LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

Hi-LAB: A New Measure of Aptitude for High-Level Language Proficiency

Jared A. Linck, Meredith M. Hughes, Susan G. Campbell, Noah H. Silbert, Medha Tare, Scott R. Jackson, Benjamin K. Smith, Michael F. Bunting, and Catherine J. Doughty

University of Maryland Center for Advanced Study of Language

Few adult second language (L2) learners successfully attain high-level proficiency. Although decades of research on beginning to intermediate stages of L2 learning have identified a number of predictors of the rate of acquisition, little research has examined factors relevant to predicting very high levels of L2 proficiency. The current study, conducted in the United States, was designed to examine potential cognitive predictors of successful learning to advanced proficiency levels. Participants were adults with varying degrees of success in L2 learning, including a critical group with high proficiency as indicated by standardized language proficiency tests and on-the-job language use. Results from a series of group discrimination analyses indicate that high-level attainment was related to working memory (including phonological short-term memory and task set switching), associative learning, and implicit learning. We consider the implications for the construct of high-level language aptitude and identify future directions for aptitude research.

Keywords language aptitude; high-level L2 proficiency; cognitive abilities; individual differences; language assessment

---

The authors would like to thank Dave Dorsey, Greg Hancock, Mike Long, Steve Ross, Marianne Gullberg, Lourdes Ortega, and three anonymous reviewers for their insightful comments on previous versions of this manuscript, Renee Meyer for her longstanding and devoted support of this project, and Jen Janesh and Rachel Davis for their assistance with data collection. This material is based on work supported, in whole or in part, with funding from the United States Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the University of Maryland, College Park and/or any agency or entity of the United States Government.

Correspondence concerning this article should be addressed to Jared Linck (co-Principal Investigator), E-mail: jlinck@casl.umd.edu or Catherine Doughty (Principal Investigator), E-mail: cdoughty@casl.umd.edu. University of Maryland Center for Advanced Study of Language, 7005 52nd Ave, College Park, MD, 20742.

© 2013 Language Learning Research Club, University of Michigan
DOI: 10.1111/lang.12011

Introduction

Few, if any, adult learners achieve nativelike proficiency in second language (L2) comprehension or production. Estimates of the number of people who reach such high proficiency vary from zero (Abrahamsson & Hyltenstam, 2008, 2009; Long, 2005, 2007, in press) to a maximum of about 5% of learners (Selinker, 1972). Yet, some adult learners do achieve *near*-native global proficiency that is virtually indistinguishable from native speakers on certain tasks, or as judged by naïve raters (Birdsong, 2009; Bongaerts, 1999; Bongaerts, Mennen & van der Slik, 2000; Hyltenstam & Abrahamsson, 2003). Available evidence suggests that these rare individuals may have an aptitude for language learning (see Granena, 2012, for a recent review).

Although existing language aptitude tests predict learning at earlier stages (e.g., Carroll, 1985; Ehrman & Oxford, 1995), very little is known about the factors that predict successful high-level learning. Indeed, there has been little systematic investigation of adult learners who have achieved high-level proficiency. A few studies have found high achievers to be similar to their less successful counterparts with respect to personality profiles and linguistic experience, suggesting that highly successful learners possess a particular aptitude—or talent—for language learning (Abrahamsson & Hyltenstam, 2008; DeKeyser, 2000; Ioup, Boustagui, El Tigi, & Moselle, 1994). For example, Ioup and colleagues (1994) described one case study of an adult native speaker of English who acquired Egyptian Arabic and, by their measures, was virtually indistinguishable from native speakers of Egyptian Arabic. Since this learner's personality traits and the linguistic input she received matched those of another learner who was not as successful at achieving high-level proficiency, they concluded that her superior success was driven by an aptitude for languages. Some have argued that this high-level aptitude is distinct from the more traditional conceptualizations of language aptitude (e.g., Schneiderman & Desmarais, 1988), which typically distinguish rates of learning at lower levels of proficiency within language classroom contexts. Although sparse, the available evidence suggests some adults have an aptitude for attaining high-level proficiency.

The purpose of this study was to obtain empirical evidence of the ability of the High-Level Language Aptitude Battery (Hi-LAB; Doughty et al., 2010) to distinguish very successful language learners from other individuals. This article is organized as follows. We begin with a discussion of components of language aptitude at early and later stages of second language acquisition (SLA). We then present potential components of high-level aptitude and provide

a definition of high-level L2 proficiency. Finally, we present our motivation and the methods, analyses, and findings of this study, and conclude with a discussion of the results.

Background to the Study

Defining Language Aptitude

Some of the original theories of language aptitude posited the importance of specific cognitive abilities to language learning success (e.g., Carroll, 1985; Pimsleur, 1966). Current theorists have continued to incorporate cognitive abilities into the set of abilities that comprise language aptitude (e.g., Robinson, 2002, 2007; Sternberg, 2002). These cognitive abilities include domain-general abilities, such as logical reasoning (e.g., Pimsleur, 1966), inductive reasoning (Carroll, 1981), working memory (e.g., Mackey, Philp, Egi, Fujii, & Tatsumi, 2002; Miyake & Friedman, 1998; Robinson, 2002), and associative memory (Carroll, 1981; Schneiderman & Desmarais, 1988). However, they also include abilities specific to the verbal domain, such as auditory or phonemic coding ability (e.g., Carroll, 1981; Pimsleur, 1966; Schneiderman & Desmarais, 1988; Skehan, 2002) and grammatical sensitivity (Carroll, 1981; see also Skehan, 2002). Cognitive abilities are likely to be critical components of any valid theory of language aptitude, which is consistent with the recent shift to an overall cognitive view of SLA in aptitude research (e.g., Dörnyei & Skehan, 2003). Although personality and motivational factors likely play a role in higher-level learning, we take the view that it is primarily cognitive and perceptual abilities that constrain a learner's highest attainable proficiency level (Doughty et al., 2010; Mislevy et al., 2009; see also Carroll, 1995).

In recent years, particular cognitive control processes have been linked to specific language processing tasks. For example, better working memory has been linked to better L2 reading comprehension performance (e.g., Fontanini & Tomitch, 2009; Harrington & Sawyer, 1992). Similar results have been found in the speech production domain, with better inhibitory control supporting language control when switching between languages (e.g., Linck, Schwieter, & Sunderman, 2012).

Psycholinguistic studies have tended to focus on either bilinguals at an intermediate level or highly proficient simultaneous bilinguals who acquired both languages from a young age (e.g., Costa & Santesteban, 2004). Critically, there is a large body of literature indicating that children are more uniformly successful in learning languages (for a recent review of age effects, see Muñoz

& Singleton, 2011) and that children's learning processes are quite distinct from those engaged during adult nonnative language learning (e.g., DeKeyser, 2000; Kersten & Earles, 2001; Newport, 1990). The research suggests that aptitude is a critical factor for adult learners, whereas for children it is mainly an issue of early and continued experience with the language (DeKeyser, Alfi-Shabtay, & Ravid, 2010; Ross, Yoshinaga, & Sasaki, 2002). Moreover, as mentioned above, aptitude factors for high-level proficiency may differ from aptitude factors at earlier stages of learning. Therefore, the approach taken in the current study was to focus on cognitive and auditory perceptual abilities of adult learners as potential components of high-level aptitude.[1]

Defining High-Level Proficiency

Prior to the development of Hi-LAB, aptitude batteries were designed primarily to predict rate advantages in the first two years of SLA, often under intensive learning conditions. In contrast, one important aim of Hi-LAB is to predict language learning advantages at the most advanced stages of SLA, ideally at ultimate L2 attainment (see Doughty, 2013, for further discussion). In the SLA literature, ultimate L2 attainment refers to the stage in SLA which is the "end state" or in which learners experience extended stablilization. However, measuring ultimate attainment is difficult, in part, because it is unclear if there is any one point when a learner can be considered to have reached such an asymptote in learning (Birdsong, 2004; Long, 2003). Therefore, in this study, we used a related criterion of high-level attainment. Certainly, attaining high proficiency levels requires a significant amount of time, and learners vary substantially in their rate of learning. Some estimate that a typical adult learner requires a minimum of 10 years' immersion. Moreover, the operationalization of high-level proficiency can further complicate measurement of ultimate attainment in this context. For example, in many of the extant studies of ultimate attainment (most of which have focused on examining age of onset of acquisition effects on learning success; see Long, in press, for a recent review), a learner's attainment has been categorically identified as either native/nativelike or not, with little attention paid to finer distinctions at the upper proficiency levels. Indeed, the question of ultimate attainment is often not even explicitly assessed (e.g., DeKeyser, 2000), or nonnative speakers across a range of proficiency levels are treated as a single nonnative group (e.g., Abrahamsson & Hyltenstam, 2008; White & Genesee, 1996). However, see Granena and Long (2013) for a recent study of age effects on windows of opportunity for learning phonology, lexis and collocations, and morphosyntax.

In this study, conducted in the United States, we defined high-level attainment as highly proficient L2 performance as demonstrated on the Defense Language Proficiency Tests (DLPT; Defense Language Institute Foreign Language Center, 2009) and/or through high-level job performance in a single or multiple languages. We took this approach in order to have standardized measures of attainment for all participants, who are United States federal employees. This information was readily available from existing personnel database records and provided equivalent measures of attainment for all participants. Since we had limited testing time and participant L2s included a large number of languages, the administration of language proficiency outcome measures as part of this study was not feasible. Both the DLPT and the job requirements of the participants were rated using the Inter-agency Language Roundtable (ILR) scale for the sub-skills of listening and reading. The ILR scale subsumes a set of descriptions of six levels of proficiency oriented toward characterizing individuals' ability to function in a work environment while using an L2, where 0 = no proficiency, 1 = elementary proficiency, 2 = limited working proficiency, 3 = general professional proficiency, 4 = advanced professional proficiency, and 5 = functionally native proficiency (see Appendix S1 of the Supporting Information online, and visit www.govtilr.org for the history of the development of the ILR scale and details of the level descriptions). The DLPT is a standardized proficiency test with separate sections and scores that assess listening and reading comprehension and is the test of record for determining the foreign language incentive pay throughout the United States federal government and military. The DLPT is most often computer delivered and is typically an objective, multiple choice test, but for some less commonly taught languages, the DLPT employs constructed responses. Listening or reading passages are presented to test takers in the L2, and the comprehension question items are given in the first language (L1), which is English. Although listening and reading comprehension scores and job performance likely indicate different focal constructs, they should both be related to language proficiency more broadly, and we wanted to operationalize high-attainment to include higher proficiency levels that are exceptional but not yet nativelike. Demonstrated capacity to perform at ILR Level 4 (Advanced Professional Proficiency) on the job provides additional evidence of high-level attainment. We elaborate further on our proficiency criteria below (see *Criteria and Procedures for Assignment Into Groups* in Methods).

In our analyses, we examined the extent to which our measures could discriminate these high-attainment learners from a comparison group of individuals believed to represent a broader sample of language aptitude, but with

Table 1 Hi-LAB constructs and measures

Construct	Measure
Working Memory	
Executive Functioning	
Updating	Running Memory Span
Inhibitory Control	Antisaccade
	Stroop
Task Switching	Task Switching Numbers
Phonological Short-term Memory	Letter Span
	Non-Word Span
Associative Memory	Paired Associates
Long-term Memory Retrieval	ALTM Synonym
Implicit Learning	Serial Reaction Time
Processing Speed	Serial Reaction Time
Auditory Perceptual Acuity	Phonemic Discrimination: Hindi, English Pseudo-Contrastive
	Phonemic Categorization: Russian

a similar profile to the high-attainment learners with respect to other critical variables (e.g., intelligence, level of education, commitment to government or military service).

Potential Components of High-Level Aptitude

For this study, we define high-level aptitude as a composite of domain-general cognitive abilities and specific perceptual abilities that, together, can support or constrain one's ability to attain high-level proficiency as an adult learner. The domain-general cognitive abilities are potentially relevant to the development of high-level proficiency, broadly construed. The perceptual abilities—two aspects of auditory perceptual acuity—were hypothesized to be particularly relevant to the development of high-level spoken language abilities. We expected this set of cognitive and perceptual abilities to relate to listening and speaking abilities most strongly, and to also relate to reading abilities to a lesser extent.[2]

The potential components of high-level language aptitude measured in Hi-LAB and the tests used to measure them are listed in Table 1 (for a detailed discussion linking these constructs to high-level aptitude, see Mislevy et al., 2010). Each construct was measured by at least one test in the battery, and some constructs were measured using two tests. In addition, a measure of processing speed was derived from response times on one of the tests.

The Current Study

The purpose of this study was to obtain a first-round indication of the validity of Hi-LAB constructs by probing whether Hi-LAB could distinguish between highly successful and moderately successful language learners. Our battery included measures of constructs that have been identified in the literature, as well as other constructs that may be important given their potential contributions to the learning of specific linguistic features found to be problematic for adult learners (e.g., perception of phonetic features). We identified nine constructs as potential components of cognitive aptitude (see Table 1). Previous work has established the theoretical construct validity of these measures in the context of high-level language aptitude research (e.g., Doughty et al., 2010) and demonstrated sufficient reliability of these measures (for test-retest stability estimates, see Mislevy et al., 2010; internal consistency estimates computed for the present study are reported in Appendix S2 in the online Supporting Information).[3]

An extreme-groups design would involve a comparison between highly proficient learners and much less proficient learners (e.g., at or below ILR level 2) while controlling for as many non-aptitude variables as possible. That is, the comparison group would include individuals who have been unable to achieve an exceptionally high level of foreign language proficiency despite continued efforts, motivation to learn the language, and language use relevant to their jobs. While we were able to recruit a population of highly proficient learners, recruiting a matched group of lower proficiency learners who had clearly reached their maximum attainment was not possible for practical reasons; people who are not able to attain their target proficiency may be moved to different jobs that do not require language proficiency. Thus, we compared a high-attainment group to a comparison group of individuals that varied considerably in their proficiency level. This mixed-attainment group represents a sample from the same well-educated professional population to which the highly proficient learners belong.

The logic of this approach is as follows. Because achieving high-level proficiency is rare (e.g., Abrahamsson & Hyltenstam, 2008), we assume that the mixed-attainment group was likely more variable than the high-attainment learners with respect to language aptitude. The high-attainment group had realized their potential for high-level language learning and was, therefore, expected to have a relatively narrow distribution of language aptitude at higher levels. We expected individuals in the mixed-attainment group generally to subtend a lower and wider range of the aptitude spectrum. However, we did not

expect the high-attainment and mixed-attainment groups to be entirely distinct in terms of language aptitude. A few participants in the mixed-attainment group could have had sufficient aptitude for high-level attainment, but we expected this number to be very small due to the postulated low incidence of high-level aptitude. To the extent that the predictors examined in this study measure components of high-level language aptitude, we would also expect the high-attainment group to have better scores relative to the mixed-attainment group, who should have lower and more variable scores, and we would expect some overlap between the two groups.

Data analysis included three major steps: group assignment, participant matching, and discriminating between groups, all discussed in greater detail below. In brief, we first excluded nonnative English speakers and individuals with extensive exposure to a foreign language during childhood. Next, we classified each participant into either the high-attainment group or the mixed-attainment group based on their language attainment. Participants were then matched on an individual-by-individual basis using a propensity score matching procedure, which facilitates the creation of groups that differ in the critical variable of interest but are matched on a set of covariates (age, gender, and level of education). Performing the matching procedure minimized the chance that any group differences revealed in the group discrimination analysis were due to differences along those covariates rather than differences in aptitude. However, we note that groups could not be matched on amount of exposure to the L2.

After matching participants, the data were subjected to logistic regression analyses to determine how well the predictors could distinguish the two groups. As mentioned above, the constructs examined in this study included a range of domain-general cognitive abilities plus perceptual abilities that are oriented towards predicting success in listening and speaking skills. In order to assess whether this collection of constructs differentially predicts these skills vs. other skill domains, three sets of analyses were conducted:

1. *Listening high-attainment analysis*: high-attainment group defined by listening proficiency only;
2. *Reading high-attainment analysis*: high-attainment group defined by reading proficiency only;
3. *Either-skill high-attainment analysis*: high-attainment group defined by high attainment in either modality (reading and/or listening).

The group discrimination analyses allowed us to see how well Hi-LAB could distinguish between the groups.

Methods

Participants

Participants were personnel from various U.S. government agencies and members of the U.S. military. They were recruited by e-mail, web postings, and word of mouth, and were given a half day of administrative leave in order to participate in the study. In advance of the study, participants were told that they would help researchers examine whether a new language aptitude test called Hi-LAB predicted language-learning outcomes by taking the measures of cognitive and perceptual ability in the battery. The testing was conducted near the participants' work sites or in the University of Maryland's Center for Advanced Study of Language lab. All participants were volunteers and provided their informed consent following procedures approved by the university's Institutional Review Board. A total of 522 individuals participated (62% male, $M_{age} = 37$ years), but some participants did not qualify for inclusion in the analyses based on our exclusion and grouping criteria.

Participants' data were excluded if they did not meet the study inclusion criteria of being a native speaker of English and not having had extensive foreign language exposure prior to the age of 10 (i.e., frequent parental input or early immersion abroad), as indicated on a language history questionnaire (LHQ). Participant exclusion was coded by multiple researchers, with an inter-coder reliability of 97.1%. In all, the data from 46 of 522 participants (8.8%) were excluded from further analysis based on these criteria, leaving 476 participants to be possible candidates for the two groups. For some participants, we had information on their language testing and job assignment histories from their employer's personnel database. For the participants without personnel database entries, group membership was determined by their responses on the LHQ administered during testing. Details on the grouping criteria are provided following the description of the materials and procedure.

Materials and Procedure

The test materials consisted of the LHQ and 11 computer-delivered cognitive tasks. The test materials were administered to participants during a single session which lasted approximately three hours. Participants were informed of the nature of aptitude tests in general and, more specifically, of the challenging nature of this particular set of tasks. The order of the tasks was set so as to keep separate any tasks that measured similar constructs, such as perceptual acuity or inhibitory control.

Language History Questionnaire

The LHQ collected self-report demographic characteristics and information on participants' experiences with languages other than English. The LHQ encompassed the following factors: exposure to languages other than English as a child or teenager, foreign languages studied, foreign language use on the job, countries lived in, biographical data, and computer and video game use. The LHQ was administered in either paper-and-pencil or electronic format.

The Eleven Cognitive and Perceptual Tests

The cognitive and perceptual tests were created and administered using the E-Prime 2.0 suite of experiment software (Psychology Software Tools, 2011a). Each testing station consisted of a Dell D630 laptop computer with a 14.1 in. screen, a five-button Serial Response Box™ (Psychology Software Tools, 2011b), a set of laminated paper templates used as button labels for the response box, a mouse, and noise-reducing headphones.

Running Memory Span Test

This test (Bunting, Cowan, & Saults, 2006) measures the updating subcomponent of executive functioning (Miyake, Friedman, Emerson, Witzki, & Howerter, 2000). Participants listen to 20 lists of 12–20 auditorily-presented letters drawn from a set of 12 consonants, presented at a rate of 3 letters per second. At the end of the list, the participants must recall the last six letters in the list, in order, using a mouse-based interface. The score is the average number of letters correctly recalled in serial order per list. The maximum possible score is six and greater scores indicate higher levels of updating ability.

Antisaccade Test

This test was developed by Unsworth, Schrock, and Engle (2004) to measure the inhibitory control subcomponent of executive functioning (Miyake et al., 2000). Eye tracking results are not collected, thus this is technically an antisaccade analogue test. A visual cue is presented on the screen to indicate the target letter's location. Fifty ms after the offset of the cue, the target letter (B, P, or R) is displayed for 100 ms before the presentation of a backward mask. Participants must indicate the letter by pressing one of three buttons on the response box. In the two critical phases, the cues and letters appear on either the right or left side of the screen. In the prosaccade phase, the cue and letter appear on the same side; in the antisaccade phase, the cue and letter appear on opposite sides of the screen. Thus, in the antisaccade phase, the participant must inhibit the tendency to look toward the cue in order to see the letter. Scoring is based

on accuracy during the antisaccade phase, and higher scores indicate greater levels of inhibition. The data from 18 participants were excluded due to the apparent use of uncooperative strategies (15 due to looking at only one side of the screen, three due to slow responding and low accuracy).

Stroop Test

The Stroop test (Stroop, 1935) measures the inhibitory control subcomponent of executive functioning. Either words ("red," "green," or "blue") or solid rectangles are shown, one at a time, in the center of the screen. The words and rectangles appear in one of three colors (red, green, or blue) and participants must press one of three buttons on the response box to indicate which color was shown, ignoring the meaning of the word. Thus, sometimes the word and color are congruent (i.e., "blue" presented in blue letters) and sometimes they are incongruent (i.e., "green" presented in red letters) thereby requiring the participant to inhibit the prepotent response based on the meaning of the word. There are four test blocks, each with 48 items. The Stroop score is computed as the difference between incongruent and congruent log-transformed reaction times (RTs) from the test blocks. This score measures the degree of slowing on incongruent trials caused by the mismatch between the meaning of the word and the color of the text. There is no inherent maximum or minimum score. A lower Stroop score indicates better inhibition abilities. If participants exhibited poor performance during the final item blocks (average response accuracy less than 80%), their data for the Stroop test was not scored and was treated as missing in the data analyses. This criterion led to the exclusion of data for two participants.

Task Switching Numbers Test

This test measures the task switching subcomponent of executive functioning (Miyake et al., 2000). During several practice sessions, participants learn two distinct tasks in which they press one of two buttons on the response box as quickly and accurately as possible in response to single digits (1 through 9, excluding 5) presented on the screen. In the odd/even task, the digits are presented on a white background, and the participant must press one button in response to an odd digit, and a second button in response to an even digit. In the low/high task, the digits are presented against a gray background, and the participant must press one button in response to a digit less than 5, and the second button in response to a digit greater than 5. The digits 0 and 5 are never presented in either task; all other digits are presented an equal number of times. The tasks are initially practiced individually, but in the critical blocks, the tasks

alternate every three trials.Two participants' test data were excluded from the data analysis due to a problematic response strategy (average response accuracy less than 70%). Two scores were produced: (a) switch costs, indicating the extra time required to complete a task because it is different from the task on the previous trial, measured as switch trial RT–non-switch trial RT; and (b) mix costs, indicating the additional time required to perform a single task within the context of the mixed blocks, measured as non-switch trial (mixed block) RT–pure block RT. Lower switch costs and mix costs indicate higher levels of task-switching ability.

Letter Span Test

This test measures phonological short-term memory and was adapted from part of the operation span task developed by Unsworth, Heitz, Schrock, and Engle (2005). Lists of letters are presented on the screen, and participants must recall the letters in order after their presentation. The letters are presented one at a time for 900 ms each. There are three lists of each length from three to nine, for a total of 21 lists, presented in a pseudorandom order. The letters are drawn from a set of 12 consonants. Participants respond using the mouse to click on-screen buttons. The score is based on the total number of letters recalled in their correct positions.

Non-Word Span Test

This test also measures phonological short-term memory and was based off Gathercole, Pickering, Hall, and Peaker (2001). Fifteen lists of seven phonotactically plausible, one- or two- syllable non-words are presented, each word shown for two seconds. At the end of each list, participants are prompted with 14 non-words, half of which were in the list, and must indicate whether the word was on the most recent list by pressing one of two buttons on the button box. There are a total of 21 non-words, so each is reused several times over the course of the test, adding difficulty. The score is the number of correct answers, with a maximum possible score of 210. Higher scores indicate greater phonological short-term memory capacity.

Paired Associates Test

This adaption of Carroll and Sapon's (1959) paired associates test measures associative memory. Participants must learn 20 word pairs, each an English noun paired with a non-word, which is presented as a word in an unspecified foreign language. Each word pair is presented five times for five seconds each time during a learning phase. In the recall test, participants are prompted

with the purportedly foreign-looking words, one at a time, and must type the corresponding English word on the keyboard. The score is the number of correctly recalled English words. Certain spelling errors that did not affect the meaning of the target words were counted as correct responses.

Available Long-Term Memory Synonym Test

This test measures associative priming of long term memory (Was & Woltz, 2007). There are two tasks, a priming task and a comparison task, that are interleaved throughout the test. In the priming task, participants listen to a list of five words and are then shown two topic words, one of which is a synonym for two words in the list and one of which is a synonym for the other three words in the list. The participants indicate which word had more synonyms in the list with a button press. Following each list of the priming task is the comparison task, in which pairs of words are presented on the screen simultaneously, and the participant must indicate with a button press whether the words have similar or different meanings. There are four unscored warm-up pairs in each set, followed by eight scored pairs. Nine of the 18 sets are primed, meaning that one or both words in each comparison pair are synonyms of one of the two topic words from the preceding priming task. The other nine sets are unprimed, meaning that none of the words in the comparison pairs are synonyms of the topic words from the preceding list from the priming task.

For scoring, the response time and accuracy for each comparison are combined into a rate score by dividing the number of correct responses within a set by the total amount of time taken for responses within that set, resulting in a score indicating correct responses per minute for that set. This rate score is computed separately for the nine primed sets and nine unprimed sets. The rate score for each primed set is regressed on the corresponding unprimed set, producing a residual priming score that removes the variance from primed sets which can be accounted for by processes that also occur in unprimed sets. *Same* comparisons are considered separately from *different* comparisons within each set, and these 18 residual difference scores are summed to create the final score. There is no maximum or minimum, but a score of 0 indicates an average amount of priming, a positive score indicates more priming, and a negative score indicates less priming. If participants exhibited poor performance during the synonym comparisons (average response accuracy less than 80%), their data was not scored and was treated as missing in the data analyses. The scores of two participants were treated as missing in this way.

Serial Reaction Time Test

This test measures sequence learning and was adapted from Willingham, Nissen, and Bullemer (1989). Four horizontally arranged boxes are shown on the screen, indicating the four positions in which an asterisk will appear. On each trial, an asterisk appears in one of the four boxes, and the participant must press the corresponding button on the response box. After a 500 ms intertrial interval, an asterisk appears in a different location from the previous trial. There were six blocks of 96 trials. In the first and sixth (final) block, the asterisks appeared in a pseudorandom order. In blocks two, three, four, and five, the asterisks appeared in a repeating pattern of length 12. Two scores were generated based on test performance. First, a reactive facilitation score was computed as the difference in median RTs in the final sequential block and the final random block, intended to indicate implicit learning. Higher scores indicate better sequence learning. Second, a general processing speed measure was computed as the mean RT in the first random block. Lower scores indicate faster processing speed. Poor performance in the fifth and sixth blocks (average response accuracy less than 70%) led to the exclusion of the Serial Reaction Time test data, which were then treated as missing in the analyses. This criterion led to the exclusion of only one person's test data.

Phonemic Discrimination: Hindi, English Pseudo-Contrastive Test

This test was developed by the authors to measure perceptual acuity for non-native speech sounds. Participants are presented with two auditory stimuli in sequence, and they are required to indicate with a button press whether the two stimuli are the same sound or two distinct sounds. This test measures an individual's ability to hear the contrast between Hindi voiced /ǰ/ and voiceless /č/. These sounds correspond roughly with the initial sounds in the English words "jeep" and "cheap." However, the stimuli used have voice-onset time (VOT) values ranging from −120 to 0 ms, which are normally in the range for just the English /ǰ/ phoneme ("j" sound). English speakers are expected to have a difficult time distinguishing these sounds, as they all fall under a single phonemic category in English. Performance on this test is scored using *d*', a standardized difference score based on signal detection theory that assesses one's ability to discriminate between two stimuli along a continuum. A separate *d*' score is computed for all five non-identical sound pairs (e.g., sound-1 and sound-2) involving both endpoints of the continuum; the mean of those 10 *d*' scores is the final score for this test. A higher *d*' score indicates better perceptual acuity.

Phonemic Categorization: Russian Test

This test was also developed by the authors to measure perceptual acuity for nonnative speech sounds. Participants listen to nine sounds, ten times each, for a total of 90 trials, presented in a pseudorandom order. The sounds would all be considered the same phoneme in English (i.e., the voiced alveolar stop in the syllable /da/), but because the VOT is manipulated, they span two different phonemes in Russian (the prevoiced alveolar stop in the syllable /da/ and the voiceless, unaspirated alveolar stop in the syllable /ta/). The sounds are assigned to three categories, based on this VOT manipulation. Participants listen to each sound, and indicate which of the three categories it belongs to with a button press. Feedback is shown after each response, indicating the correct category for the preceding sound. The score is the number of correctly categorized sounds, with a maximum possible score of 90. Higher scores indicate greater perceptual acuity.

Criteria and Procedures for Assignment Into Groups

Participants assigned to the high-attainment group had demonstrated high-level proficiency. Our criteria allowed participants to qualify for the high-attainment group in three ways: (1) testing at or above an ILR level 4 on the DLPT in any language, (2) working two or more job assignments which were characterized at a difficulty level of ILR level 4 or higher in any language, or (3) demonstrating competent multilingualism by testing at or above ILR level 3 on the DLPT in two or more languages. Meeting at least one of these criteria qualified a participant for the high attainment group.

To prepare the data for the three main analyses (discriminating on listening high-attainment, reading high-attainment, and either-skill high-attainment), the high-attainment group criteria were applied in three rounds—first to just listening test scores and job assignments, next just reading test scores and job assignments, and finally to both skill modalities. For the "either-skill" grouping criterion, a test score in any modality could qualify a participant for the first criterion, a job assignment in any modality could qualify for the second criterion, and test scores in any combination of modalities could qualify for the third criterion (e.g., a combination of testing at ILR level 3 in Russian reading and ILR level 3 in Arabic listening). Participants selected for these three analysis represented different, but overlapping, subsets of the participants.

Participants assigned to the mixed-attainment group had extensive language training experience, but failed to meet the criteria for the high-attainment group. Specifically, participants were considered to have extensive experience if one or more of the following criteria were met: (1) They rated their ability in a foreign

language as "good," "very good," or "excellent"; (2) they reported earning test scores of ILR level 2 or higher; (3) they reported ever working as a language analyst in any language; (4) they reported working on a job assignment which required at least a level 2 language ability and they reported being able to perform all of their duties on that job; or (5) they reported extensive foreign language study experience (based on where they studied foreign language, how long they studied foreign language, and how much time they spent living abroad in a country where a language other than English was spoken). Participants were also considered to have extensive language training experience if they had taken more than three semesters of foreign language courses in college (e.g., majoring in a foreign language), studied language in the military at the Defense Language Institute Foreign Language Center, or lived abroad in a non-English speaking country for more than six months as an adult. High school foreign language classes and two or three semesters of foreign language study in college were not considered sufficient experience to qualify an individual for the mixed-attainment group. The mixed-attainment grouping criteria were also applied in three separate rounds for the reading, listening, and either-skill analyses.

It is important to note that, because participants could qualify for the high-attainment group in various ways, a given individual could be classified as having high attainment by one criterion (e.g., for the listening analysis) but then be classified as part of the mixed-attainment group by a different criterion (e.g., for the reading analysis).

Analyses

Grouping Criteria

The high-attainment grouping criteria yielded 81 (listening), 98 (reading), and 113 (either-skill) participants. The mixed-attainment grouping criteria yielded 303 (listening), 286 (reading), and 271 (either-skill) participants. These groups were then subjected to the propensity score matching procedure.

Propensity Score Matching

We used propensity score matching to obtain high-attainment and mixed-attainment groups (for each analysis) that were balanced with respect to three covariates: age, gender, and level of education. Creating matched samples is a way of reducing differences between the groups due to covariates, or due to an unobserved factor for which the covariate can act as a proxy. These three covariates were chosen because there is clear potential for all three to be related to L2 attainment and/or performance on the Hi-LAB tests. Age is likely

to be related more or less directly to exposure to and opportunities to learn foreign languages. Age is also known to be correlated with overall response times (e.g., Fozard, Vercruyssen, Reynolds, Hancock, & Quilter, 1994), which form the basis of test scores for some Hi-LAB measures. Gender has been identified as a potential factor influencing language learning, with documented gender differences in learning strategies (Oxford, 1993) and suggestive evidence that females outperform males in L2 listening skills (Larsen-Freeman & Long, 1991). Individuals with higher levels of education may also have a greater exposure to L2 learning opportunities or may be more likely to self-select into careers in foreign affairs that require language training (Ehrman & Oxford, 1995).

Another important factor for attaining high-level proficiency is the total amount of language training and exposure. Some self-report data was collected, but in examining this data, we found a number of problematic issues in interpreting it. In some cases, participants failed to report the years of foreign language study, they indicated a number but not the unit of measurement (e.g., months, semesters, or years), and, most importantly, the length of study may have been interpreted as targeting the length of formal language training and may not have fully reflected informal language training, study or experience. The types and quality of training are very heterogenous in this sample, which means that a report of *N* months of training could describe vastly different experiences for two different people (e.g., an immersion experience vs. low-intensity self-study).

Additionally, due to the cross-sectional design of the study, it is impossible to draw firm conclusions about potential effects of exposure. On the one hand, it may be that greater exposure leads directly to higher proficiency. On the other hand, individuals who failed to reach high-attainment levels of proficiency may have opted against pursuing additional advanced language training, since there may be little incentive to improve beyond the professional-level proficiency they already attained. Ultimately, we decided that including the self-report exposure data as either a covariate in the propensity score matching procedure or as a predictor in the classification analyses would be likely to add more noise than useful information. The relative impact of training time can only be adequately addressed by a longitudinal design. It is also worth noting that all participants had considerable L2 experience (see grouping criteria above), so the absence of exposure time in the present study is unlikely to introduce substantial limitations to this design.

Participants in the two groups were matched according to their propensity scores, which are the conditional probabilities of belonging to one group

given the observed covariate values (Rosenbaum & Rubin, 1983). This allows participants with very similar covariate profiles to be paired without having to match exactly on each covariate, and it creates groups with conditional covariate distributions that are independent of group membership.

Propensity scores were computed by estimating group membership probabilities with logistic regression models with age (interval), gender (categorical), and level of education (ordinal) as predictors. Matching was done with a nearest-neighbor matching approach, such that each participant in the smaller high-attainment group was matched with the participant from the larger mixed-attainment group with the closest propensity score. Matching was performed without replacement, so a participant in the high-attainment group could be matched to no more than one participant in the mixed-attainment group. If a suitable match could not be found for a given participant from the high-attainment group, then that participant was not included in the analysis dataset. The absolute difference between propensity scores for a pair of matched participants was restricted to be no greater than .075. Larger values produced worse matching without substantial increases in the size of the analysis dataset, whereas smaller values reduced the sample size while only slightly increasing closeness of covariate matching.

The propensity score matching procedure was carried out for each of the three definitions of high attainment: listening, reading, and either-skill. Because the initial samples and matching parameters were somewhat different for each definition, the resulting datasets consisted of different subsets of participants for each analysis. Matching produced 76, 94, and 103 matched pairs for the listening, reading, and either-skill matching procedures, respectively.

We assessed the effectiveness of the propensity score matching procedure by comparing each covariate's within-groups distributions before and after matching and by testing for significant group differences in the distributions of the three covariates in the unmatched and matched samples. Table 2 shows that any significant group differences in age and education in the unmatched datasets were no longer significant in the matched datasets. These results suggest that the propensity score matching procedure successfully produced samples with balanced covariate distributions.

Data Preparation: Outliers and Missing Data

Extreme outliers in the data, defined as scores more than three times the interquartile range beyond the first and third quartiles of the distribution of the score, were removed to prevent them from having undue influence over the results of any analyses. Only four scores were identified as outliers and were treated

Table 2 Group differences on covariates before and after matching

			Before Matching			After Matching		
Skill Analysis	Covariate	Test	Statistic	*df*	*p*	Statistic	*df*	*p*
Listening	Gender	χ^2	0.699	1	.403	0.738	1	.390
	Education	χ^2	**12.180**	5	.032	0.962	4	.916
	Age	*t*	**−4.620**	117.383	< .001	−0.721	149.850	.472
Reading	Gender	χ^2	0.001	1	.980	0.196	1	.658
	Education	χ^2	7.886	5	.163	2.003	5	.849
	Age	*t*	**−5.525**	155.543	< .001	0.269	185.099	.788
Either-skill	Gender	χ^2	0.416	1	.519	0.185	1	.667
	Education	χ^2	8.225	5	.144	1.598	5	.901
	Age	*t*	**−5.768**	197.093	< .001	0.226	202.296	.821

Note. Significant test statistics (in bold) indicate lower education levels and younger age for the mixed-attainment group before matching.

as missing: three extremely low Paired Associates scores and one extremely low Running Memory Span score, which likely indicated technical problems or lack of motivation on those particular tests. No scores were so high as to be considered outliers.

Covariates and scores on the 13 Hi-LAB measures could be missing for a variety of reasons, including computer errors during testing or score rejections during the scoring process. Casewise deletion of participants with missing data is known to create bias, as are other methods such as mean replacement (Little & Rubin, 2002). Thus, we used a multiple imputation approach, such that multiple complete datasets were created and analyzed, with plausible but distinct values imputed for each missing data point in each complete dataset (Rubin, 1987). When using a multiple imputation approach, the results of the analyses (e.g., regression coefficients) are pooled across imputed datasets, taking both within- and between-imputation variability into account. Parameters and their variance estimates are adjusted to reflect the uncertainty about the parameter estimate due to missing values. The ability to incorporate uncertainty due to the imputation procedure itself in the final parameter estimates is one of the primary advantages of multiple imputation over other single-imputation methods such as hot-decking.

To ensure that the majority of any given participants' data had been observed, participants could have missing values on no more than three variables

to be included in the multiple imputation process and following analyses. The median number of missing values for a given variable was four (range = 1 to 18). The imputation model should include all variables to be entered into the analysis model, or else one risks biasing the analysis results towards zero (van Buuren, Boshuizen, & Knook, 1999). Thus, all predictors, covariates (age, gender, and education), and group membership variables for the group discrimination analysis were included in the imputation models. For each imputed dataset, we ran 10 iterations, which an examination of covariance matrices showed was sufficient to reach stable multivariate relationships across iterations. The result of the multiple imputation procedure was 10 complete datasets, which were then analyzed separately and the results were pooled.

Multiple imputation was performed in the R statistical software environment (R development Core Team, 2009) using the *mice* package (Multivariate Imputation by Chained Equations; van Buuren & Groothuis-Oudshoorn, 2011).

Results

We used logistic regression to measure the utility of the 13 cognitive and perceptual measures for discriminating between the two groups. A separate analysis was carried out for three skill attainment measures—listening, reading, and either-skill. In each analysis, an indicator variable was set equal to one for each member of the high-attainment group and zero for each (matched) member of the mixed-attainment group (i.e., the models predict high attainment). We then fit a logistic regression model with the group indicator as the dependent variable and the set of Hi-LAB scores as the independent variables.

Prior to fitting the logistic regression model to each set of matched data, the Hi-LAB scores were standardized (centered and scaled by subtracting the mean and dividing by the standard deviation across the matched groups), which enables direct comparison of the estimated parameters. The magnitudes and signs[4] of the fitted β parameters indicate which Hi-LAB scores provided the most information with respect to group membership and how changes in the scores related to the probability of being in each group. For each skill definition we present parameter estimates for two sets of analyses. First, we present parameter estimates for single-predictor models, each of which consists of an intercept and a single Hi-LAB score. This indicates how each predictor is related to group membership in isolation, thereby avoiding interpretive difficulties introduced by correlations between the predictors (the correlation matrix for the 11 subtests of the Hi-LAB Test are reported in Appendix S3 of the

Table 3 Betas (with *SE*s) for the single-predictor analyses

	Analysis		
Predictor	Listening[a]	Reading[b]	Either-skill[c]
Running Memory Span	0.176 (0.165)	0.201 (0.149)	†**0.269** (0.143)
Antisaccade	− 0.057 (0.163)	− 0.055 (0.146)	− 0.006 (0.140)
Stroop	− 0.073 (0.163)	− 0.015 (0.146)	− 0.096 (0.140)
Task Switching Mix Cost	0.031 (0.163)	0.121 (0.147)	0.055 (0.140)
Task Switching Switch Cost	† **− 0.296** (0.168)	− 0.227 (0.149)	− 0.155 (0.141)
Letter Span	***0.422** (0.173)	***0.356** (0.152)	****0.458** (0.149)
Non-Word Span	†**0.289** (0.168)	†**0.256** (0.150)	***0.327** (0.145)
Paired Associates	****0.679** (0.210)	****0.437** (0.163)	****0.563** (0.167)
ALTM Synonym	0.072 (0.163)	0.012 (0.146)	0.099 (0.140)
Serial Reaction Time	***0.390** (0.176)	†**0.261** (0.151)	***0.348** (0.149)
Processing Speed	0.067 (0.163)	− 0.008 (0.146)	0.034 (0.140)
Phonemic Discrimination	0.164 (0.165)	− 0.008 (0.146)	− 0.054 (0.140)
Phonemic Categorization	0.119 (0.164)	0.145 (0.148)	0.184 (0.142)

Note. Bold indicates significant predictors. [a]$N = 152$. [b]$N = 188$. [c]$N = 206$.
†$p < .10$.*$p < .05$. **$p < .01$.

Supporting Information online). Second, we present parameter estimates for models containing all 13 Hi-LAB scores. Finally, overall model performance is summarized by classification accuracy and patterns of correct and incorrect classification of high-attainment and mixed-attainment group members.

Note that the fitted parameters of a logistic regression model can be difficult to interpret directly. Gelman and Hill (2007, p. 82) suggest that dividing fitted logistic regression parameters by 4 provides an estimate of the maximum possible change in predicted probability per unit change in the associated independent variable. So, for example, a fitted β parameter value of 1.00 would correspond to a maximum possible change in predicted probability of .25 (e.g., a positive one standard deviation change in a predictor could correspond to a change from .50 probability of high-attainment to .75 probability of high-attainment). Because the data were standardized prior to analysis, each fitted parameter can be interpreted with respect to a change of one standard deviation in the associated Hi-LAB score.

Single Predictor Models

Table 3 shows the fitted parameter estimates for the single predictor models for the listening, reading, and either-skill analyses. The single-predictor model

Table 4 Betas (with *SE*s) for the full-model analyses

	Outcome		
Predictor	Listening[a]	Reading[b]	Either-skill[c]
Intercept	− 0.026 (0.182)	− 0.004 (0.155)	− 0.007 (0.152)
Running Memory Span	− 0.216 (0.246)	− 0.078 (0.204)	− 0.063 (0.202)
Antisaccade	− 0.324 (0.228)	− 0.265 (0.201)	− 0.238 (0.188)
Stroop	− 0.136 (0.188)	− 0.084 (0.163)	− 0.217 (0.159)
Task Switching Mix Cost	− 0.212 (0.220)	− 0.045 (0.186)	− 0.147 (0.184)
Task Switching Switch Cost	†**− 0.404** (0.215)	− 0.279 (0.183)	− 0.235 (0.180)
Letter Span	†**0.483** (0.262)	***0.447** (0.224)	***0.477** (0.219)
Non-Word Span	− 0.045 (0.226)	0.030 (0.199)	0.054 (0.194)
Paired Associates	****0.675** (0.244)	***0.403** (0.179)	****0.503** (0.183)
ALTM Synonym	0.118 (0.205)	− 0.099 (0.178)	0.038 (0.173)
Serial Reaction Time	***0.508** (0.206)	***0.402** (0.172)	****0.455** (0.170)
Processing Speed	0.051 (0.206)	− 0.047 (0.194)	− 0.067 (0.181)
Phonemic Discrimination	0.160 (0.206)	− 0.082 (0.174)	− 0.151 (0.164)
Phonemic Categorization	− 0.073 (0.212)	0.069 (0.181)	0.111 (0.173)

Note. Bold indicates significant predictors. [a]$N = 152$. [b]$N = 188$. [c]$N = 206$.
†$p < .10$. *$p < .05$. **$p < .01$.

fits indicate that measures of phonological short-term memory (Letter Span and Non-Word Span), implicit learning (Serial Reaction Time), and associative memory (Paired Associates) were robust predictors of high attainment when considered in isolation. Executive functions were also marginally predictive (Switch cost in the listening analysis; Running Memory Span in the either-skill analysis).

Full Models

The single predictor models provide a measure of the relationship between each individual Hi-LAB score and each attainment indicator, independent of the other predictors. Table 4 shows the fitted parameters for the full-model (i.e., multiple regression) analyses, which indicate the relationships between the Hi-LAB scores and the attainment indicators in the context of the full aptitude battery. The full-model results paralleled the single-predictor results, with contributions from measures of phonological short-term memory (Letter Span, but not Non-Word Span), implicit learning (Serial Reaction Time), and associative memory (Paired Associates). In addition, executive functions, as measured by Task Switching Switch Costs, were marginally predictive in the

Table 5 Classification accuracy with statistical significance of associated likelihood ratio test, and pseudo-R^2

	Outcome		
	Listening	Reading	Either-Skill
Classification accuracy	70.4%	59.2%	67.2%
Pseudo-R^2	.176**	.109	.146**

Note. Classification accuracy was computed as the percentage of participants correctly classified as members of the high-attainment group or the comparison group. Analyses were conducted using a classification criterion of 0.5 (see text for details). Pseudo-R^2 computed using the Nagelkerke method.
**$p < .01$

listening analysis, although the coefficient was negative in both the single-predictor and full-model analyses. We return to this point in the discussion.

Classification Accuracy

We can summarize the overall performance of the full models by considering their ability to classify cases correctly, and we can gain further insight into model performance by analyzing both types of misclassifications (high-attainment cases misclassified as mixed-attainment, and mixed-attainment cases misclassified as high-attainment). Table 5 shows overall classification accuracy rates and the statistical significance of associated likelihood ratio tests for the three skill-attainment group indicators. For each skill attainment indicator, the likelihood ratio tests compared an intercept-only model to the full model that includes all 13 Hi-LAB scores.

The fitted model produces a more or less continuous range of predicted high attainment probabilities. In order to classify high-attainment and mixed-attainment observations, we must select a criterion on the predicted probability scale. Any observation with a predicted probability above a given criterion is classified as high-attainment, and any observation with predicted probability below the criterion is classified as mixed-attainment. Hence, there are two types of possible correct classification and two types of possible misclassification: high-attainment observations correctly classified as high-attainment (i.e., hits), mixed-attainment observations correctly classified as mixed-attainment (i.e., correct rejections), high-attainment observations incorrectly classified as mixed-attainment (i.e., misses), and mixed-attainment observations classified as high-attainment (i.e., false alarms).

If the cost of a miss is equal to the cost of a false alarm, and if the benefit of a hit is equal to the benefit of a correct rejection, optimal classification for a given model is based solely on predicted class membership probabilities. For a given model, the optimal classification rule classifies as high-attainment any observation that is predicted to be more likely to be high-attainment (i.e., if $\hat{p}_i > .5$) and classifies as mixed-attainment any observation that is predicted to be more likely to be mixed-attainment (i.e., if $\hat{p}_i \leq .5$). In logistic regression, predicted class membership probabilities are a function of a linear combination of measured predictors, as described above.

If the cost of a miss is not equal to the cost of a false alarm, and/or if the benefit of a hit is not equal to the benefit of a correct rejection, then a classification criterion may be chosen to reflect the relative costs and benefits of incorrect and correct classifications. For all analyses reported here, we employed a classification criterion of .50.

Table 5 shows the classification accuracy of the full models. Classification accuracy was highest for the listening attainment indicator and lowest with the reading attainment indicator. When interpreting the classification accuracies reported in Table 5, it is important to keep in mind that the mixed-attainment group consisted of educated professionals, whom we would expect to score above average (with respect to the general population) on any of a number of cognitive measures. Given how rare high-level foreign language proficiency is, and given the characteristics of the two comparison groups, the classification rates given in Table 5 are very likely underestimates of the ability of Hi-LAB to correctly classify people who are capable of high attainment in a foreign language relative to the general population. Additionally, as a consequence of the cross-sectional design of the study, all participants have undergone some amount of foreign language training—some with extensive experience.

As noted above, overall accuracy consists of two components: correct classification of high-attainment learners as high attainers and correct classification of mixed-attainment learners as not having achieved high attainment. Table 6 provides a detailed breakdown of the classification performance of the full models for each attainment indicator. For each combination of attainment indicator and comparison group, correct classifications are shown in shaded cells and errors are shown in unshaded cells. Note that the classification rates in Table 6 are conditioned on the observed class. Hence, within each row, the numbers represent the percentages within each observed class. So, for example, for the listening analysis (left portion of the table), 68.7% of observed mixed-attainment cases were classified correctly, while 31.4% were incorrectly classified as high attainers; for the observed high-attainment cases, 72.1% were

Table 6 Classification performance for each outcome, computed as the percentage of correct and incorrect classifications of mixed-attainment and high-attainment group members

	Outcome					
	Listening		Reading		Either-skill	
Observed	Predicted Mixed	Predicted High	Predicted Mixed	Predicted High	Predicted Mixed	Predicted High
Mixed	68.7	31.4	59.6	40.4	65.6	34.4
High	27.9	72.1	41.2	58.8	31.3	68.7

Note. Classification percentages for the mixed-attainment observations are given in the top row, and analogous percentages for the high-attainment observations are given in the bottom row. Percentages for the listening, reading, and either-skill indicators are given in the leftmost, middle, and rightmost columns, respectively. Within each pair of columns corresponding to a given skill attainment indicator, percentages of observations predicted to be non-high are given in the left column, and percentages of observations predicted to be high are given in the right column.

classified correctly, while 27.9% were incorrectly classified as non-high attainers. Because the classification percentages in Table 6 are conditioned on observed group membership, the overall accuracy percentages given in Table 5 are the averages of the two types of correct classifications from the corresponding attainment indicators in Table 6. Overall, for each attainment indicator, the models correctly classified roughly equal percentages of the two groups.

Qualitative analysis of the LHQ data indicated that a fairly large number of the misses (i.e., high-attainment learners classified as not being high attainers) learned foreign languages through non-standard methods (e.g., missionary work in foreign countries) and had somewhat lower education levels, whereas a number of the false alarms (i.e., mixed-attainment learners classified as high attainers) had unusually high levels of education. If this pattern is replicated in future studies, this might suggest two issues worth consideration. First, formal education may upwardly bias composite aptitude scores based on these measures, which may need to be accounted for in any real-world applications of aptitude tests that measure these constructs. Second, particular features of extensive, focused language training in an in-country immersion learning environment may support attainment of higher proficiency levels than expected based on one's aptitude alone. Future research on immersion learning could focus on measuring specific features of different immersion learning environments

and linking them to language learning outcomes to better understand the interactions between the language learning context and high-level language aptitude.

Summary of Main Results

Logistic regression models were used to predict high attainment in listening, reading, and either skill, and the results indicated that the information available from Hi-LAB test scores enabled classification of members of the high-attainment and mixed-attainment groups. Across skills, classification accuracy ranged from 58.8% to 72.1%. Consideration of the fitted model parameters indicated that Paired Associates, Serial Reaction Time, and Letter Span provided substantial classification information across the listening, reading, and either-skill analyses, consistent with the idea that associative memory, implicit learning, and phonological short-term memory play an important role in achieving high attainment. In the listening analysis, Switch Cost also provided useful information, though the relationship between this score and high attainment was opposite of the expected direction.

Discussion

The purpose of this study was to examine whether the set of cognitive and perceptual abilities measured by the Hi-LAB test battery can distinguish very successful language learners from other individuals. Results from a series of analyses indicate that the tests correctly classified high-attainment learners with up to 70% classification accuracy. The classification accuracies for the listening and either-skill outcomes were statistically significantly well above chance. That is, high-attainment learners could be reliably classified based on their performance on these measures of cognitive and perceptual abilities.

These results are all the more striking when we recall that, due to recruitment constraints, our analyses were based on a comparison group (i.e., the mixed-attainment group) comprised of language learners with varying levels of proficiency, including relatively high proficiency—up to ILR level 3+. A more ideal comparison group for this type of analysis would be a group of language learners who have had ample learning opportunities but who are more distinct from the high-attainment group in their proficiency—such as individuals unable to surpass ILR level 2 despite continued effort, opportunity, and motivation. The fact that these measures were able to distinguish high attainment from mixed attainment is all the more remarkable, and suggests that Hi-LAB may be even better at distinguishing high-level learners from more typical adult language learners. That is, the classification accuracies from this

study may underestimate the true ability of Hi-LAB to predict attainment of high-level proficiency. This must be examined in a longitudinal study in which performance on these measures is assessed prior to participants achieving high attainment.

Another point to consider is that our indicators of proficiency level are less than perfect, which may be further suppressing the true validity of the test battery. The proficiency measures are based on scores that come from different languages and different DLPT test versions, and the ILR scale is a general proficiency rather than fine-grained measure. Furthermore, while performance on standard proficiency tests is an important goal for this population, it is not the only goal, nor is predicting proficiency test scores the only possible application for Hi-LAB. For example, Hi-LAB test scores can be leveraged to generate learner aptitude profiles, which could enhance language training through specific training interventions that leverage learner strengths (e.g., by identifying optimal pedagogical approaches; see Brooks, Kempe, & Sionov, 2006; Perrachione, Lee, Ha, & Wong, 2011; Vatz, Tare, Jackson, & Doughty, 2013) and overcome learner weaknesses (e.g., through improvement of specific abilities; see Brehmer, Westerberg, & Bäckman, 2012). Additional measures that are related to job performance, real-world task performance, or more discrete measures of linguistic ability (all of which may differ substantially from the tasks that comprise general proficiency exams) would provide complementary evidence for the validity of Hi-LAB constructs, and may show a different pattern of which individual predictors appear most important.

The Hi-LAB project constitutes a major advance in foreign language aptitude research. Prior to this study, we developed the first theoretical model of high-level language aptitude and established construct validity for the model (Mislevy et al., 2009). This study provides the first empirical evidence of criterion validity. The finding that Hi-LAB successfully distinguished high-attainment learners from their colleagues lends support to our model; nonetheless, further research is needed to gather additional evidence of the utility of Hi-LAB and to refine the theoretical model. Moreover, due to the cross-sectional nature of the research design, it is impossible to rule out the possibility that some of the cognitive and perceptual abilities measured by Hi-LAB are enhanced through the very process of attaining high-level proficiency. As pointed out by an anonymous reviewer, if results reveal a certain ability in all high-level attainers but also in a few non-high-level attainers, then it might be concluded that this is not a consequence of high-level attainment, but rather an innate ability that some—but not all—learners have taken advantage of. In cases where an ability is found only among high-level attainers, it is not clear whether the

ability was present at the outset or developed with learning. A longitudinal study that tracks language proficiency progress as well as any potential changes in aptitude is needed to directly address this concern.

Component Abilities of High-Level Aptitude

Looking across analyses, the pattern of results identifies a clear group of constructs that contributed to successful classification of participants with a high degree of consistency when examining listening, reading, and either-skill outcomes. Specifically, associative memory, implicit learning, and phonological short-term memory all positively distinguished the high-attainment language learners, with better performance indicating a greater likelihood of high-level attainment. Clearly, implicit and explicit learning mechanisms along with memory storage and retrieval processes play a role in achieving high-level proficiency.

The executive functions measures did not show the predicted positive relationship with high-attainment outcomes. In fact, the Switch Cost component of Task Switching was negatively predictive in the listening analysis, indicating that individuals with greater flexibility in mental shifting (i.e., smaller switch costs) were less likely to be high attainers in listening. Being capable of focusing one's attention squarely on the L2—preventing both controlled and uncontrolled attentional shifts to the L1—might be important to developing high-level L2 proficiency. Perhaps being able to easily switch back to the L1 impedes reaching very high levels of L2 proficiency because one can rely on the L1 more readily, thereby preventing deeper L2 processing. This seems particularly relevant to L2 listening, where the transient auditory input is available for processing for only a brief moment. This contrasts with reading, where the visual input typically is present longer and shifts in attention can be overcome by rereading previous portions in the text. This hypothesis might also explain why the listening analysis results were strongest in terms of model classification. The (negative) predictive contribution of switch costs appears to have provided skill-specific enhancements to the listening analysis. Although this hypothesis is purely speculative at this point, if future studies replicate this negative relationship, then this suggests specific and theoretically interesting constraints on the contributions of executive functions to high-level language attainment.

Differential Skill Prediction

Many of the constructs examined in this study, such as executive functions, are likely to be relevant to multiple aspects of language learning. However,

some of the examined constructs (and some of the specific measures of these constructs) were selected specifically for purposes of enhancing prediction of spoken language skills. For example, we included two auditory perceptual acuity measures, and some of the executive function measures (e.g., Running Memory Span) used auditory stimuli. Thus, Hi-LAB was expected to show discriminant validity, with stronger relations to listening outcomes than reading outcomes. Indeed, Hi-LAB components were related to both listening and reading proficiency, but a stronger relationship was found when focusing on listening attainment. As discussed above, executive functions (task set switching) seem to have contributed specifically to the listening analysis. Note also that implicit learning and associative memory contributed to all three analyses but had the strongest coefficients in the listening analysis, providing some support to our goal of optimizing prediction to listening and speaking outcomes. Additional research is needed to test the hypothesis that the constructs examined here should also predict high-level attainment in speaking, and to examine other constructs that may enhance prediction of reading or speaking specifically. That is, it is feasible that a slightly different set of measures could be combined to enhance discriminant validity for reading skills. For example, constructs related to visual perceptual acuity may be relevant to the development of skilled reading in languages with non-Roman scripts and, therefore, may add incremental validity in the prediction of high-level reading proficiency attainment. To enhance the differential prediction of speaking proficiency, constructs such as speech planning (fluency) could be incorporated. Considering a range of constructs—some of which are skill-specific and some of which are skill-non-specific—and identifying the constructs that contribute to differential skill prediction will inform the development of more sophisticated theories of language aptitude and strengthen the state of the science of aptitude measurement.

High-Level Attainment vs. Initial Stages of Learning

Prior to the conceptualization of Hi-LAB, language aptitude tests were designed primarily to predict rate of language learning at initial stages (i.e., first two years) under intensive, instructed SLA conditions. The three most widely used tests—the Modern Language Aptitude Test (MLAT), Defense Language Aptitude Battery (DLAB), and Pimsleur Language Aptitude Battery (PLAB)—have four constructs in common: phonetic coding ability, grammatical sensititivity, rote learning ability, and inductive language learning ability. The PLAB and the MLAT also include measures of L1 vocabulary knowledge as a proxy for verbal ability, and the PLAB includes a measure of phonemic discrimination.

These tests have been shown to predict outcomes such as grades in courses and scores on general proficiency tests (e.g., DLPT, particularly listening and reading) after instruction in high schools, universities and at the Defense Language Institute Foreign Language Center. In contrast, Hi-LAB was designed to predict the attainment of high-level proficiency—rather than initial rate of learning—with the expectation that high-level acquisition requires going beyond the classroom setting, for instance by participating in an immersion experience. Most of the constructs in Hi-LAB were designed to capture potential for language learning processes that operate in such non-instructional settings, where superior cognitive and perceptual abilities of the learner may enhance the processing of language input and facilitate the mapping in memory of apperceived forms, meaning and function. Hi-LAB does not measure phonetic coding ability (important in learning to read in a foreign language) or grammatical sensitivity (important in explicit language instruction), since those measures already exist and are hypothesized to be more relevant at initial stages. Instead, the focus was on measuring potential for dealing with the remaining language learning problems, such as mastering complex linguistic systems and perceiving non-salient language features. Therefore, we have focused on cognitive and perceptual abilities that are hypothesized to support more advanced aspects of L2 learning that are required to attain high-level proficiency.

Limitations

As with any cross-sectional study, we cannot be certain about the direction of causality in the links between our examined predictors and the outcomes of interest. For example, it is intuitively apparent that better associative memory ability could enhance learning of the vast amounts of declarative knowledge (e.g., vocabulary) that is needed to advance one's language proficiency. But it is equally possible that the process of learning a foreign language to high levels provides ample opportunity to enhance one's associative memory abilities through practice. While we acknowledge this limitation, we highlight that this study was designed as a preliminary exploration of discriminant validity and so provides critical empirical evidence of the relationship between the predictors and high-level learning outcomes. The results suggest that high-level language aptitude is comprised, in part, of cognitive abilities related to attention, memory, and implicit and explicit learning. As noted previously, future studies employing longitudinal designs will allow us to disambiguate the causal direction of these relationships and show that high-level language aptitude, measured using the constructs in Hi-LAB, can predict future language

learning potential. Furthermore, such a design would provide an empirical examination of whether language learning experiences can affect any components of high-level language aptitude. This remains an open question in the field.

Another limitation of this study pertains to the lack of a comparison of Hi-LAB against other aptitude tests or a measure of general intelligence. No existing aptitude tests were designed to predict high-level proficiency, and the extant research suggests that high-level language aptitude may differ from traditional conceptualizations of aptitude (which focus on predicting rate of learning at earlier stages) or general intelligence. However, no study to date has provided conclusive evidence that the constructs comprising aptitude for high-level proficiency differ from those comprising aptitude for initial stages of learning. To provide empirical evidence in support of these claims, the constructs measured by Hi-LAB must be examined alongside other aptitude tests and a test of general intelligence to determine whether Hi-LAB enhances prediction of high-level attainment. Logistical constraints precluded this possibility in the current study. Therefore, future studies should also include one or more alternative measures of aptitude and general intelligence in addition to Hi-LAB.

Future Directions

As noted previously, the results of this study should be further validated within a longitudinal study. A longitudinal design has a number of benefits. It will provide unambiguous evidence of the predictive utility of Hi-LAB by measuring aptitude on a number of individuals prior to their extensive higher-level language training. A currently planned longitudinal study should allow for the measurement of aptitude at multiple time points, along with language proficiency, in order to better understand the causal direction of relationships between aptitude components and language learning outcomes. That is, the study should allow us to disentangle the effects of aptitude on learning from any effects that the language learning experience may have on aptitude. As no such study has been conducted to date, this would represent a major contribution to the field of SLA that has the potential to make significant advances to theoretical models of language aptitude. It would also inform the current debate on the cognitive benefits of bilingualism (e.g., Bialystok, 2010).

Additional work is clearly needed to improve the criterion model—that is, the measurement of high-level language learning outcomes. The current study employed a coarse criterion measure, which may be a potential limiting factor on these results. Future studies should employ a set of more fine-grained

criterion measures to provide more robust measurement of language learning outcomes. This would also allow for more specific predictions regarding the role of subcomponents of aptitude for particular outcomes. An enhanced criterion model would also increase the statistical and inferential power of any analyses through its richer performance metric.

Conclusion

We have provided the first empirical evidence that highly successful adult language learners can be distinguished from other individuals with moderate success based on a set of cognitive and perceptual abilities that were hypothesized a priori to be related to high-level language learning outcomes. Specifically, our results indicate that working memory (specifically, task set switching), phonological short-term memory, associative memory, and implicit learning all contributed substantially to the group discrimination analyses, and the use of listening-oriented measures—including auditory measures of working memory and measures of auditory perceptual acuity—likely optimized the validity of Hi-LAB for high-level listening attainment. Thus, these abilities are plausible candidate components of the construct of high-level language aptitude.

Final revised version accepted 6 March 2013

Notes

1 It should be noted that grammatical sensitivity is often predictive of adult success in instructed settings.
2 Other constructs are likely to contribute to successful high-level learning (e.g., inductive reasoning ability), particularly when considering abilities specifically relevant to reading. We are examining additional constructs in ongoing and future studies.
3 We also computed internal consistency reliability estimates separately for each subgroup to ensure that between-group differences in reliabilities were not impacting our analyses. Reliability estimates were highly similar across the subgroups.
4 Prior to analysis, scores for which lower values indicate better performance were reverse coded (i.e., multiplied by -1), so that for every score, higher is better. Thus, for all predictors, a positive β value indicates that predicted high attainment increases with an increase in the corresponding score, whereas a negative β value indicates that predicted high attainment decreases with an increase in the corresponding score.

References

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

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning*, *59*, 249–306.

Bialystok, E. (2010). Bilingualism. *WIREs Cognitive Science*, *1*, 559–572. doi: 10.1002/wcs.43

Birdsong, D. (2004). Second language acquisition and ultimate attainment. In A. Davies & C. Elder (Eds.), *Handbook of applied linguistics* (pp. 82–105). Malden, MA: Blackwell.

Birdsong, D. (2009). Age and the end state of second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *The new handbook of second language acquisition* (pp. 401–424). Bingley, UK: Emerald Group.

Bongaerts, T. (1999). Ultimate attainment in L2 pronunciation: The case of very advanced late L2 learners. In D. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (pp. 133–159). Mahwah, NJ: Lawrence Erlbaum.

Bongaerts, T., Mennen, S., & van der Slik, F. (2000). Authenticity of pronunciation in naturalistic second language acquisition: The case of very advanced learners of Dutch as a second language. *Studia Linguistica*, *54*, 298–308.

Brehmer, Y., Westerberg, H., & Bäckman, L. (2012). Working-memory training in younger and older adults: Training gains, transfer, and maintenance. *Frontiers in Human Neuroscience*, *6:63*, doi: 10.3389/fnhum.2012.00063

Brooks, P. J., Kempe, V., & Sionov, A. (2006). The role of learner and input variables in learning inflectional morphology. *Applied Psycholinguistics*, *27(2)*, 185–209.

Bunting, M. F., Cowan, N., & Saults, J. S. (2006). How does running memory span work? *The Quarterly Journal of Experimental Psychology*, *59*, 1691–1700.

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

Carroll, J. B. (1985). Second-language abilities. In R. J. Sternberg (Ed.), *Human abilities: An information-processing approach* (pp. 83–103). New York: W. H. Freeman.

Carroll, S. E. (1995). On the irrelevance of verbal feedback to language learning. In L. Eubank (Ed.), *The current state of interlanguage studies in honor of William E. Rutherford* (pp. 73–88). Amsterdam: John Benjamins.

Carroll, J., & Sapon, S. M. (1959). *Modern language aptitude test*. New York: Psychological Corporation.

Costa, A., & Santesteban, M. (2004). Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and L2 learners. *Journal of Memory and Language*, *50(4)*, 491–511.

Defense Language Institute Foreign Language Center. (2009). *Defense language proficiency testing system 5 framework*. Retrieved from http://www.dliflc.edu/file.ashx?path=archive/documents/Framework_Document_Sep_10_09.pdf

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

DeKeyser, R. M., Alfi-Shabtay, I., & Ravid, D. (2010). Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics*, *31*, 413–438.

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

Doughty, C. (2013). Assessing aptitude. In A. Kunnan (Ed.), *The companion to language assessment* (pp. 25–46). Oxford, UK: Wiley-Blackwell.

Doughty, C., Campbell, S., Bunting, M., Mislevy, M., Bowles, A., & Koeth, J. (2010). Predicting near-native L2 ability. *Proceedings of the 2008 Second Language Research Form*. Cascadilla Press. http://www.lingref.com/cpp/slrf/2008/index.html

Ehrman, M. E., & Oxford, R. L. (1995). Cognition Plus: Correlates of language learning success. *Modern Language Journal*, *79*, 67–89.

Fontanini, I., & Tomitch, L. M. B. (2009). Working memory capacity and L2 university students' comprehension of linear texts and hypertexts. *International Journal of English Studies*, *9*, 1–18.

Fozard, J. L., Vercruyssen, M., Reynolds, S. L., Hancock, P. A., & Quilter, R. E. (1994). Age differences and changes in reaction time: The Baltimore longitudinal study of aging. *Journal of Gerentology*, *49*, 179–189.

Gathercole, S. E., Pickering, S. J., Hall, M., & Peaker, S. M. (2001). Dissociable lexical and phonological influences on serial recognition and serial recall. *The Quarterly Journal of Experimental Psychology*, *54A*, 1–30.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge.

Granena, G. (2012). *Age differences, cognitive aptitudes and ultimate L2 attainment*. Unpublished doctoral dissertation, University of Maryland, College Park, MD.

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*(1). DOI: 10.1177/0267658312461497

Harrington, M., & Sawyer, M. (1992). L2 working memory capacity and L2 reading skill. *Studies in Second Language Acquisition*, *14*, 25–38.

Hyltenstam, K., & Abrahamsson, N. (2003). Maturational constraints in SLA. In C. J. Doughty & M. H. Long (Eds.), *Handbook of second language acquisition* (pp. 538–588). Oxford: Blackwell.

Ioup, G., Boustagui, E., El Tigi, M., & Moselle, M. (1994). Reexamining the critical period hypothesis: A case study of successful adult SLA in a naturalistic environment. *Studies in Second Language Acquisition*, *16*, 73–98.

Kersten, A. W., & Earles, J. L. (2001). Less really is more for adults learning a miniature artificial language. *Journal of Memory and Language*, *44*, 250–273.

Larsen-Freeman, D., & Long, M. H. (1991). *An introduction to second language acquisition research*. Oxford: Oxford University Press.

Linck, J. A., Schwieter, J. W., & Sunderman, G. (2012). Inhibitory control predicts language switching performance in trilingual speech production. *Bilingualism: Language and Cognition*, *15*, 651–662.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, NJ: Wiley.

Long, M. H. (in press). Maturational constraints on child and adult SLA. In G. Granena & M. H. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment*. Amsterdam: John Benjamins.

Long, M. H. (2003). Stabilization and fossilization in interlanguage development. In C. J. Doughty & M. H. Long (Eds.), *Handbook of second language acquisition* (pp. 487–535). Oxford: Blackwell.

Long, M. H. (2005). Problems with supposed counter-evidence to the critical period hypothesis. *IRAL*, *43*, 287–317.

Long, M. H. (2007). Age differences and the sensitive periods controversy in SLA. In M. H. Long (Ed.), *Problems in SLA* (pp. 43–74). Mahwah, NJ: Lawrence Erlbaum.

Mackey, A., Philp, J., Egi, T., 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* (Vol. 2, pp. 181–209). Amsterdam: John Benjamins.

Mislevy, M., Annis, R., Koeth, J., Campbell, S., Linck, J., Bowles, A., & Doughty, C. J. (2009). *Final Hi-LAB assessment utilization argument*. Center for Advanced Study of Language Technical Report. College Park: University of Maryland.

Mislevy, M., Linck, J., Campbell, S., Jackson, S., Bowles, A., Bunting, M., & Doughty, C. J. (2010). *Predicting high-level foreign language learning: A new aptitude battery meets reliability standards for personnel selection tests*. Center for Advanced Study of Language Technical Report. College Park: University of Maryland.

Miyake, A., & Friedman, N. P. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. F. Healy & L. E. Bourne, Jr. (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 339–364). Mahwah, NJ: Lawrence Erlbaum.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., & Howerter, A. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49–100.

Muñoz, C., & Singleton, D. (2011). A critical review of age-related research on L2 ultimate attainment. *Language Teaching*, *44*, 1–35.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*, 11–28.

Oxford, R. L. (1993). Instructional implications of gender differences in language learning styles and strategies. *Applied Language Learning*, *4*, 65–94.

Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning novel phonological contrasts depends on interactions between individual differences and training paradigm design. *Journal of the Acoustical Society of America*, *130*, 461–472.

Pimsleur, P. (1966). Testing foreign language learning. In A. Valdman (Ed.), *Trends in language teaching*. New York: McGraw-Hill Company.

Psychology Software Tools. (2011a). E-Prime™ [Computer Software]. Pittsburgh, PA: Psychology Software Tools, Inc.

Psychology Software Tools. (2011b). Serial Response Box™ [Apparatus]. Pittsburgh, PA: Psychology Software Tools, Inc.

R Development Core Team (2009). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. URL http://www.R-project.org

Robinson, P. (Ed.). (2002). *Individual differences and instructed language learning*. Amsterdam: John Benjamins.

Robinson, P. (2007). Aptitudes, abilities, contexts, and practice. In R. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 256–286). New York: Cambridge University Press.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

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.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: J. Wiley & Sons.

Schneiderman, E. I., & Desmarais, C. (1988). The talented language learner: Some preliminary findings. *Second Language Research*, *4*(2), 91–109.

Selinker, L. (1972). Interlanguage. *IRAL: International Review of Applied Linguistics in Language Teaching*, *10*(3), 209–231.

Skehan, P. (2002). Theorising and updating aptitude. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 69–93). Amsterdam: John Benjamins.

Sternberg, R. J. (2002). The theory of successful intelligence and its implications for language aptitude testing. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 13–43). Amsterdam: John Benjamins.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*, 498–505.

Unsworth, N., Schrock, J. C., & Engle, R. W. (2004). Working memory capacity and the antisaccade task: Individual differences in voluntary saccade control. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *30*, 1302–1321.

van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, *18*, 681–694.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

Vatz, K., Tare, M., Jackson, S., & Doughty, C. J. (2013). Aptitude-treatment interaction studies in second language acquisition: Findings and methodology. In G. Granena & M. H. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 273–292). Amsterdam: John Benjamins.

Was, C. A., & Woltz, D. J. (2007). Reexamining the relationship between working memory and comprehension: The role of available long-term memory. *Journal of Memory and Language*, *56*, 86–102.

White, L., & Genesee, F. (1996). How native is near-native? The issue of ultimate attainment in adult second language acquisition. *Second Language Research*, *12*, 233–265.

Willingham, D. B., Nissen, M. J., & Bullemer, P. (1989). On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 1047–1060.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. The Interagency Language Roundtable (ILR) proficiency level descriptors

Appendix S2. Hi-LAB score reliabilities and descriptive statistics

Appendix S3. Correlation matrix for the Hi-LAB test components