John Wiley & Sons.

*Burgos-Kohler, N. (2002). *The effect of a selected group of language learning strategies upon language development* (Unpublished PhD dissertation). University of Texas at Austin.

*Carpenter, H. S. (2008). *A behavioural and electrophysiological investigation of different aptitudes for L2 grammar in learners equated for proficiency level* (Unpublished PhD dissertation). Georgetown University.

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

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

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). Rowley, MA: Newbury House.
Carroll, J. (1990). Cognitive abilities in foreign language aptitude: Then and now. In T. Parry & C. Stansfield (Eds.), *Language aptitude reconsidered* (pp. 11–29). Englewood Cliffs, NJ: Prentice Hall.
Carroll, J., & Sapon, S. (1959). *Modern language aptitude test*. New York, NY: The Psychological Corporation/Harcourt Brace Jovanovich.
*Carroll, J., & Sapon, S. (2002). *Manual for the MLAT*. Second Language Testing. N. Bethesda, MD.
Carroll, J., Sapon, S., Reed, D., & Stansfield, C. (2010). *Manual for the MLAT*. Second Language Testing. N. Bethesda, MD.
Child, J. (1998). Language aptitude testing: Learners and applications. *Applied Language Learning*, *9*, 1–10.
Clément, R., & Kruidenier, B. (1985). Aptitude, attitude and motivation in second language proficiency: A test of Clément's model. *Journal of Language and Social Psychology 4*, 21–37.
*Cochran, J. L., McCallum, R. S., & Bell, S. M. (2010). Three A's: How do attributions, attitudes, and aptitude contribute to foreign language learning? *Foreign Language Annals*, *43*, 566–582.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.
*Cooper, C. (1964). Some relationships between paired-associates learning and foreign-language aptitude. *Journal of Educational Psychology*, *55*, 132–138.
Cronbach, J., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.
Cross, C., Copping, L., & Campbell, A. (2011). Sex differences in impulsivity: A meta-analysis. *Psychological Bulletin*, *137*(1), 97–130.
*Curtin, C., Avner, A., & Smith, L. A. (1983). The Pimsleur Battery as a predictor of student performance. *The Modern Language Journal*, *67*, 33–40.
Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*, 422–433.
de Graaff, R. (1997). The eXperanto experiment: Effects of explicit instruction on second language acquisition. *Studies in Second Language Acquisition*, *19*, 249–276.
*DeKeyser, R. (1993). The effect of error correction on L2 grammar knowledge and oral proficiency. *The Modern Language Journal*, *77*, 501–514.
*DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, *22*, 499–533.
DeKeyser, R., & Koeth, J. (2011). Cognitive aptitudes for second language learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 395–406). New York, NY: Routledge.
Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates.
*Ehrman, M. (1998). The modern language aptitude test for predicting learning success and advising students. *Applied Language Learning*, *9*, 31–70.
*Ehrman, M. E., & Oxford, R. L. (1995). Cognition plus: Correlates of language learning success. *The Modern Language Journal*, *79*, 67–89.
*Eisenstein, M. (1980). Childhood bilingualism and adult language learning aptitude. *Applied Psychology*, *29*, 159–172.
Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*, 141–172.
Engel de Abreu, P. M., & Gathercole, S. E. (2012). Executive and phonological processes in second-language acquisition. *Journal of Educational Psychology*, *104*, 974–986.
*Erlam, R. (2005). Language aptitude and its relationship to instructional effectiveness in second language acquisition. *Language Teaching Research*, *9*, 147–171.
Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, *27*, 464–491.
Fischer, R., & Boer, D. (2011). What is more important for national well-being: Money or autonomy? A meta-analysis of well-being, burnout, and anxiety across 63 societies. *Journal of Personality and Social Psychology*, *101*, 164–184.

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

*Fujii, A. (2005). *Individual differences in task performance: Aptitude profiles, orientation to form, and second language production in the EFL classroom* (Unpublished PhD dissertation). Georgetown University.
Ganschow, L., & Sparks, R. (1995). Effects of direct instruction in Spanish phonology on the native-language skills and foreign-language aptitude of at-risk foreign-language learners. *Journal of Learning Disabilities*, *28*, 107–120.
*Ganschow, L., & Sparks, R. (1996). Anxiety about foreign language learning among high school women. *Modern Language Journal*, *80*, 199–212.
*Garcia, R. T. (1996). *The relation among college student attitudes, motivation and aptitude to foreign language achievement: A test of Gardner's model* (Unpublished PhD dissertation). University of Houston.
Gardner, R. C. (1985). *Social psychology and second language learning: The role of attitudes and motivation*. London: Edward Arnold.
*Gardner, R. C., & Lambert, W. E. (1965). Language aptitude, intelligence, and second-language achievement. *Journal of Educational Psychology*, *56*, 191–199.
*Gardner, R., & Lysynchuk, L. (1990). The role of aptitude, attitudes, motivation, and language use on second-language acquisition and retention. *Canadian Journal of Behavioural Science*, *22*, 254–270.
*Gardner, R., Moorcroft, R., & Metford, J. (1989). Second language learning in an immersion program: Factors influencing acquisition and retention. *Journal of Language and Social Psychology*, *8*, 287–305.
*Gardner, R., Tremblay, P., & Masgoret, A. (1997). Towards a full model of second language learning: An empirical investigation. *The Modern Language Journal*, *81*, 344–362.
*Goodman, J. F., Freed, B., & McManus, W. J. (1990). Determining exemptions from foreign language requirements: Use of the modern language aptitude test. *Contemporary Educational Psychology*, *15*, 131–141.
*Gordon, M. E. (1980). *Attitudes and motivation in second-language achievement: A study of primary school students learning English in Belize, Central America* (Unpublished PhD dissertation). University of Toronto.
Granena, G. (2012). *Age differences and cognitive aptitudes for implicit and explicit learning in ultimate second language attainment* (Unpublished PhD dissertation). University of Maryland.
Granena, G. (2013). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning*, *63*, 665–704.
Granena, G. (2015). Cognitive aptitudes for implicit and explicit learning and information-processing styles: An individual differences study. *Applied Linguistics*, online. DOI: 10.1017/S0142716415000120
*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*, 311–343.
Grigorenko, E., Sternberg, R., & Ehrman, M. (2000). A theory-based approach to the measurement of foreign language learning ability: The Canal-F theory and test. *The Modern Language Journal*, *84*, 390–405.
Groth-Marnat, G. (2003). *Handbook of psychological assessment*. Hoboken, NJ: John Wiley & Sons.
*Harley, B., & Hart, D. (1997). Language aptitude and second language proficiency in classroom learners of different starting ages. *Studies in Second Language Acquisition*, *19*, 379–400.
*Harley, B., & Hart, D. (2002). Age, aptitude, and second language learning on a bilingual exchange. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 302–330). Amsterdam and Philadelphia: John Benjamins Publishing Company.
*Hauptman, P. C. (1971). A structural approach vs. a situational approach to foreign-language teaching. *Language Learning*, *21*, 235–244.
*Horwitz, E. K. (1980). *The relationship of conceptual level to the development of communicative competence in French* (Unpublished PhD dissertation). University of Illinois at Urbana-Champaign.
Horwitz, E. K. (2001). Language anxiety and achievement. *Annual Review of Applied Linguistics*, *21*, 112–126.
*Hsieh, C. (2004). *Relationships between non-linguistic variables in second language acquisition and the attained English language proficiency of Taiwanese university students* (Unpublished PhD dissertation). Lynn University.

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

*Hummel, K. 2009. Aptitude, phonological memory, and second language proficiency in nonnovice adult learners. *Applied Psycholinguistics*, *30*, 225–249.
Hwu, F., & Sun, S. (2012). The aptitude-treatment interaction effects on the learning of grammar rules. *System*, *40*, 505–521.
In'nami, Y., & Koizumi, R. (2010). Database selection guidelines for meta-analysis in applied linguistics. *TESOL Quarterly*, *44*, 169–184.
Juffs, A., & Harrington, M. (2012). Aspects of working memory in L2 learning. *Language Teaching*, *44*, 137–166.
*Keitges, D. J (1986). *Relationships between foreign-language aptitude, attitudes and motivation, and personality traits, and the attained English language proficiency of Japanese university students* (Unpublished PhD dissertation). University of Texas at Austin.
*Kormos, J., & Trebits, A. (2012). The role of task complexity, modality, and aptitude in narrative task performance. *Language Learning*, *62*, 439–472.
Krashen, S. (1981). Aptitude and attitude in relation to second language acquisition and learning. In K. Diller (Ed.), *Individual differences and universals in language learning aptitude* (pp. 155–175). Rowley, MA: Newbury House.
*Lado, B. (2008). *The role of bilingualism, type of feedback, and cognitive capacity in the acquisition of non-primary languages: A computer-based study* (Unpublished PhD dissertation). Georgetown University.
*Lalonde, R. N., & Gardner, R. C. (1984). Investigating a causal model of second language acquisition: Where does personality fit? *Canadian Journal of Behavioural Science 16*, 224–237.
*Lee, S. (2008). *Salience, frequency, and aptitude in the learning of unaccusativity in a second language: An input enhancement study* (Unpublished PhD dissertation). University of Hawaii.
*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*, 634–654.
Li, S. (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.
Li, S., Shintani, N., & Ellis, R. (2012). Doing meta-analysis in SLA: Practices, choices, and standards. *Contemporary Foreign Language Studies*, *384*, 1–17.
Linck, J., et al. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning*, *63*, 530–566.
Linck, J., Osthus, P., Koeth, J., & Bunting, M. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review*, *21*, 861–883.
Lipsey, M., & Wilson, D. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.
MacIntyre, P. (1995). How does anxiety affect second language learning? A reply to Sparks and Ganschow. *The Modern Language Journal*, *79*, 90–99.
*Mackey, A., Philp, J., Egi, Takako, Fujii, A., & Tatsumi, T. (2002). Individual differences in working memory, noticing of interactional feedback and L2 development. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 181–209). Amsterdam and Philadelphia: John Benjamins Publishing Company.
Masgoret, A., & Gardner, R. (2003). Attitudes, motivation, and second language learning: A meta-analysis of studies conducted by Gardner and associates. *Language Learning*, *53*, 123–163.
*McMullin, J. J. (1988). *The effects of gender and the total physical response strategy on the retention of French commands by seventh-grade black students* (Unpublished PhD dissertation). Temple University.
Meara, P. (2005). *LLAMA Language Aptitude Tests*. Swansea: Lognostics.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). New York, NY: Macmillan.
Miyake, A., & Friedman, N. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. Healy & L. Bourne (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 339–364). Mahwah, NJ: Erlbaum.
*Myer, B. J. (1985). *Relationships among selected measures of auditory foreign language aptitude and achievement, musical aptitude and experience, prior academic achievement,*

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

sex and handedness factors at the secondary school level (Unpublished PhD dissertation). Ohio State University.
*Natelson, E. R. (1975). *The predictive validity of each of the five parts of the modern language aptitude test: The effect of previous language training on MLAT scores: The correlation of MLAT scores and achievement in specific language groups* (Unpublished PhD dissertation). Georgetown University.
Neisser, U., Boodoo, G., Bouchard, T., Boykin, A., Brody, N., Ceci, S., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*, 77–101.
Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*, 417–528.
Oswald, F., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, *30*, 85–110.
*Parry, T., & Child, J. (1990). Preliminary investigation of the relationship between VORD, MLAT and Language Proficiency. In P. Thomas & C. Stansfield (Eds.), *Language aptitude reconsidered* (pp. 30–66). Englewood Cliffs, NJ: Prentice Hall Regents.
Petersen, C., & Al-Haik, A. (1976). The development of the Defence Language Aptitude Battery (DLAB). *Educational and Psychological Measurement*, *36*, 369–380.
*Phillips, S. K. (1988). *Predictive validity for the modern language aptitude test* (Unpublished PhD dissertation). Indiana University of Pennsylvania.
Pimsleur, P. (1966). *Pimsleur Language Aptitude Battery (PLAB)*. San Diego: Harcourt Brace Jovanovich.
Pimsleur, P., Reed, D., & Stansfield, C. (2004). *Pimsleur Language Aptitude Battery (PLAB)*. N. Bethesda, MD: Second Language Testing.
Piske, T., MacKay, I., & Flege, J. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics*, *29*, 191–215.
Plonsky, L., & Brown, D. (2015). Domain definition and search techniques in meta-analyses of feedback (Or why 18 meta-analyses of feedback have different results). *Second Language Research*, *31*, 267–278.
Plonsky, L., & Oswald, F. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.
Popham, J. (2000). *Modern educational measurement*. Needham, MA: Allyn and Bacon.
*Rachlin, S. (2005). *Reading, mathematics, language, second language aptitude scores and personality type as predictors of success in beginning second language study* (Unpublished PhD dissertation). Hofstra University.
Ranta, L. (2002). The role of learners' language analytic ability in the communicative classroom. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 159–180). Amsterdam: John Benjamins.
Révész, A. (2012). Working memory and the observed effectiveness of recasts on different L2 outcome measures. *Language Learning*, *62*, 93–132.
Robinson, P. (2001). Individual differences, cognitive abilities, aptitude complexes and learning conditions in second language acquisition. *Second Language Research*, *17*, 368–392.
*Robinson, P. (2002). Effects of individual differences in intelligence, aptitude and working memory on adult incidental SLA. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 212–266). Amsterdam and Philadelphia: John Benjamins Publishing Company.
Robinson, P. (2005). Cognitive abilities, chunk-strength, and frequency effects in implicit artificial grammar and incidental L2 learning: Replications of Reber, Walkenfield, and Hernstatd (1991) and Knowlton and Squire (1996) and their relevance for SLA. *Studies in Second Language Acquisition*, *27*, 235–268.
*Roehr, K., & Gánem-Gutiérrez, G. (2009). The status of metalinguistic knowledge in instructed adult L2 learning. *Language Awareness*, *18*, 165–181.
*Ross, S., Yoshinaga, N., & Sasaki, M. (2002). Aptitude-exposure interaction effects on Wh-movement violation detection by pre-and-post-critical period Japanese bilinguals. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 267–299). Amsterdam: John Benjamins.
*Sachs, R. (2010). *Individual differences and the effectiveness of visual feedback of reflexive binding in L2 Japanese* (Unpublished PhD dissertation). Georgetown University.
*Sáfár, A. & Kormos, J. (2008). Revisiting problems with foreign language aptitude. *International Review of Applied Linguistics in Language Teaching*, *46*, 113–136.

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

Sasaki, M. (1996). *Second language proficiency, foreign language aptitude, and intelligence: Quantitative and qualitative analyses*. New York, NY: Peter Lang.
Seipp, B. (1990). Anxiety and academic performance: A meta-analysis of findings. *Anxiety Research*, *91*, 27–41.
Sheen, Y. (2007). The effects of corrective feedback, language aptitude, and learner attitudes on the acquisition of English articles. In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp. 301–322). New York, NY: Oxford University Press.
Skehan, P. (2002). Theorizing and updating aptitude. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 69–94). Amsterdam: John Benjamins.
Skehan, P. (2012). Language aptitude. In S. Gass & A. Mackey (Eds.), *Handbook of second language acquisition* (pp. 381–395). New York, NY: Routledge.
Snow, R. E. (1991). Aptitude-treatment interaction as a framework for research on individual differences in psychotherapy. *Journal of Consulting and Clinical Psychology*, *59*, 205–216.
*Sparks, R. L., & Patton, J. (2013). Relations of L1 skills and L2 aptitude to L2 anxiety on the foreign language classroom anxiety scale. *Language Learning*, *63*, 870–895.
*Sparks, R. L., Ganschow, L., & Patton, J. (1995). Prediction of performance in first-year foreign language courses. *Journal of Educational Psychology*, *87*, 638–655.
*Sparks, R. L., Patton, J., Ganschow, L., & Humbach, N. (2009). Long-term relationships among early first language skills, second language aptitude, second language affect, and later second language proficiency. *Applied Psycholinguistics*, *30*, 725–755.
*Sparks, R. L., Patton, J., Ganschow, L., & Humbach, N. (2011). Subcomponents of second- language aptitude and second-language proficiency. *Modern Language Journal*, *95*, 253–273.
*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.
Sparks, R. L., Patton, J., Ganschow, L., Humbach, N., & Javorsky, J. (2006). Native language predictors of foreign language proficiency and foreign language aptitude. *Annals of Dyslexia*, *56*, 129–160.
*Sparks, R. L., Artzer, M., Ganschow, L., Siebenhar, D., Plageman, M., & Patton, J. (1998). Differences in native-language skills, foreign-language aptitude, and foreign-language grades among high-, average-, and low-proficiency foreign-language learners: Two studies. *Language Testing*, *15*, 181–216.
Speciale, G., Ellis, N. C., & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistic*, *25*, 293–332.
Suzuki, Y., & DeKeyser, R. (2015) Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language Learning*, online. DOI: 10.1111/lang.12138
Thompson, A. (2013). The interface of language aptitude and multilingualism: Reconsidering the bilingual/multilingual dichotomy. *Modern Language Journal*, *97*, 685–701.
Tobias, S. (1985). Test anxiety: Interference, defective skills, and cognitive capacity. *Educational Psychologist*, *20*, 135–142.
Valentine, J., Pigott, T., & Rothstein, H. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics*, *35*, 215–247.
Wasserman, J. (2012). A history of intelligence assessment: The unfinished tapestry. In D. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 3–55). New York, NY: The Guilford Press.
Wen, Z. (2015). Working memory in second language acquisition and processing: The phonological/executive model. In Z. Wen, B. Mailce, & A. McNeill (Eds.), *Working memory in second language acquisition and processing* (pp. 41–62), Bristol, UK: Multilingual Matters.
*Wesche, M., Edwards, H., & Wells, W. (1982). Foreign language aptitude and intelligence. *Applied Psycholinguistics*, *3*, 127–140.
Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology*, *84*, 608–618.
Williams, J. N. (2009). Implicit learning in second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *The new handbook of second language acquisition* (pp. 319–353), Bingley, UK: Emerald Group Publishing Limited.

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press

*Winke, P. (2005). *Individual differences in adult Chinese second language acquisition: The relationships between aptitude, memory, and strategies for language learning* (Unpublished PhD dissertation). Georgetown University.
*Winke, P. (2013). An investigation into second language aptitude for advanced Chinese language learning. *Modern Language Journal*, *97*, 109–130.
*Yoshizawa, K. (2002). *Relationships among strategy use, foreign language aptitude, and second language proficiency: A structural equation modelling approach* (Unpublished PhD dissertation). Temple University.
Zikmund, W., Babin, B., Carr, J., & Griffin, M. (2013). *Business research methods*. Mason, OH: Cengage Learning.

https://doi.org/10.1017/S027226311500042X Published online by Cambridge University Press