Testing the three-stage model of second language skill acquisition

Ryo Maie[1] and Aline Godfroid[2]

[1]Tohoku University

[2]Michigan State University

Author Note

Ryo Maie https://orcid.org/0000-0001-6119-0360

Aline Godfroid https://orcid.org/0000-0001-8011-7935

We have no conflict of interest to disclose.

Correspondence concerning this article should be addressed to Ryo Maie, Department of Theoretical Linguistics, Graduate School of International Cultural Studies, Tohoku University, 41 Kawauchi, Aoba-ku, Sendai, Miyagi 980-0862, Email: ryo.maie.e5@tohoku.ac.jp

Acknowledgements

We thank Drs. Shawn Loewen, Koen Van Gorp, Paula Winke, and Phillip Hamrick for invaluable feedback on the first author's Ph.D. dissertation on which the current article is based. We also thank Dr. Caitlin Tenison (Educational Testing Service) for sharing the Python code for running the hidden Markov modeling analysis and Surya Narayan (The University of Tokyo) for programming on Google Colaboratory. The current study is based on a larger project funded by the Doctoral Dissertation Research Improvement Grants (Linguistics Program) from National Science Foundation, and NFMLTA/MLJ Dissertation Writing Support Grants by the National Federation of Modern Language Teachers Associations.

Testing the three-stage model of second language skill acquisition

Abstract

Skill acquisition theory conceptualizes second language (L2) learning in three stages (declarative, procedural, and automatic), yet competing theoretical models with fewer stages also exist, and the number of stages has never actually been tested. We tested the validity of the three-stage model by investigating the number and nature of learning stages in L2 skill acquisition. Seventy-three participants deliberately learned grammar and vocabulary of a miniature language through explicit-deductive instruction. They systematically practiced comprehending the language until their accuracy and speed of performance did not improve anymore. Participants received a battery of tests assessing individual differences in their declarative and procedural learning ability. We first applied hidden Markov modeling to participants' reaction time data (obtained from the language practice) to compare rival hypotheses on the number of stages in L2 skill acquisition. We then examined which cognitive variables predicted participants' performances (accuracy and speed) in each stage. Our results indicated that participants indeed acquired L2 skills in three stages and that their performance correlated initially with declarative learning ability, but there was a tendency for procedural learning ability to take over in the later stages. Our findings provide the first formal evidence for the influential three-stage model of L2 skill acquisition.

Introduction

Skill acquisition theory conceptualizes second language (L2) learning as a form of skill acquisition (see DeKeyser, 2020; Suzuki, 2022 for reviews). Its basic tenet is the three-stage model, proposing that the acquisition of L2 skills follows three stages of development, described as the declarative, procedural, and automatic stage (Anderson, 1982, 1983). Although current L2 research has provided indirect evidence that supports the existence of such developmental stages (e.g., Hamrick, 2015; Morgan-Short et al., 2014; Pili-Moss et al., 2020), the stages themselves have never become the object of inquiry. In this study, we tested the validity of the three-stage model by (a) identifying how many stages learners go through while learning a foreign language, and (b) examining whether cognitive abilities hypothesized to underlie each stage of learning do indeed play a role in that stage.

Participants in our study received explicit-deductive instruction on a new foreign language and subsequently practiced their comprehension of the language until their performance reached an asymptote (i.e., the point at which no further significant improvement is expected). To formally test the number of stages shown in their practice, we drew on recent research in cognitive psychology to model the skill acquisition process mathematically (e.g., Tenison & Anderson, 2016). We used this computational modeling to pit rival hypotheses about the number of skill acquisition stages (i.e., one, two, or three stages) against each other. Participants also completed a battery of cognitive tests measuring their individual differences in declarative and procedural learning abilities. We tested the nature of the identified learning stages by determining which learning ability predicted participants' performances in each stage.

In the following sections, we first introduce a cognitive theory of skill acquisition, along with its rival models, which have formed the basis of skill acquisition theory in L2 learning. We

then review the main claims of skill acquisition theory, particularly in reference to the three-stage model. After reviewing relevant empirical L2 research, we discuss a methodological barrier that has prevented L2 researchers from testing the theoretical models. We lastly draw on a methodology employed by Tenison and Anderson (2016) in cognitive psychology as a solution that can provide a way to directly juxtapose the theoretical models in L2 contexts.

Literature Review

Cognitive Theories of Skill Acquisition

It takes many hours of practice to learn a language. There is a consensus that as people practice, their performance changes in predictable ways: they become more accurate, faster, and more consistent in their performance. What is less clear, however, is whether these quantitative changes (e.g., continuous improvement in accuracy and speedup) also correspond with qualitative changes in the mechanisms underlying the skilled performance. Such qualitative changes would reflect different "stages" in the acquisition of a skill. In other words, when learners improve (or become faster in performing) the same psychological process, they do so within a developmental stage (a quantitative change). However, when learners progress from one stage to another, they qualitatively change the underlying cognitive process to a more efficient process (see DeKeyser, 2001 for distinguishing between quantitative and qualitative changes in L2 skills).

Skill acquisition theory in L2 learning provides an account of how L2 learning proceeds through practice. It is primarily based on Anderson's Adaptive Control of Thought (ACT) theory (Anderson, 1982, 1983) and its subsequent computer-implemented cognitive architecture, Adaptive Control of Thought-Rational (ACT-R; see Anderson, 2005, 2007). ACT-R posits that declarative and procedural knowledge play fundamental roles in skill acquisition. Declarative knowledge refers to explicit representations of factual and episodic information, whereas

procedural knowledge pertains to skill-specific routines that underlie efficient skill performance. For instance, declarative knowledge in language may be conscious knowledge of grammatical rules, such as adding *-s* at the end of a verb in the third-person singular form in English, whereas procedural knowledge pertains to applying this rule in use. In ACT-R, declarative knowledge is represented as chunks of information, whereas procedural knowledge takes the form of (skill-specific) production rules.[1] A production rule is a primitive rule in the form of a condition-action pair (or, in other words, an IF-THEN conditional) that encodes a cognitive contingency such that when the condition is met, the action is performed (Anderson, 1982). An example of a condition may be a preverbal (or intended) message one must communicate (Levelt, 1989), and the resulting response is the communication of the message through utterance. The major assumption in ACT-R is that most human behaviors are done by production rules.

Skill acquisition in ACT-R begins by acquiring declarative knowledge about a skill through instruction or observing someone else's performance. For instance, L2 learners can be explicitly taught a rule of how to conjugate verbs based on thematic roles. Initially, in the declarative stage, learners perform a skill using declarative knowledge, for instance by carefully applying the conjugation rule. However, applying declarative knowledge is a laborious task because it requires the knowledge to be retrieved from long-term memory and maintained in working memory. Consequently, learners develop skill-specific procedures, or procedural knowledge, to optimize their performance. This process, termed proceduralization, creates specialized production rules committed to the target skill. Procedural knowledge obviates the need for retrieving declarative knowledge because it provides a direct mapping between a condition and an action. Hence, in the procedural stage, learners directly produce a conjugated form without drawing on the rule. However, this procedural knowledge is weak in its representation and may still be inaccurate. They

go through a process called production tuning, through which learners incrementally weigh the applicability of procedural knowledge in specific contexts to increase the accuracy and speed of the knowledge application. This gradual tuning of performance eventually leads to automaticity (the automatic stage), enabling one to execute the skill quickly and effortlessly without the need for focal attention.

ACT is considered a rule-based approach due to its reliance on production rules as a basic form of knowledge representation. This feature of the model contrasts with two rival models of skill acquisition frequently cited in L2 research (DeKeyser, 2001): the race model and the component power laws (CMPL) theory, both of which adopt an item-based approach. The race model, proposed as part of the instance theory of automatization (Logan, 1988, 2002), claims that learners store and represent each experience of skill performance as an instance. An instance is stored in episodic memory and encodes contextual and task-specific cues relevant to the skill. For example, in L2 learning, learners can store memories of conjugating a verb in specific linguistic contexts (e.g., thematic roles, number, and gender). The more instances learners accumulate in memory, the faster their performance becomes. When learners encounter the same task again, each of the previously stored instances races against each other, with the fastest one producing the behavior. The race model suggests a single-stage model, theorizing that skill acquisition is solely driven by the continuous amassing of instances, which corresponds to a quantitative (rather than a qualitative) change in the mechanism.

The CMPL theory (Rickard, 1997; for a recent review, see Bajic & Rickard, 2011) similarly posits instance learning as the underlying mechanism. However, the CMPL theory additionally proposes that initially, learners can choose from two different strategies: either applying rule-based, algorithmic processing to figure out the task or retrieving the past solution (i.e., an instance) from

memory. The CMPL theory conceives automatization as the result of the transition from algorithmic processing to memory retrieval. This is where CMPL theory differs from the race model by hypothesizing a two-stage model based on the transition between two different mechanisms. Note that neither the race model nor the CMPL theory incorporates procedural knowledge (or memory) in their learning mechanisms, as episodic memory (storing instances) is a type of declarative memory (Squire & Zola, 1996).

Table 1 summarizes predictions from the three theoretical models of skill acquisition reviewed here. ACT-R differs from the two rival models in terms of (a) the number of cognitive stages (three vs. one vs. two) and (b) which type of memory (or knowledge) is involved (declarative/procedural vs. declarative only). In L2 research, ACT-R and the associated learning mechanisms have informed the skill acquisition theory in L2 learning, to which we will turn next.

Table 1

Prediction from the Three Models of Skill Acquisition

	Stage 1	Stage 2	Stage 3
The race model	declarative	–	–
The CMPL theory	declarative	declarative	–
ACT-R	declarative	procedural	automatic

Skill Acquisition Theory

Skill acquisition theory (DeKeyser, 2001, 2020; Suzuki, 2022) states that L2 learning can be explained as a form of skill acquisition and that the development of L2 skills follows three distinct cognitive stages: (a) the declarative stage, in which learners acquire declarative knowledge about a language and use this knowledge to work out performance; (b) the procedural stage, in which learners develop procedural knowledge that dramatically facilitates performance routines;

and (c) the automatic stage, in which those routines become automatic, allowing the skill to be performed effectively as a reflex.

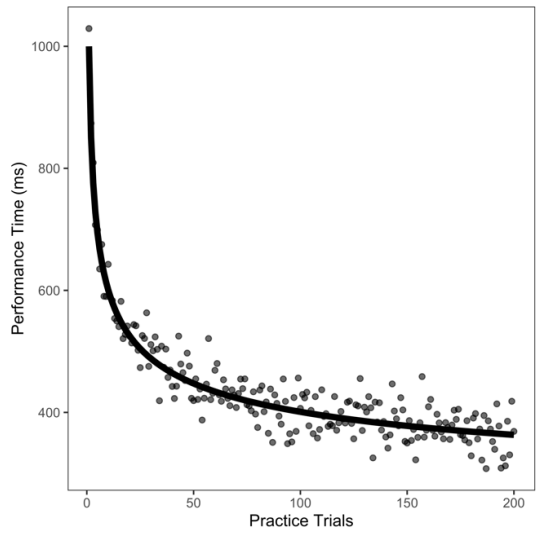
Researchers agree that achieving automatic L2 use entails prolonged practice using the language. With practice, behavioral indicators of skill acquisition, especially performance times, are known to decrease following the power-law of practice (DeKeyser, 1997, 2020; see Newell & Rosenbloom, 1981 for initial evidence in cognitive psychology). This scientific law states that the time it takes to perform a skill decreases with practice, and the decrease follows a specific non-linear curve defined by a power function. Figure 1 illustrates an example of a power function applied to skill acquisition data. The power-law of practice is characterized by an initial sharp decrease in performance times followed by a gradual decrease until it reaches the asymptote. Although theoretical interpretations of this nonlinear development vary by researchers, DeKeyser (1997, 2001) suggests that the initial rapid decrease reflects the transition from the declarative to procedural stage (proceduralization), and the subsequent incremental decrease reflects smoothing out performances to automatize a behavior (production tuning).

Skill acquisition theory constitutes one of the major theoretical approaches in L2 research (DeKeyser, 2020; Suzuki, 2022). Yet, its core claims have rarely been tested (for exceptions, see Robinson, 1997; Robinson & Ha, 1993), and no studies have tested the three-stage model in L2 contexts, let alone compared its validity against other models (see Figure 1). In reviewing skill acquisition theory, DeKeyser (2020, p.88) highlighted this fact, stating "not much research in the field of second language learning has explicitly set out to gather data [...] to test (a specific variant of) Skill Acquisition Theory." Currently, a group of studies have investigated the general roles of declarative and procedural memory systems in L2 learning. The findings from these studies suggest

that the three-stage model may be applicable in L2 contexts, thereby opening avenues for further exploration.

Figure 1

An Example of a Power Function Applied to Skill Acquisition Data



Note. The figure shows data simulated from a power function $T = 200 + 800N^{-0.3} + N(0, 30)$, where T is performance times, N is the number of practice trials, and N(0, 30) is a normal distribution with the mean of 0 and the standard deviation of 30 to add sampling error. For this dataset, R^2 = .99.

Declarative and Procedural Memory in L2 Learning

Studies examining the roles of declarative and procedural memory systems in L2 learning provide valuable insights into how skill acquisition may proceed in L2 learning. At the conceptual level, these studies draw on Ullman's (2004, 2020) declarative/procedural (D/P) model, which shares with ACT-R the basic premise of a declarative-procedural transition.[2] For L2 learning, the D/P model specifically predicts that learners initially rely on declarative memory to acquire idiosyncratic knowledge, such as vocabulary items, morphosyntactic rules, and pragmatic functions, but some aspects of grammar (and some other features requiring probabilistic learning)

can simultaneously be learned by procedural memory. As learners gain proficiency and increase automaticity in processing, procedural memory gradually takes priority.

Several empirical studies support these predictions of the D/P model (e.g., Hamrick, 2015; Morgan-Short et al., 2014; Pili-Moss et al., 2020). For instance, Morgan-Short et al. (2014) explored how learners' declarative and procedural learning abilities influenced grammar learning in an artificial language practiced over an extended period (over 20 sessions). They found that declarative learning abilities (i.e., as measured by Part V of the Modern Language Aptitude Test and the Continuous Visual Memory Task) predicted grammar learning in the initial stages, while procedural learning measures (the Tower of London and Weather Prediction Tasks) correlated with test scores in later stages of learning. Hamrick (2015) demonstrated that these findings hold even when learners were merely exposed to the language under incidental learning conditions. Additionally, a meta-analysis by Hamrick et al. (2018) confirmed this dynamic role of declarative and procedural learning abilities in L2 learning, particularly among adult learners (but see also Oliveira et al., 2023 for counterevidence).

More recently, Pili-Moss et al. (2020) analyzed the practice data (rather than the posttest data, which were the focus of Morgan-Short et al.'s original publication) collected in Morgan-Short et al. (2014). They examined how participants' declarative and procedural learning abilities predicted accuracy and the degree of automatization during practice in an artificial language. They found that declarative learning measures consistently predicted participants' accuracy of performance across all the practice sessions. For automaticity, procedural learning abilities showed a positive correlation in the later two of three (self-identified) stages of practice, but only for learners with higher declarative learning abilities in the initial stage. Although Pili-Moss et al. interpreted their results primarily within the D/P model, these findings also align with the learning

mechanism in ACT-R, suggesting that the role of procedural learning in later stages presumes successful declarative learning in initial stages.

However, one critical limitation in Pili-Moss et al. (2020) was their presupposition of the number of stages (i.e., three) and the intuitive division of practice sessions. Admittedly, it is methodologically challenging to identify the number of stages solely based on behavioral data. To a large degree, this difficulty has prevented SLA researchers from testing rival models of skill acquisition (DeKeyser, 2020, p. 88). In our study, we aimed to address this challenge by drawing on research in cognitive psychology that utilized computational modeling of RT data to model the process of skill acquisition (Anderson, 2005; Tenison & Anderson, 2016; Tenison et al., 2016). Inspired by Tenison and Anderson (2016), our empirical approach focuses on determining the number of skill acquisition stages most consistent with our data.

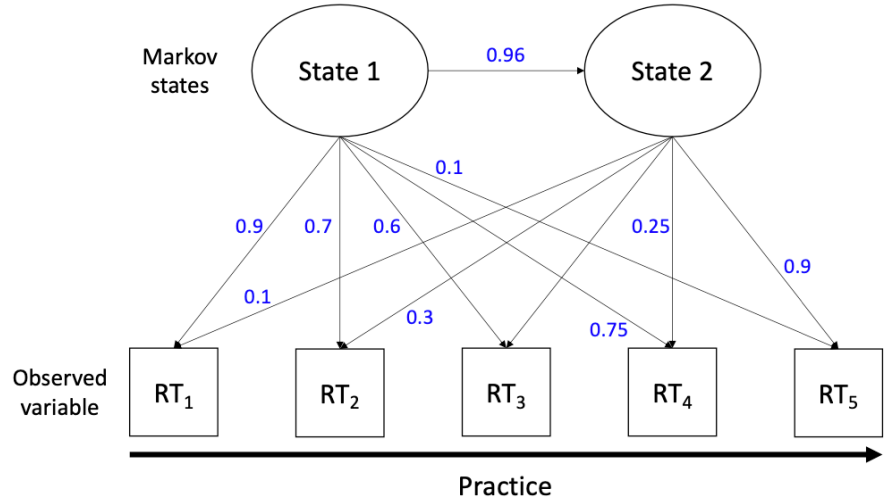
Modeling Skill Acquisition Processes

The model research for our study was Tenison and Anderson's study (2016), who investigated how learners develop fluency in an arithmetic task called the Pyramid problem. A typical problem in the Pyramid task takes the form of base$height, for instance, 8$3. Participants start with the base as the starting number and repeatedly add as many numbers as the height, with each term being one less than the previous one. For instance, the solution for 8$3 can be found as 8 + 7 + 6 = 21. In the study, participants explicitly learned the structure of the problem through instruction and then practiced solving each problem over six blocks (36 repetitions for each practiced item). Upon seeing an item (e.g., 8$3), participants solved the problem, typed the solution (e.g., 21 for 8$3), and received correctness feedback on their answer. Participants practiced the items in an MRI scanner that recorded their neural signatures during the task in addition to their accuracy and speed of performance.

The first step of the analysis focused on testing the number of learning stages. In doing so, Tenison and Anderson adopted hidden Markov modeling to model the reaction time (RT) history of each individual participant on each practice item. The hidden Markov model (HMM) is a stochastic time-series model consisting of a Markov chain, a mathematical system representing a sequence of states. The HMM is a special type of Markov chain that treats the actual states as hidden (and hence unobservable), but their probability can be estimated using observed data (in their study, RT data). HMMs are commonly used as a pattern-recognition method in computational linguistics, including in speech recognition research where HMMs are used to segment and identify words based on the temporal characteristics of speech. In Tenison and Anderson's study (2016), the hidden Markov states represented different learning stages in solving the arithmetic task. For an illustration, Figure 2 shows the structure of a HMM applied to skill acquisition data.

Figure 2

A Visual Representation of a Hidden Markov Model Applied in Skill Acquisition Research



In this example, a HMM was specified to have two Markov states, corresponding to two learning stages based on five trials of practice producing observable data. Using a vector of RTs as the dependent variable (RT_1–RT_5), the HMM estimates two types of likelihoods: (a) transition probability, defined as the probability of individuals eventually moving from one state to another (i.e., from State 1 to State 2), and (b) emission probability, defined as the probability of a data point (RT) being generated by a given state. The transition probability between the two stages in Figure 2 is .96, meaning that learners eventually transition to the second stage with a probability of .96. Additionally, the emission probability of RT_5 in State 2, for example, is .90, which means that there is a .90 probability that learners have already transitioned to the second stage at the fifth practice trial. In our study, we were especially interested in the emission probability because it can be used as an estimate of which learning stage participants occupy after a particular number of trials. A HMM analysis considers all possible trajectories of learners transitioning to subsequent stages after each practice trial (see Analysis and Appendix S5 for more details). Tenison and Anderson (2016) tested one- to five-state models, with the first three models representing the race model, the CMPL theory, and ACT-R, respectively, and the remaining four- and five-state models tested for exploration. They found that the three-state model showed the best fit, while the one- and two-state models provided substantially poorer fits, and the four- and five-state models did not improve on the three-state model.

Tenison and Anderson further hypothesized that the identified three stages would correspond to the three-stage model in ACT-R. Specifically, they predicted that the following cognitive processing were involved in each stage: direct mathematical calculations to find the answer in stage one, an effortful retrieval of the past solution from memory in stage two, and an automatic execution of the solution in stage three. To validate their predictions, they drew on their

participants' neuronal signatures to examine how activation patterns in their brains changed as the participants transitioned through the three stages. The results showed that the regions known to subserve the hypothesized cognitive processings (i.e., calculation, effortful retrieval, and automatic reflex) were indeed highly activated in the corresponding stages. These findings, combined with the results of the HMM analysis, provided converging evidence that skill acquisition (at least of the Pyramid problem) is a three-stage process and that the cognitive processes involved in the three stages are consistent with the predictions from ACT-R.

In L2 research, studies have similarly provided evidence suggesting that the existence of different developmental stages (e.g., Hamrick et al., 2018; Morgan-Short et al., 2014; Pili-Moss et al., 2020), yet a limitation has been that there was no proper methodological option to identify learning stages based on behavioral data (such as RT data). The study by Tenison and Anderson (2016) was innovative because they adopted hidden Markov modeling to test rival models of skill acquisition and validate the learning mechanisms of the three-stage model. With our study, we aim to bring this innovation to the field of L2 research. We adopt the analytical method of Tenison and Anderson (2016) to identify the number of learning stages and the research design of L2 research (such as Pili-Moss et al., 2020) to examine the role of declarative and procedural learning abilities in each identified stage.

The Current Study

We conducted an experiment in which participants deliberately learned and practiced a novel miniature language until their accuracy and speed of performance no longer improved. We measured the participants' individual differences in their declarative and procedural learning abilities to investigate which ability, declarative and/or procedural learning, may underlie skill

acquisition in each learning stage. To investigate the three-stage model of L2 skill acquisition, we addressed the following research questions:

RQ1. How many stages of skill acquisition do L2 learners go through while they learn and practice comprehending a novel miniature language?

RQ2. Which learning abilities, declarative and/or procedural learning, predict participants' performances in each stage of skill acquisition?

Research Question 1 addressed the number of learning stages in L2 skill acquisition. We hypothesized that participants would acquire L2 skills in three stages following the dominant view in SLA (DeKeyser, 2020; Suzuki, 2022). Assuming the three stages, we also hypothesized for Research Question 2 that we would observe the declarative-procedural transition proposed in ACT-R. Specifically, we predicted that participants' individual differences in declarative learning would predict their accuracy of performance in the first stage, whereas individual differences in procedural learning would predict the speed of performance (RT) in the second stage. The differential prediction for accuracy and RT is because each measure best represents development in a different stage of skill acquisition (i.e., accuracy in initial stages and speed in later stages) (DeKeyser, 1997, 2020). Because performance in the third (automatic) stage is highly stabilized, we hypothesized that neither declarative nor procedural learning would predict performance accuracy or speed in the third stage.

Methods

Participants

We recruited seventy-three L1 speakers of English who had not studied any case-marking languages. We excluded eight participants because they either did not complete the study or provided responses that were psychophysically implausible for the experimental task at hand (e.g.,

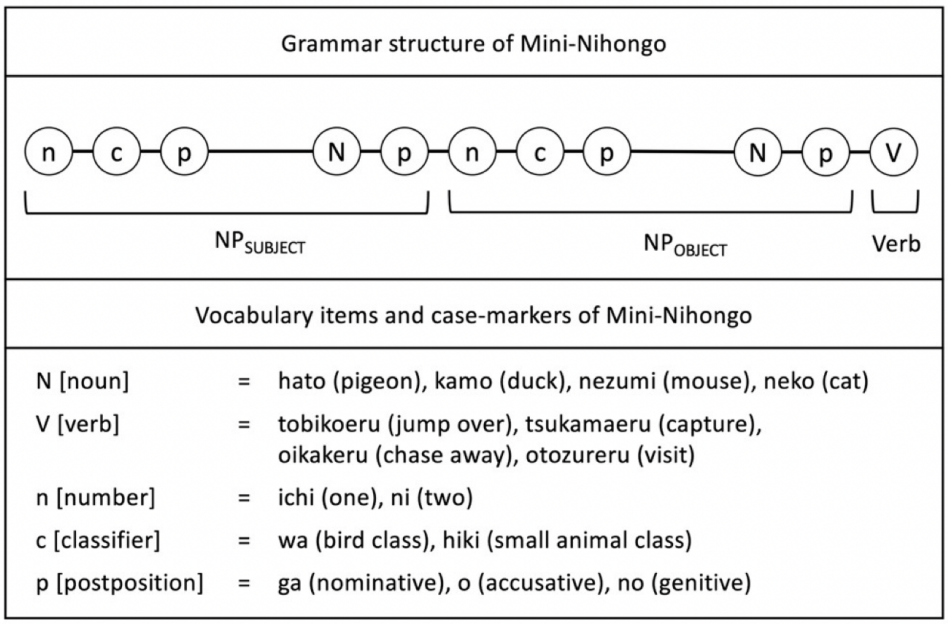
RT lower than 300 milliseconds in 90.26% of trials in the final practice session). The final sample consisted of 65 participants (46 female, 14 male, and 5 non-binary or not specified) with a mean age of 20.35 years old (*SD* = 2.61, Range = 18–30). On average, participants knew 1.2 additional languages (*SD* = 0.79, Range 0–4). Most were students at Michigan State University who were not majoring in linguistics or language teaching fields. They received 60 U.S. dollars upon the completion of the study.

Language

We used a miniature version of Japanese, called Mini-Nihongo (which translates as "Mini-Japanese"), originally developed by Mueller and colleagues (see Mueller, 2006 for a review). The researchers showed that grammatical and semantic violations in Mini-Nihongo elicit ERP (i.e., event-related potential) signatures that are identical to those that L1 speakers show when comprehending Japanese. The language thus preserves an appropriate level of complexity and naturalness in its system while at the same time allowing L2 learners to be trained to advanced proficiency within a relatively brief period of time. Figure 3 illustrates the entire structure of Mini-Nihongo adopted in our study.

Figure 3

The Structure of Mini-Nihongo



Mini-Nihongo consisted of five grammatical categories: four nouns, four verbs, two numerals, two numerical classifiers, and three postpositions. Although Japanese in general allows scrambling of words, we used only the Subject-Object-Verb order, which is canonical in Japanese. Hence, a sentence always contained two noun phrases followed by a main verb, with the first noun phrase corresponding to the grammatical subject and the second to the object. A noun phrase consisted of a case-marked head noun modified by a numeral and a classifier. In Japanese, numbers are not marked morphologically and hence must be conveyed by numerical classifiers. The choice between the two classifiers depended on whether the noun they marked was a bird (*hato*, *kamo*) or another type of small animal (*nezumi*, *neko*). The postposition -*ga* was the nominative marker, -*o* was the accusative marker, and -*no* was the genitive marker. Numerals and classifiers must be marked by the genitive marker in order to connect to the head noun (e.g., ichi-hiki *no* neko: “a

cat"). The entire language consisted of 256 unique sentences. Each sentence was matched with a colored picture. Because each practice session consisted of 128 practice trials, we divided the stimulus list into two sets (List A and List B) and counterbalanced the order of the stimuli across practice sessions. All the stimuli and the corresponding pictures can be found at: https://osf.io/x9u6h/.

General Procedure

Table 1 illustrates the procedure of the study. We collected data over six sessions (Day 1–6). In principle, participants completed the study over six consecutive days, but they were granted a two-day interval in case of emergencies. On average, participants completed the study in 6.13 days (*SD* = 0.39). We administered the entire study online on Gorilla™ (https://app.gorilla.sc/). Because participants completed the entire study at home, we monitored their performance each day so that any unusual responses (such as those qualifying for exclusion, see Participants section) could be flagged for review.

Table 1

The Procedure of the Entire Study

Day 1 (39 minutes)	
Task	Min
1. Background questionnaire	1
2. Two-choice reaction time task	3
3. Alternating serial reaction time task	15
4. Statistical learning task	20
Day 2 (60 minutes)	
1. Continuous visual memory task	10
2. LLAMA-B	10
3. Explicit instruction of Mini-Nihongo	20
4. Vocabulary and grammar tests	5
5. Warmup practice of Mini-Nihongo	15
Day 3 (70 minutes)	
1. Vocabulary and grammar tests	5
2. Comprehension practice	20
3. Production practice	45
Day 4 (65 minutes)	
Task	Min
1. Vocabulary and grammar tests	5
2. Production practice	40
3. Comprehension practice	20
Day 5 (60 minutes)	
1.Vocabulary and grammar tests	5
2. Comprehension practice	20
3. Production practice	35
Day 6 (55 minutes)	
1. Vocabulary and grammar tests	5
2. Production practice	30
3. Comprehension practice	20

Note: This study is part of a larger project, for which participants also completed an additional (production) practice task and a cognitive test (two-choice RT task). The results for these tasks are detailed Maie (2022).

On Day 1, participants first completed a consent form and filled out a background questionnaire (https://osf.io/x9u6h/). They then completed two tasks of procedural learning, namely an alternating serial reaction time task followed by a statistical learning task. On Day 2, participants first took two declarative learning tasks, the Continuous Visual Memory Task and LLAMA-B (see Cognitive Tests section for details on the cognitive measures). Afterward, they received explicit-deductive instruction in Mini-Nihongo by watching a 19-minute video (see Online Supporting Materials for the explicit instruction). Vocabulary and grammar knowledge tests followed the instruction to ascertain that participants had indeed developed explicit, declarative knowledge of the language (see Figure S5 for results). Participants were then guided to do warmup practice, which served to familiarize them with the format of the tasks used for language practice

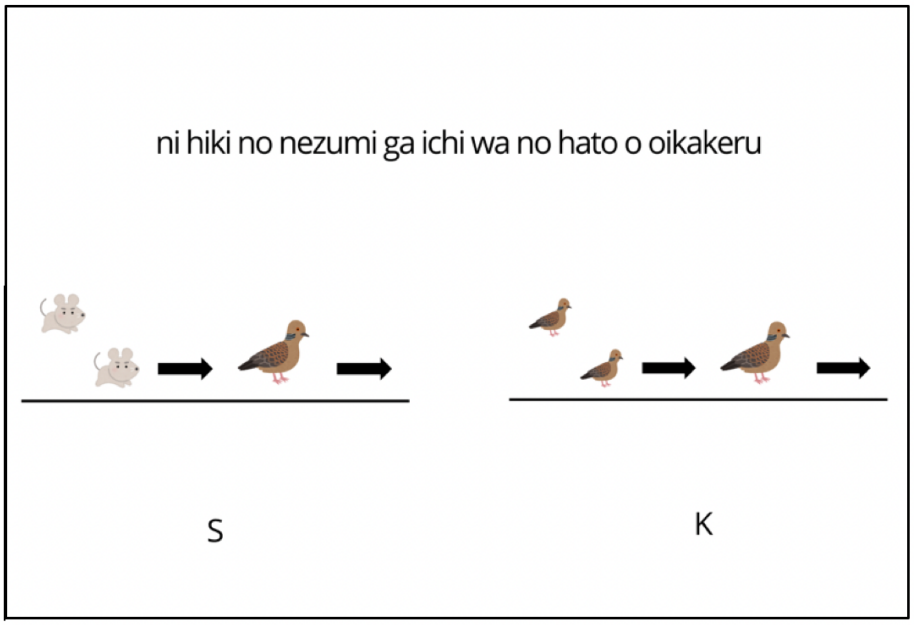
(see the next section for details). The remaining Day 3–Day 6 had an identical structure: the vocabulary and grammar knowledge tests were administered first and then came language practice.

Language Practice

In Day 3–6, participants engaged in comprehension practice of Mini-Nihongo. The practice was a sentence-picture matching task. Participants saw a sentence with two pictures and chose which picture matched the written sentence displayed above by pressing the corresponding (S, for left or K, for right) key. Figure 4 shows an example of the task. In our study, the picture options differed in terms of either (a) the subject noun, (b) the number on the subject (i.e., one or two), (c) the object noun, (d) the number on the object, or (e) the verb. Note that word order and case markers could not be tested directly because their positions were fixed in Mini-Nihongo. However, making a correct decision in the task required participants to understand and process those features; that is, they must recollect the word order (i.e., S-O-V) and the case-marking rules to understand which noun phrase in the sentence corresponded to the subject or the object. We randomized the presentation of test items, as well as the position of the correct answer (left or right) on the screen. Each of the five critical features was practiced through an equal number of test items. Participants received accuracy feedback throughout the practice sessions: they saw a green checkmark for correct responses and a red cross for incorrect responses.

Figure 4

Example of the Comprehension Practice Task



Note. The correct answer is the left picture in this example as the subject *ni hiki no nezumi* indicated by the subject marker *-ga* means "two mice".

Before the main practice sessions in Day 3–6, participants were guided to do an initial warmup practice in Day 2. This was a familiarization period during which participants became used to the format of the comprehension task. Hence, the warmup practice took the same format as the main practice task except that trials dealt with individual words (i.e., nouns and verbs) and noun phrases ([number] [animal class] [possessive] [noun]) and then built up progressively to full sentences. When practicing individual words or noun phrases, participants saw a word (or a noun phrase) together with two pictures and responded by choosing the matching picture. There were 24 word-level trials (eight content words repeated three times), eight phrase-level trials, and 16 sentence-level trials.

Each practice session (Day 3–6) consisted of 128 trials, presented in 8 blocks of 16 trials each. After each block, participants were allowed to take a 3-5 minute break. Combining all four practice sessions and the warmup practice (16 trials), participants completed a total of 528 practice trials in 33 blocks, all of which were included in the data analysis. Participants practiced the same list of items, but the order of presentation was randomized within the list.

Cognitive Tests

In our study, we focused on declarative and procedural learning as cognitive abilities underlying the acquisition of cognitive skills (Anderson, 1982, 1983). We predicted that they are also important in L2 skill acquisition (Pili-Moss et al., 2020). There were two tasks for each ability: the Continuous Visual Memory Task (CVMT) and LLAMA-B for declarative learning, and an alternating serial reaction time task (ASRT) and a statistical learning task (SL) for procedural learning. Within each dimension, the former task was a non-linguistic measure and the latter task was a linguistic measure. In Online Supporting Materials (Appendix S2), we provide further details on the cognitive tests.

Cognitive Visual Memory Task

The CVMT tests one's ability for nonverbal declarative learning using a visual recognition paradigm (Trahan & Larrabee, 1988; see Buffington & Morgan-Short, 2019 for a review in SLA). During the task, participants saw a series of complex abstract designs (k = 123, composed of 11 practice trials and 112 test trials), and they were tested on their ability to recognize seven target designs repeated seven times (k = 49) among the other designs (k = 63), which served as distractors. In the CVMT, learning ability is measured using d-prime scores, which is a standardized measure that represents a participant's ability to discriminate old designs from new designs in our study. We calculated d-prime scores by subtracting the z-score for the proportion of old (repeating) items

that were incorrectly labeled as new items (i.e., false alarms) from the *z*-score for the proportion of new items that were correctly labeled as new items (i.e., hits). A d-prime score of 0 indicates a complete lack of ability (i.e., chance-level performance). Scores can range from 4.65 (the effective limit: 99%) to –4.65 (1%), with a higher score indicating better declarative learning ability. The internal consistency of the task based on the Kuder-Richardson Formula 20 (KR-20) (equivalent to Cronbach's alpha but used for dichotomous items) was acceptable at KR20 = .72 [.62, .81].

LLAMA-B

LLAMA-B assesses one's ability to learn the names of unfamiliar objects (Meara & Rogers, 2019). In the task, participants were given two minutes to associate 20 unfamiliar objects with their names and then tested on how many of the new form-meaning associations they were able to recall. The original task comes with a unique graphical user interface that allows test-takers to move a cursor over an object to see its name. However, this feature was not available in Gorilla; instead, we presented an array of objects with their names together in a single screen. This presentation format made the task more similar to Part V of the Modern Language Aptitude Test, a conceptual model of LLAMA-B. The testing phase consisted of 20 items with no time limit. We used percentile scores as participants' declarative learning scores. The internal consistency of the task was similarly acceptable at KR20 = .76 [.66, .84].

Alternating Serial Reaction Time Task

ASRT examines one's ability for implicit (or procedural) sequence learning (Howard & Howard, 1997). Participants saw an array of four circles aligned horizontally in the center of the screen. The circles were empty, but one of them was filled with an orange bird and then became empty again before another circle filled. The sequence in which the circles were filled followed a second-order conditional rule where (systematic, rule-based) pattern trials were interleaved with

random trials (e.g., 3r1r2r4r, where *r* denotes a random position). Participants pressed a key corresponding to the filled circle position as quickly and accurately as possible, mirroring the position of the orange bird with their fingers. The task consisted of 15 blocks of 88 trials each. The first eight trials were random practice trials. We followed Godfroid and Kim (2021) and calculated the learning score by taking the mean of each participant's RT over the final block (Block 15) and subtracted the means on the pattern trials from those on the random trials. Any incorrect responses or those with RT lower than 100 milliseconds or not within the range of an individual's mean $\pm$ 3 standard deviations (SD) were removed from the analysis (1.6% of the dataset). To estimate the reliability of the learning scores, we took the mean of participants' RT on the pattern and random trials and calculated Cronbach's alpha between the two items. The internal consistency of the scores was = .95 [.91, .96].

Statistical Learning Task

The statistical learning task measured one's ability to learn adjacent and non-adjacent relationships. We adopted a statistical learning task of Romberg and Saffran (2013, Experiment 1) as a linguistic measure of the ability (i.e., verbal statistical learning) that underlies proceduralization. The target stimuli consisted of a list of three-word phrases in the form of A-X-B. There were three words for A words (*pel*, *vot*, and *dak*), three words for B words (*rud*, *jic*, and *tood*), and sixteen words for X words (*balip*, *benez*, *deecha*, *fengle*, *gensim*, *gople*, *hiftam*, *kicey*, *loga*, *malsig*, *plizet*, *puser*, *roose*, *skiger*, *suleb*, and *vamey*). Crucially, each A word was paired with a B word as a categorical non-adjacent dependency frame (*pel_rud*, *vot_jic*, and *dak_tood*), but the relationship between A words and X words, and B words and X words was probabilistic (see Online Supporting Materials for details). Participants were exposed to a list of 72 three-word phrases (i.e., A-X-B) that were repeated four times (288 trials in total) and were then tested on their

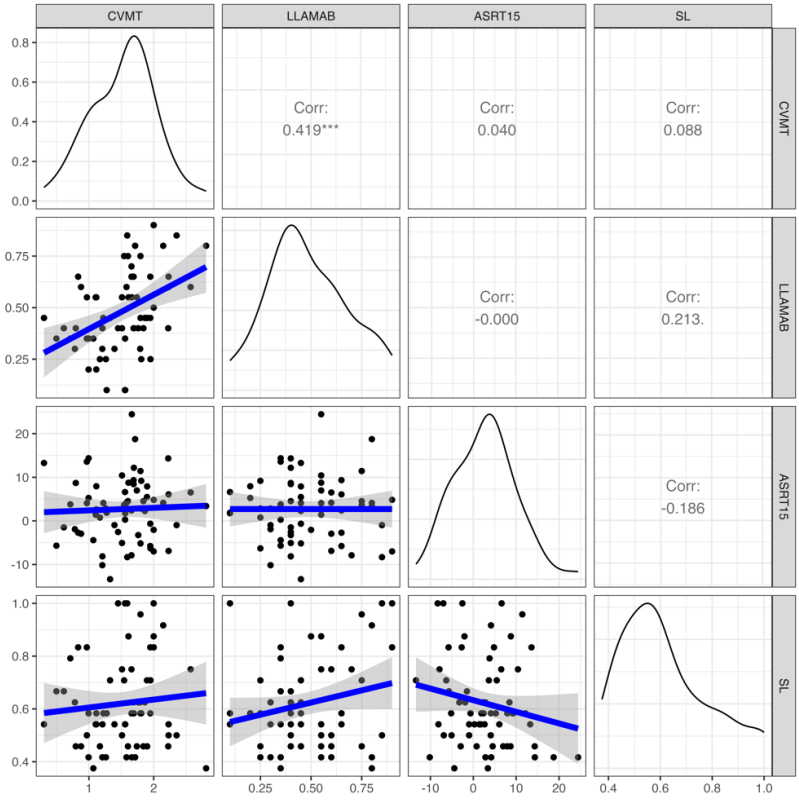
knowledge of adjacent and non-adjacent dependencies. In the test, participants saw two phrases in a sequence and decided which one sounded more familiar to them. There were 30 items: 12 items testing non-adjacent dependencies, 12 items testing adjacent dependencies, and 6 items as attention-checking items. The test stimuli can be found at https://osf.io/x9u6h/. The internal consistency based on KR-20 was acceptable at .70 [.60, .81].

Relationships between the Cognitive Variables

Figure 5 shows the correlations among the four cognitive tests. As expected, participants' scores on CVMT and LLAMA-B showed a correlation because both are hypothetical measures of declarative learning ability (r = .42, 95% CI [.19, .60], p < .001). The strength of their association was moderate and of comparable magnitude or larger than those reported in previous research (e.g., Morgan-Short et al., 2014 with r = .149 for MLAT-CVMT; Buffington et al., 2021 with r = .469 for DecLearn-CVMT and r = .273 for MLAT-CVMT). However, a similar correlation was not observed between the two procedural learning tasks (i.e., ASRT and SL: r = -.18, 95% CI [-.41, .06], p = .13). There could be multiple reasons for this (see Li & DeKeyser, 2021 for a recent overview of SLA-focused research, finding similar issues with the construct validity of implicit language aptitude), but because of the lack of correlation, we did not attempt data reduction on the procedural learning measures. We did combine the scores from CVMT and LLAMA-B into a single declarative learning ability score. Specifically, because the CVMT -LLAMA-B correlation aligned with our theory-based predictions, we conducted an exploratory factor analysis on the cognitive tests and extracted the corresponding factor scores for declarative learning ability (see Online Supporting Materials for details). We retained the ASRT and SL scores as two distinct individual differences measures of nonverbal and verbal procedural learning, respectively.

Figure 5

Correlations among the Cognitive Test Scores



Analysis

We analyzed our dataset in two steps: (a) hidden Markov modeling and (b) regression modeling. In step one, we first identified the number of skill acquisition stages most consistent with our dataset. Then, in step two, we identified the nature of these stages by investigating which cognitive variables, declarative and/or procedural learning ability, predicted the accuracy and speed of performance in each learning stage.

Hidden Markov Modeling

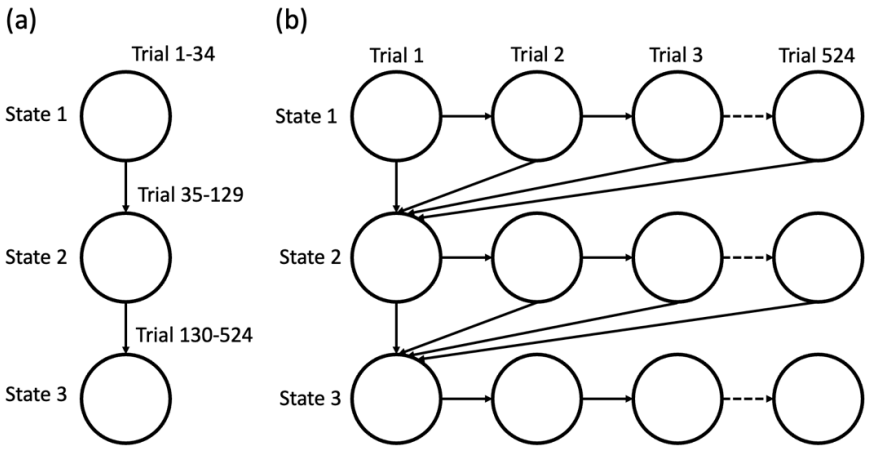
We adopted the hidden Markov modeling analysis to address Research Question 1, which aimed to identify the most probable number of skill acquisition stages consistent with participants' RT data (Tenison & Anderson, 2016). Although the use of coefficient of variation (CV) of RT has been prevalent in SLA research, our trial-level analysis of skill development using hidden Markov modeling required a dependent variable that was available at the trial level. This requirement excluded CV, which was calculated based on the participant mean and standard deviation of RT at the block level. For interested readers, we present our analysis of CV in Appendix S3.

A hidden Markov model is a stochastic system consisting of a series of states, called Markov states, which are unobservable (hidden), but their probability can be estimated (modeled) using observed data. In this study, Markov states represented skill acquisition stages. We used each participant's history of RT over 528 trials, taken from the warmup (16 trials in Day 2) and main practice tasks (512 trials in Days 3–6), as the dependent variable for modeling purposes. Here, we use the term *state* to refer to the hypothesized learning stage represented by the computational model and *stage* to refer to the actual learning stage learners were assumed to be in. Although readers may view the two terms as equivalent, we distinguish between them because (a) we wanted to align our terminology with that of Tenison and Anderson (2016); and more importantly, (b) our HMMs were unlike typical HMMs and hence the number of hidden Markov states did not exactly correspond to the number of stages. Figure 6 illustrates the difference between a typical HMM (Figure 6a) and the model we adopted following Tenison and Anderson (2016) (Figure 6b). Since the transition between stages requires multiple times of practice, one may typically allow learners (or their RT data) to remain in the same state for some number of trials (Figure 6a). However, such a model design would violate the fundamental assumption of a Markov state positing that behaviors (or probability thereof) in a future only depend on the current state. In our model, each practice

trial was hence associated with all hypothesized states (Figure 6b) and we estimated the probability for all states that were consistent with RT data (see Appendix S5 for more explanation). This meant that when we hypothesized k stages, we had $528k$ stages in the model. For interested readers, we share full mathematical details of our hidden Markov modeling in Online Supporting Materials.

Figure 6

Comparison of (a) Typical Hidden Markov Model and (b) the Current Model



Following Tenison and Anderson (2016), we fitted a series of HMMs to the participants' RT data and estimated the probability of each participant residing in a given state on each practice trial. We fitted three HMMs that were informed by the three theoretical models of skill acquisition (i.e., the race model, the CMPL theory, and ACT-R) and each assumed one, two, or three states, respectively. We compared the models by examining how well they fit the data based on its associated Akaike Information Criterion corrected for small sample sizes (AICc) and Bayesian Information Criterion (BIC) (Wagenmakers & Farrell, 2004). AICc and BIC were defined as:

$$AICc = -2logL + 2k + \frac{2k(k+1)}{n-k-1}$$
$$BIC = -2logL + klog(n)$$

where log$L$ is the log likelihood of obtaining the observed data under the model, k is the number of parameters in the model, and n is the sample size. Because examining the values of AICc or BIC per se does not indicate how well the best model compares to its rival models, we also calculated the so-called Akaike weight and the BIC model weight. These weights can be interpreted as the conditional probability of a model when compared to the other candidate models in the set (with the probability value bound between 0 and 1). We followed the formulation of the weights summarized in Wagenmakers and Farrell (2004):

$$w_i(\text{index}) = \frac{exp\left\{-\frac{1}{2}\Delta_i(\text{index})\right\}}{\sum_{k=1}^{K} exp\left\{-\frac{1}{2}\Delta_k(\text{index})\right\}}$$

where Δ(index) is the difference between AICc or BIC of the best model and that of a model in focus. The primary purpose of using both AICc and BIC was to gather and triangulate multiple sources of information. All parameters associated with the HMMs were estimated using the BFG algorithm (Bacri et al., 2023). All parameters started from initial neutral values and the algorithm iteratively searched for the best values for optimization. The HMM fitting was done on Google Colaboratory, an online platform to program and execute the Python language. We publicly share our code for hidden Markov modeling at:

https://colab.research.google.com/drive/1vwqR4dLYylRA-nhydEr14IGzWma_WONQ?usp=sharing.

Regression Modeling

The regression analysis addressed Research Question 2, which sought to explore the nature of learning stages by examining which cognitive variables predicted participants' performances in each of the learning stages identified in the HMM analysis. We modeled participants' response accuracy and speed (RT) of performance using generalized linear mixed models. The accuracy model was a binomial model, which took accuracy (0 or 1) as the dependent variable and in which

we regressed the probability of correct responses on the predictor variables. We used the logit-link function to map the probability (0−1) to the logit of probability (−∞ to +∞) and to model the linear relationship between the dependent and predictor variables. The RT model was a normal (or Gaussian) model, which assumed normality of the dependent variable. Note that we log-transformed RT because RT data by nature produce positively skewed distributions.

Predictor variables included practice trials (1-528), Stage 2 (dummy coded as 0 or 1), Stage 3 (dummy coded as 0 or 1), Declarative (factor scores), ASRT (z-score), and SL (z-score). We also let Stage 2 and Stage 3 interact with the three cognitive variables to model the potentially differential effects of the cognitive variables in Stage 2 and Stage 3, respectively, relative to Stage 1 (baseline). Note that practice trials were log-transformed in the RT model to model the linear relationship between RT and practice trials. For random effects, we let the intercept vary among participants. We also modeled random slopes of practice trials that can vary among participants.

We estimated the model parameters using Bayesian inference. In Bayesian analysis, prior knowledge in the form of probability distributions is combined with data to generate the posterior distribution of model parameters (Gelman et al., 2013). For prior distributions, we chose weakly informative priors to strike a balance between incorporating general knowledge about our problem of interest (e.g., the general range of RTs in our dataset) and allowing the data to drive the result. We used the R-package brms (version 2.19.0, Büerkner, 2017) to estimate the posterior distributions through the Markov chain Monte Carlo (MCMC) simulation, consisting of four MCMC chains of 10,000 iterations each, with the first 1,000 iterations discarded as a warmup period. We monitored the value of $\hat{R}$ associated with each parameter to assess whether the MCMC simulation converged on a stable solution (Gelman & Rubin, 1992). We adopted the mean of the posterior distribution as the point estimate, and the highest posterior density interval as the interval

estimate of the parameter (i.e. 95% credible intervals [CrI]). Additionally, we computed the posterior probability of whether a given parameter value is larger or smaller than 0. The posterior probability directly indicates the probability of an effect being present; for instance, Pr(b>0) = .998 for Declarative (in Stage 1) means that the slope of declarative learning scores in Stage 1 is positive with a probability of 99.8%. We share our dataset and R-code publicly on the Open Science Foundation page at https://osf.io/x9u6h/.

Results

Descriptive Results

Figure 7 and 8, respectively, show the mean accuracy rates and RT across the entire 528 practice trials. For Figure 8, the blue line shows the average RT (over participants) and the black bars show the 95% confidence intervals of the means. Participants showed highly accurate performances (i.e., 90%) even from the first few trials, indicating that they had already developed solid declarative knowledge after the instruction (see Figure S5 in Online Supporting Materials for the accuracy of their performances on the vocabulary and grammar tests immediately after explicit instruction). They increased the speed of performance following the power-law of practice. At the trial level, regressing the logarithm of mean RT on the logarithm of practice trials yielded R^2 = .90. To compare the result with that of previous studies, we repeated the same analysis at the block level, which showed R^2 = .97 (cf. DeKeyser, 1997 with R^2 = .97 and Pili-Moss et al., 2020 with R^2 = .92). Finally, the flattening of the curve in both figures indicated that participants reached asymptotic levels of accuracy and speed of performance.

Figure 7

Participants' Mean Accuracy Rates across Practice Trials

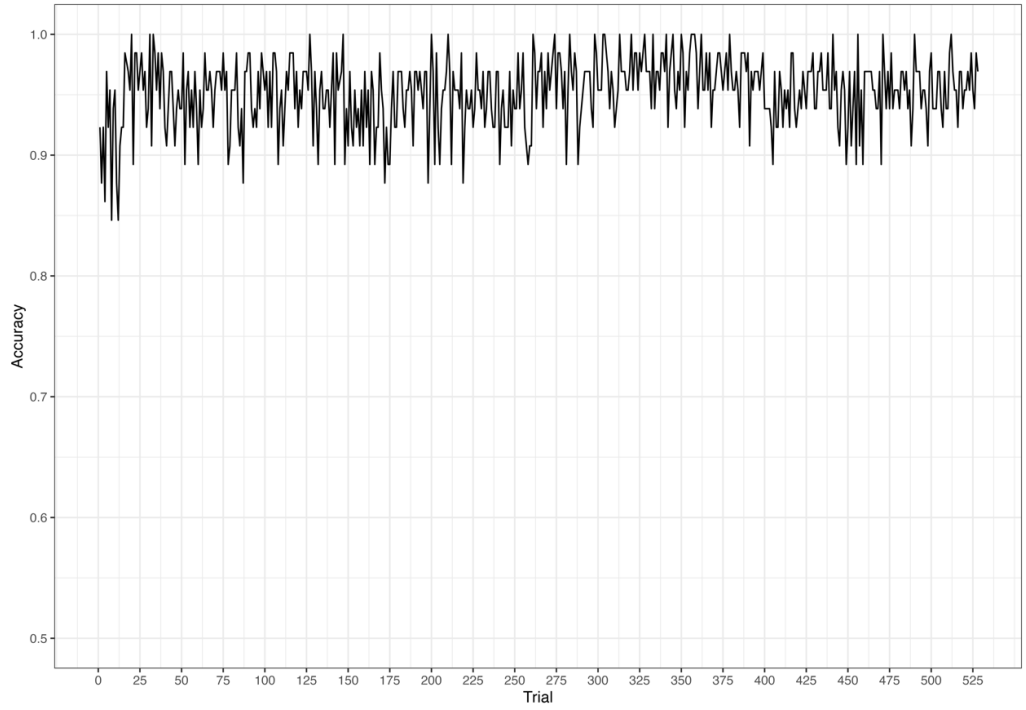
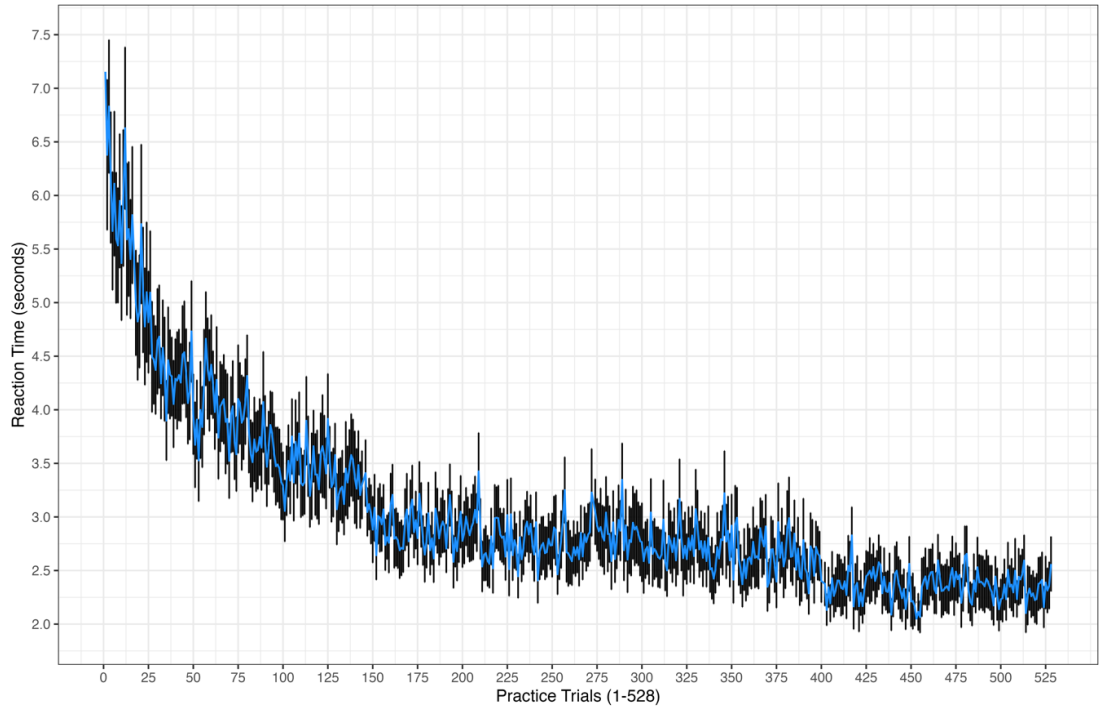


Figure 8

Participants' Mean Reaction Time across Practice Trials



Note. Blue line = mean RT over participants and black bars = 95% CIs.

Hidden Markov Modeling

The results of the hidden Markov modeling are shown in Table 2. The table displays the log-likelihood (log(L)) of the models, BIC and its weights (*w*(BIC)), as well as AICc and its weights (*w*(AICc)). BIC and AICc indicate the deviation of the models from the RT data (and thus lower values indicate better fit), while their weights show the relative likelihood of the models compared to the other models. Model comparisons based on BIC and AICc indicated a clear advantage of the three-state HMM, suggesting that participants indeed acquired L2 skills in three stages. The model weights for the three-state HMM were almost identical to 1, indicating the model's overwhelming advantage over the other two models.

Table 2

The Results of Hidden Markov Modeling

	log(*L*)	BIC	*w*(BIC)	AICc	*w*(AIC)
One-state	-45369	90768	.000	90741	.000
Two-state	-27502	55044	.000	55008	.000
Three-state	-27436	54921	≈ 1.000	54876	≈ 1.000

Note. log(*L*) = log likelihood, *w*(BIC) = BIC model weight, and *w*(AICc) = AICc weight.

One advantage of the HMM analysis is that it estimates the probability of individual participants residing in each stage after each practice trial. Based on the three-state HMM, we therefore estimated the participants' state occupancy by assuming that they resided in the state that had the highest probability. Figure 9 summarizes the average state probability across the entire sample. All participants started from Stage 1 (a constraint of the HMMs), but they eventually moved to Stage 2 and most participants also moved to Stage 3 with practice over time. Stage 2 became the majority state at Trial 19, and Stage 3 became the majority state at Trial 183. Figure 10 presents the individual trajectories of stage occupancy by participant. 16.9% of the participants showed wiggly (highly variable) ends of the power-function curve in RT (see individual curves in Appendix S3) and did not reach the final stage.

Figure 9

Participants' Average State Probability across Practice Trials

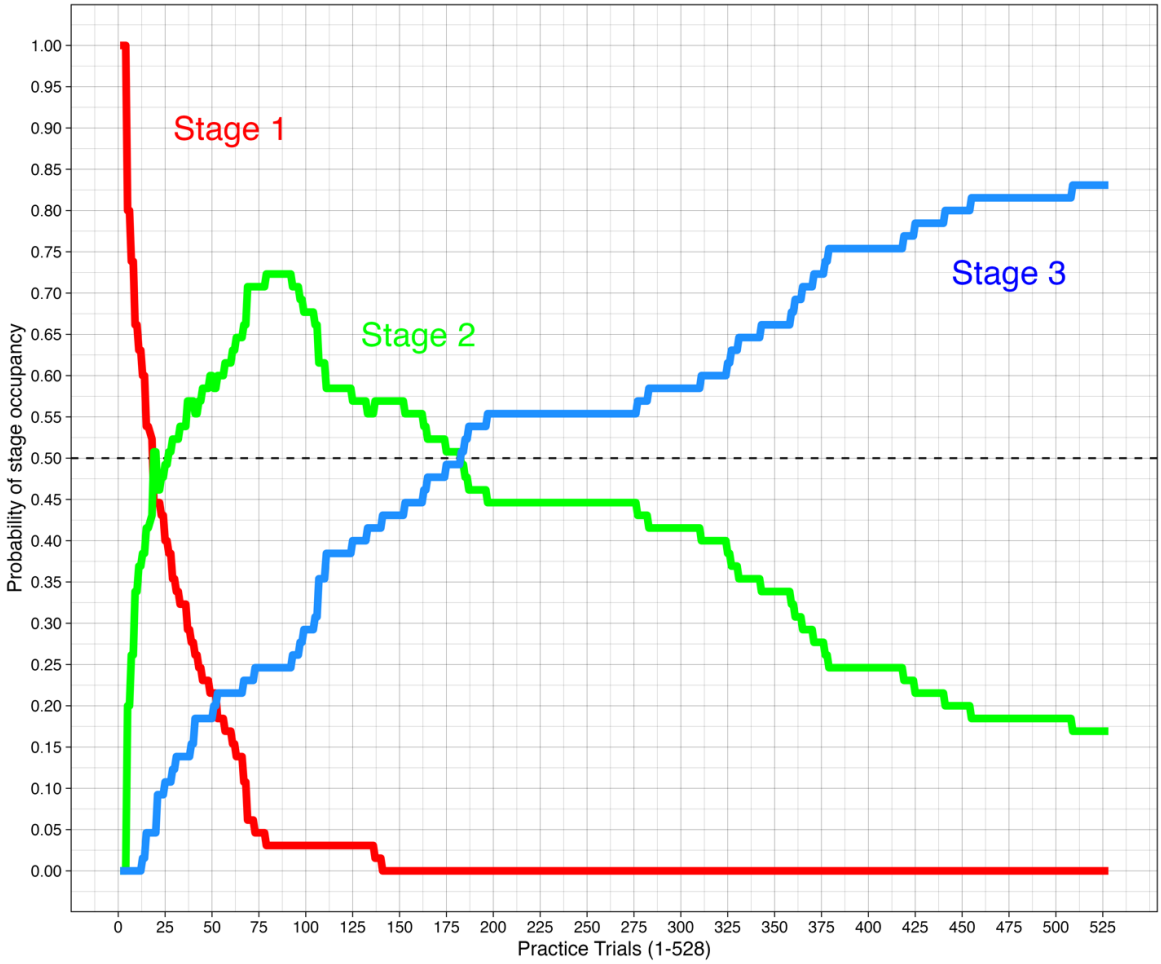
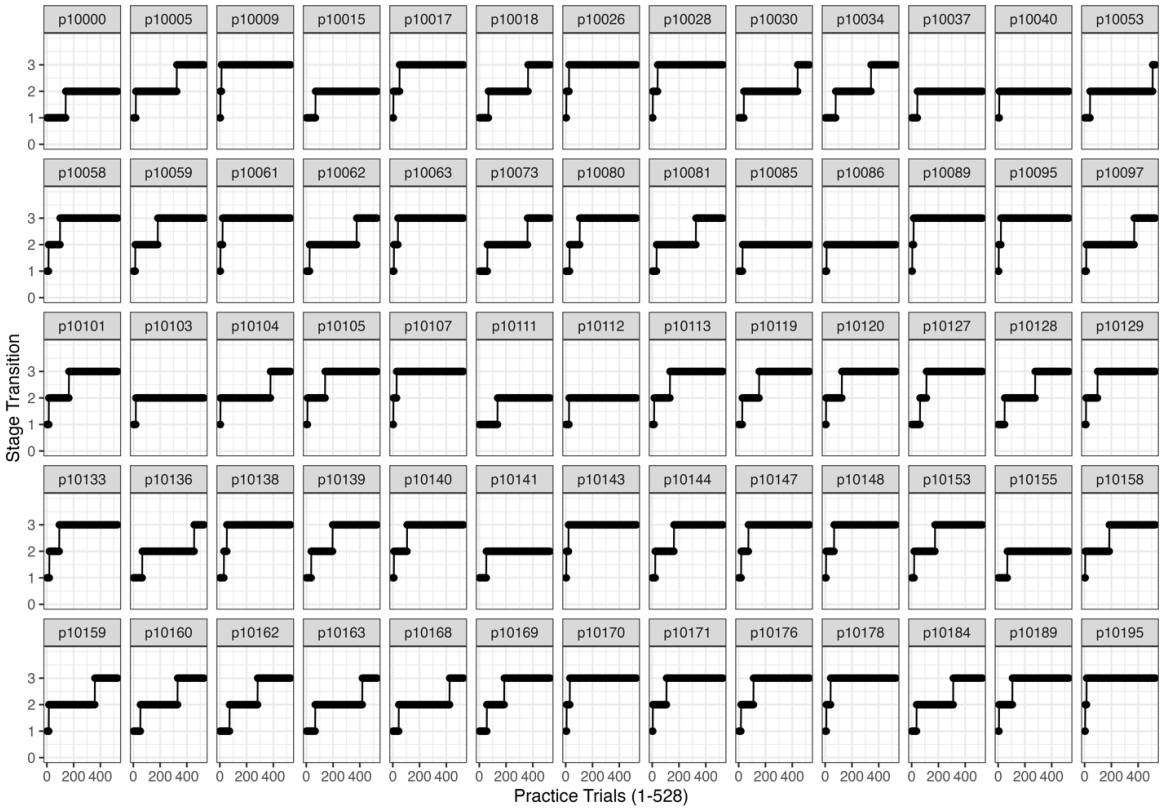


Figure 10

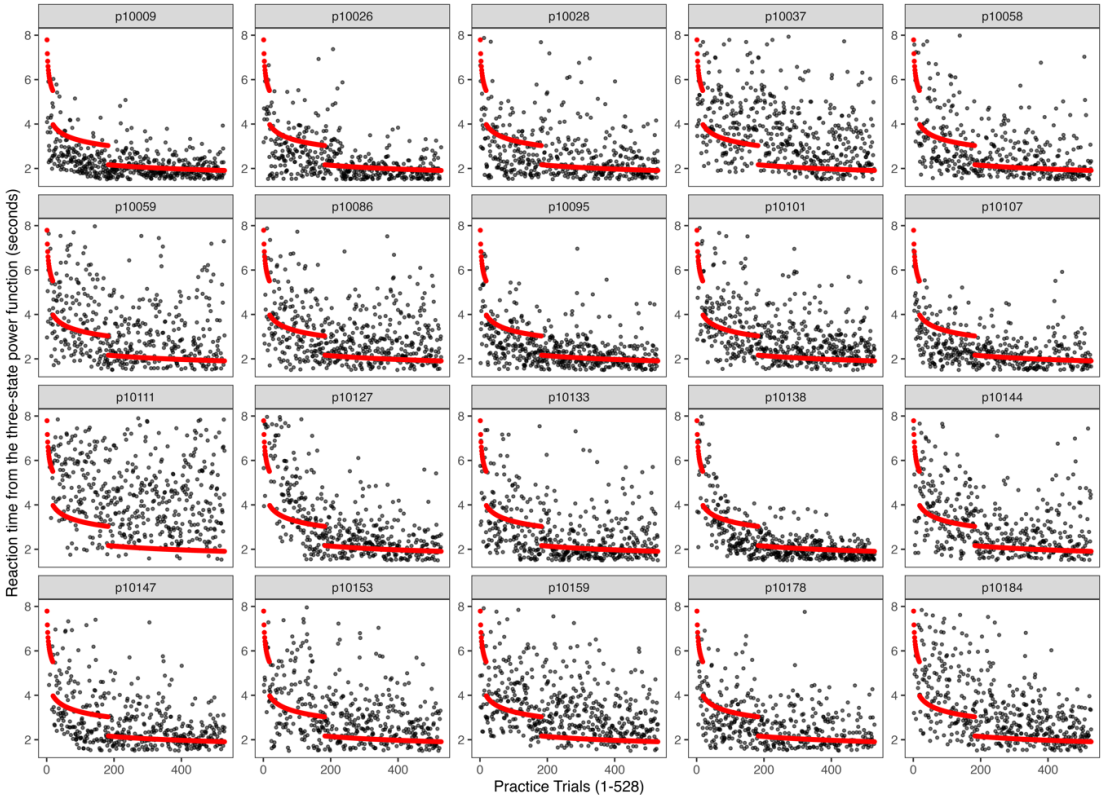
Individual Participants' State Transitions across Practice Trials



The three-state HMM additionally estimated the predicted values of RT based on a power function (see Online Supporting Materials for mathematical details). Figure 11 shows the predicted values of RT at the group level (red line) overlaid on participants' raw data points. Note that we randomly selected a sample of 20 participants for visualization purposes, but Figure S12 in Online Supporting Materials shows the same data for all 65 participants. Overall, participants decreased their RT from approximately 8 seconds (s) to 6s within Stage 1. The transition from Stage 1 to Stage 2 meant an additional decrease of 2s (from 6s to 4s). In Stage 2, participants decreased their response time from 4s to 3s, and the following transition from Stage 2 to Stage 3 represented an additional 1s decrease, after which almost no further improvement was expected.

Figure 11

Predicted Values of RT (Red Line) Overlaid on Participants' Raw Data Points



Regression Modeling

Here, we report the results of the regression analysis to answer Research Question 2, asking which cognitive capacity, declarative and/or procedural learning ability, predicted the participants' accuracy and speed (RT) of performance in each stage identified through hidden Markov modeling. From the three-state HMM, we extracted the probability of each participant residing in stage one, two, or three on a given practice trial. We assumed that on a given trial, participants occupied the stage that had the highest probability associated with it in the model (see Figure 10). We estimated model-based predicted values of accuracy and RT as a function of learning stages (Stage 1–3), practice trials (1–528), and the cognitive variables (declarative learning scores, ASRT, and SL).

See Figures 12 and 13. We computed the posterior probability of whether the effect of a given variable was larger or smaller than 0 (i.e., Pr(b > or < 0)). This probability encoded the likelihood of a given effect being present and the direction of the effect. For interested readers, Tables S3 and S4 show the entire list of the parameter estimates for the accuracy and RT models.

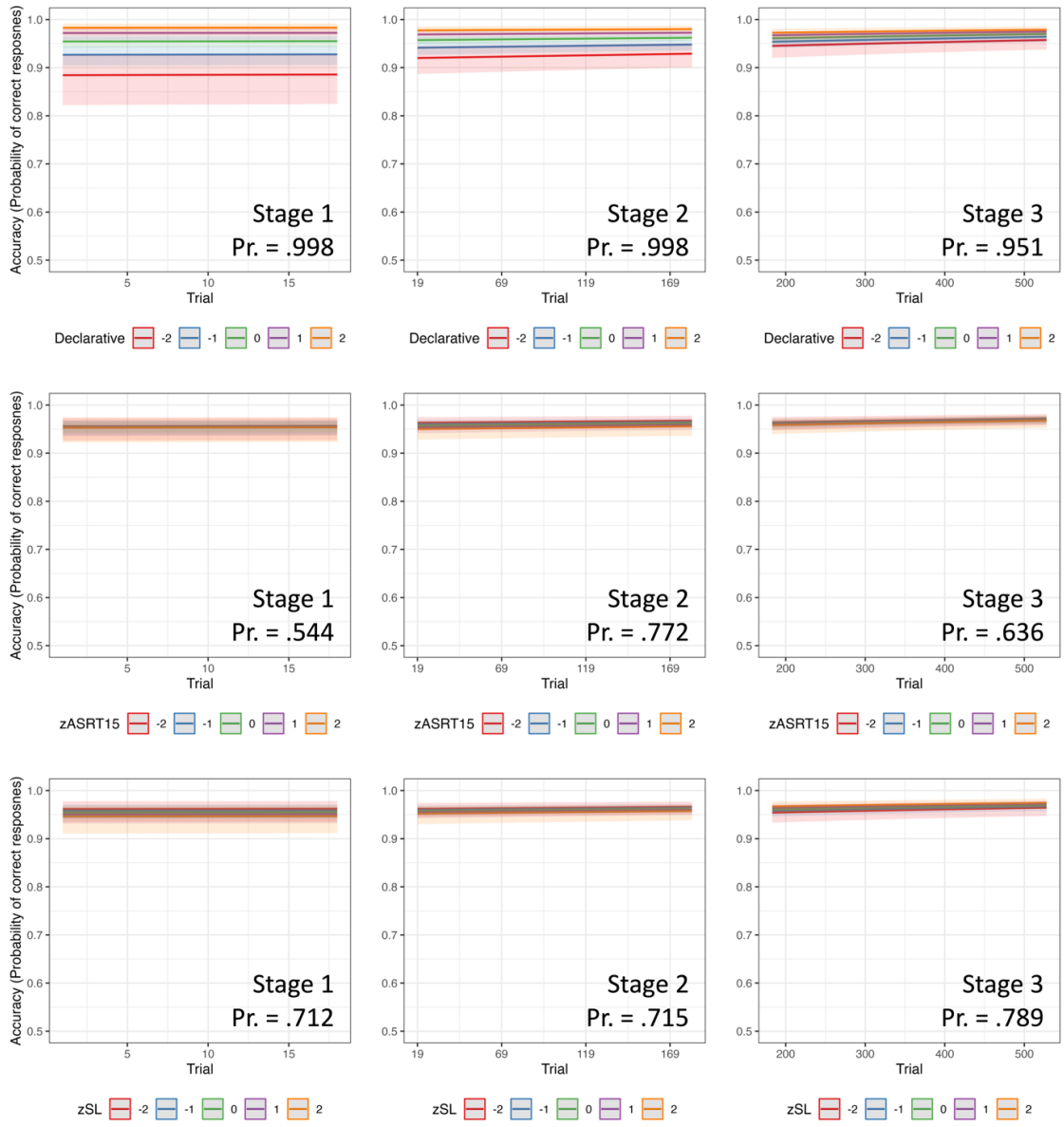
Accuracy

Figure 12 shows the predicted values of accuracy (y-axis) based on practice trials (x-axis), learning stages (panels from left to right), and the three cognitive variables (panels from top to bottom). We plotted different ability levels in the cognitive variables as separate lines. Specifically, each line in Figure 12 represents a 1SD increase in ability, from –2SD, –1SD, mean, and +1SD, to +2SD. The only variable that was reliably associated with participants' accuracy of performance was declarative learning scores, which was positively related with accuracy in all three stages, with a very high (posterior) probability of .998 (Stage 1), .998 (Stage 2), and .951 (Stage 3). This result indicated that participants with higher declarative learning ability provided more accurate responses than those with lower declarative learning scores. The comparison of the declarative learning effects among the three stages further showed that the magnitude of the effect decreased following the transition from Stage 1 to Stage 2 and from Stage 2 to Stage 3. The difference in accuracy rates between individuals with the highest and lowest declarative learning abilities, respectively (i.e., at the point of +2SD and –2SD away from the mean) was 8.21%, (95% CrI [4.71, 13.25]) in Stage 1, 4.81% (95% CrI [3.21, 7.02]) in Stage 2, and 2.57% (95% CrI [1.62, 4.13]) in Stage 3. When comparing the slope of declarative learning between stages, the posterior probability of the difference was .878 for Stage 1–Stage 2, .977 for Stage 1–Stage 3, and .949 for Stage 2–Stage 3. This decreasing impact is also visible in Figure 12 (top row): the differences between the

lines become narrower following the transition from Stage 1 to Stage 2 and from Stage 2 to Stage 3.

Figure 12

Predicted Values of Accuracy Based on Practice Trials, Learning Stages, and the Three Cognitive Variables



Reaction Time

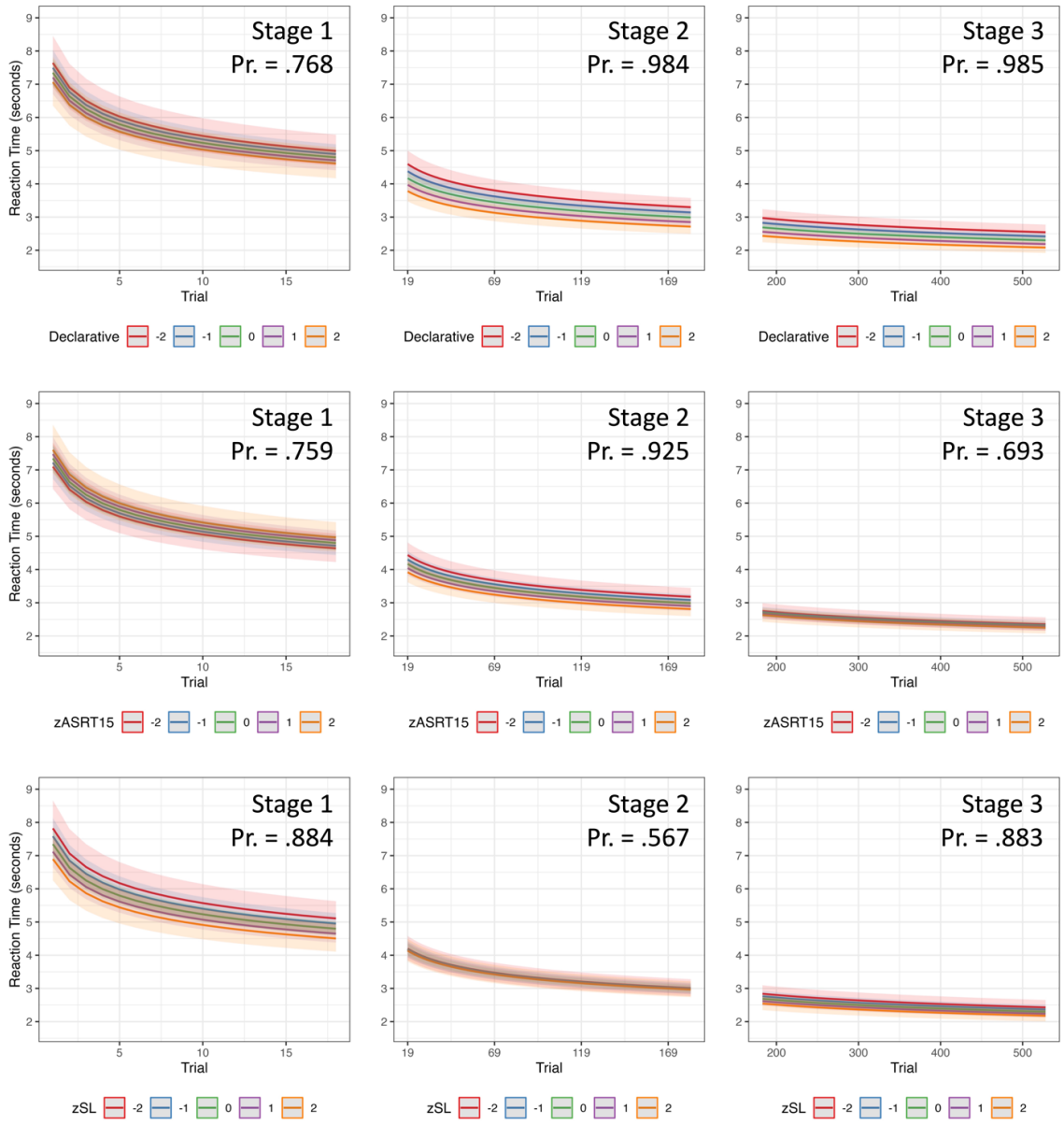
Figure 13 displays the predicted values of RT (y-axis) based on practice trials (x-axis), learning stages (panels from left to right), and the three cognitive variables (panels from top to bottom). Declarative learning scores reliably predicted RT in Stage 2 (Pr(b<0) = .984) and in Stage 3 (Pr(b<0) = .985). This result indicated that those participants who had higher declarative learning capacity performed faster than those with lower declarative learning scores in Stage 2 and Stage 3. Note that the posterior probability of the difference (or whether the difference is larger than zero) between the effects of declarative scores in Stage 2 and Stage 3 was .543, indicating that declarative learning was as important in Stage 3 as in Stage 2 for maintaining the speed of performance. The difference in RT between individuals with the highest and lowest declarative learning scores (at +2SD and –2SD away from the mean) was 0.55 (95% CrI [0.52, 0.60]) and 0.51 (95% CrI [0.44, 0.56]) seconds in Stage 2 and Stage 3, respectively.

Lastly, the results in Figure 13 showed that procedural learning capacity also predicted RT, but ASRT and SL showed different patterns of results. ASRT predicted RT in Stage 2 (Pr(b<0) = .925), showing that higher procedural learning (measured by ASRT) coincided with faster performances in Stage 2. The difference in RT for individuals with the highest and lowest ASRT scores (at +2SD and –2SD away from the mean) (with higher scores showing higher capacity) was 0.35 seconds in Stage 2. Similarly, SL scores were predictive of RT such that higher SL scores were associated with faster performances, but the effect was present in Stage 1 (Pr(b<0) = .884) and in Stage 3 (Pr(b<0) = .883). The difference in RT between individuals with the highest and lowest SL scores (at +2SD and –2SD points away from the mean) was 0.40 and 0.27 seconds in Stage 1 and Stage 3, respectively.

Comparing the effects of declarative and procedural learning scores, the effect of declarative learning seemed larger than that of ASRT scores (the difference in RTs at $\pm$2SD = 0.55 seconds for Declarative versus 0.35 seconds for ASRT) in Stage 2, whereas the effect of SL seemed larger than that of declarative learning in Stage 3 (0.51 for SL versus 0.27 seconds for Declarative). However, the statistical evidence for these differences was not clear. The posterior probability of the two effects being different was .715 (Declarative versus ASRT in Stage 2) and .736 (Declarative versus SL in Stage 3).

Figure 13

Predicted Values of Reaction Time Based on Practice Trials, Learning Stages, and the Three Cognitive Variables



Summary

Our results from the hidden Markov modeling showed that participants acquired comprehension skills in Mini-Nihongo in three stages (Research Question 1). Our regression analysis further revealed that participants' individual differences in declarative and procedural learning differentially contributed to the accuracy and speed of performance in the three stages (Research Question 2). In terms of participants' performance accuracy, declarative learning ability emerged as the sole predictor, playing a crucial role in all three stages. However, the effect of declarative learning was particularly prominent in Stage 1, followed by Stage 2 and then Stage 3, in that order. For RT, both declarative and procedural learning abilities were predictive of the speed of performance. When it came to procedural learning, ASRT was a predictor in Stage 2, whereas SL was influential in Stage 1 and 3. In both cases, better procedural learning ability coincided with faster performance. Although our findings did not match our hypotheses exactly (see The Current Study), these results are consistent with the declarative-procedural transition proposed in the three-stage model.

Discussion

We tested the three-stage model of L2 skill acquisition by investigating the number and nature of developmental stages in L2 learning. We hypothesized that participants would learn to comprehend the miniature Japanese language in three stages and these stages would follow the declarative-procedural transition characteristic of the three-stage model. Our findings largely confirm these hypotheses and hence corroborate the validity of the three-stage model. However, it is important to examine whether the three-stage model and its most prominent example, ACT-R, can account for specific details in our results.

The Number of Learning Stages

Our first research question addressed the number of learning stages in L2 skill acquisition. We showed through our hidden Markov modeling (Table 2) that participants indeed acquired comprehension skills in Mini-Nihongo following three stages. This finding aligns with the learning mechanism of the three-stage model, or ACT-R (DeKeyser, 2020; Suzuki, 2022), and is at odds with that of the race model (a single-stage model) and the CMPL theory (a two-stage model). Focusing on the mean number of practice trials necessary for someone to reach the second and the third stage (Figure 9), we found that our participants required 18 trials on average to move to Stage 2 and an additional 165 trials to reach Stage 3. Assuming that the Stage 1–2 transition corresponds to proceduralization (see the next section for our discussion of evidence), this number (18 trials) resembles the estimate proposed by DeKeyser (1997), who reported in his experiment that participants required only 16 trials to proceduralize the target morphosyntactic rules. We further extended this finding by showing that after proceduralization, our participants on average took an additional 165 trials to reach the third, final stage. Without further research, it is difficult to assess the degree to which our findings generalize to other learning scenarios involving other linguistic features, learning conditions, and learner samples. However, the similarity between our results and those reported in DeKeyser (1997) is striking.

Our analysis of the predicted RT from the three-state HMM (Figure 11) further revealed that participants decreased their performance times both within and between stages at comparable rates. A key assumption in the ACT-R learning mechanism is that most performance improvements occur through stage transitions or qualitative shifts in the underlying cognitive mechanism (Anderson, 2005; Tenison & Anderson, 2016). However, our results did not support this pattern. For example, participants reduced their RT from 8s to 6s within Stage 1, while the transition from Stage 1 to Stage 2 resulted in a similar decrease of 2s (from 6s to 4s). Our results

therefore suggest that both quantitative (within-stage) and qualitative (between-stage) changes in cognitive mechanisms contribute equally, or nearly so, to L2 skill acquisition, which contrasts with previous findings on cognitive skill acquisition (Tenison & Anderson, 2016).

The Nature of Learning Stages

Our second research question asked about the nature of learning stages. More specifically, we investigated which cognitive abilities, declarative and/or procedural learning ability, predicted the accuracy and speed of performance in each learning stage identified in the hidden Markov modeling analysis. Following skill acquisition theory, we hypothesized that declarative learning would predict accuracy in Stage 1, whereas procedural learning would predict RT in Stage 2. Here, we interpret our results vis-à-vis the learning mechanisms proposed in ACT-R.

Accuracy

We found a steady but decreasing impact of declarative learning on maintaining accuracy throughout the three stages of L2 skill acquisition. This finding, although deviating from our prediction, replicated the finding by Pili-Moss et al. (2020), who showed that declarative learning ability is important for maintaining accuracy throughout the entire L2 learning process. Nevertheless, our finding also does not contradict recent implementations of ACT-R, which allow for the involvement of declarative memory even after proceduralization has taken place (Anderson, 2005, 2007). In fact, ACT-R includes a mechanism of *declarative strengthening* at work in the later two stages, which functions to increase the accuracy and speed of retrieving declarative knowledge (Anderson, 1982, 1983). The existence of declarative strengthening implies that learners can rely on declarative memory even after proceduralization, enabling them to fall back on a more general (but less efficient) mode of skill performance when in need. The decreasing impact of declarative

memory thus makes sense because in the latter two stages learners could rely on declarative memory as a fail-safe.[3]

Additionally, complex cognitive tasks such as language processing almost always involve the use of declarative knowledge. In L2 research, Ellis (1994) highlighted a dissociation whereby the retrieval of vocabulary-form knowledge, for instance, is procedural in nature and hence can be learned implicitly, but the meaning aspects are declarative in nature and require explicit processes. This view is consistent with the ACT-R cognitive mechanisms whereby some aspects of performance routines can be proceduralized, but others always remain declarative in nature (Anderson, 2005, 2007). Suppose learning how to comprehend the subject of a sentence in Mini-Nihongo (e.g., *neko-ga* [cat-nominative]) as an example. Initially, learners must retrieve declarative knowledge of (a) the form-meaning mapping of the word *neko* and (b) the rule that *-ga* is a case marker (nominative) to be appended to the noun. At this stage, learners slowly work to translate these two pieces of declarative knowledge into the target behavior. However, through practice, learners develop procedural knowledge that combines *neko* and *-ga* into one chunk so that these words can now be readily called up together in the context of the sentential subject. Nevertheless, the meaning of the chunk itself is declarative in nature and must therefore always be retrieved from declarative memory. Hence, proceduralization applies to the retrieval of the knowledge (*neko* + *ga* → *neko ga*, which is a behavior) but not to the knowledge itself (*neko ga, "the cat*$_{SUBJECT}$*"*). Our findings support this view and show that the declarative-procedural transition does not make one's performance dependent solely on one system (also see endnote 2 for a note on idealization in skill acquisition theory).

Reaction Time

For speed of performance (RT), we hypothesized that participants' individual differences in procedural learning would predict RT in the second stage. We found that ASRT was influential in the second stage, suggesting that proceduralization is taking place, but SL emerged as a predictor in the first and third stages. One explanation for the differential roles of ASRT and SL may be related to the underlying constructs that these tasks were designed to measure. ASRT assesses one's ability to implicitly learn and routinize non-adjacent categorical dependencies in non-verbal stimuli through repetitive motor movements, while SL assesses both non-adjacent categorical and adjacent probabilistic relationships through passive exposure to verbal stimuli. It may be that the process of proceduralization draws upon abilities that are captured better by ASRT than SL. Regardless of the specific nature of our procedural learning tasks, the findings point to the emerging role of procedural memory in later stages of L2 skill acquisition, which is consistent with skill acquisition theory, the D/P model, as well as the findings of previous research on the role of declarative and procedural memory in L2 learning (e.g., Hamrick, 2015; Morgan-Short et al., 2014; Pili-Moss et al., 2020).

Discussions for Future Research

An immediate logical extension of our study is to replace behavioral measures of declarative and procedural learning abilities by neural measures such as fMRI to obtain more direct views of the mechanism of L2 skill acquisition within the declarative/procedural paradigm (Morgan-Short et al., 2015). The original study by Tenison and Anderson (2016) took this approach. More importantly, the use of neural measures would enable studying skill acquisition not only at the level of learning mechanisms, which may change over multiple practice trials (as we observed in this study), but also at the level of a single trial, by offering a window into the cognitive processing stages that may take place during thattrial. As a case in point, Tenison et al.

(2016) examined how changes in learning stages in mathematical problem-solving (e.g., solving 5$3 as 5 + 4 + 3 = 12) impacted stages of cognitive processing happening within a single trial (problem encoding, problem solving, and producing a response). Triangulating hidden Markov modeling with fMRI data, they found that transitioning from the cognitive to procedural stage significantly reduced the duration of the problem solving stage due to proceduralization, and transitioning to the automatic stage further shortened the encoding stage because participants at this level were able to perform the task almost as a reflex. We believe that research with the same granularity of analysis is necessary to further delve into the process of skill acquisition in L2 learning.

It is worth recalling that ACT theory as cited in SLA research originated from cognitive psychology over 40 years ago (Anderson, 1982, 1983). Since then, SLA research has advanced sufficiently as a scientific discipline so skill acquisition theory can and, we would argue, should be further developed and specified in relation to L2 learning. This does not mean that SLA should disregard insights from cognitive psychology. Rather, SLA researchers should assess the applicability of any imported cognitive theories in L2 learning (as we did in this study) and extend them if needed. L2 learning is undoubtedly more complex than typical cognitive skill acquisition studied in psychology experiments, because successful learning involves mastering various levels of linguistic knowledge (e.g., vocabulary, morphosyntax, and pragmatics), which need to be coordinated simultaneously during performance. There are many theoretical models that attempt to explain the process of L2 acquisition. Future researchers could fruitfully attempt to integrate these models with a skill acquisition view of L2 learning. For instance, following Levelt's (1989) model of speech production (updated by Kormos, 2006 to suit L2 speech), how would different stages of speech production (conceptualizing, formulating, and articulating) change as learners go

through the three stages of skill acquisition? And how would reading comprehension change with increased mastery, for instance, if one adopts a model of reading proposed by Khalifa and Weir (2009) with eight stages of cognitive processes underlying reading? Drawing on existing models, but seeing how component processes might change (qualitatively or quantitatively) as a result of practice, could be a fruitful way forward to enable the development of a skill acquisition theory or theories dedicated specifically to L2 learning.

In this study, we found three stages most suitable for characterizing the learning of Mini-Nihongo, but the learning primarily focused on learning morphosyntactic rules and comprehending the language based on those rules. We acknowledge that our findings may not generalize to other linguistic structures (e.g., collocations, pronunciation, or pragmatic routines) and modes of language use (e.g., production). For instance, item-based learning models such as the race model and the CMPL theory may better account for vocabulary learning, which often does not involve rule learning. Additionally, learners can draw on multiple processes for learning the target language–explicit and deductive, as we examined in this experiment, but also incidental, meaning as a by-product of meaningful exposure to the target language. We would not necessarily expect that participants would similarly go through three stages if they learned the target language inductively, without provision of explicit instruction. These varied learning targets and conditions may guide future research to properly assess the scope of the three-stage model and any possible constraints on its generalizability in L2 learning.

Conclusion

In this study, we reported on the first empirical test of the three-stage model of L2 skill acquisition, derived from cognitive psychology. We accomplished this by (a) identifying the number of stages learners progress through while learning and practicing a new foreign language,

and (b) investigating whether participants' individual differences in declarative and/or procedural learning abilities would predict their performances in each stage. We combined hidden Markov modeling analysis (addressing the number of stages) with regression analysis (addressing the nature of stages) to answer our research questions. Overall, our findings lend support for the three-stage model of L2 skill acquisition, which consists of a declarative, a procedural, and an automatic stage in L2 learning. However, specific details in our results warrant further empirical testing in future studies. Additionally, work in this area will benefit from further development of skill acquisition theory itself beyond its cognitive psychological origins to reflect the sheer complexity of L2 learning. But overall, our results showed that skill acquisition theory provides a solid foundation for understanding L2 acquisition. We invite other researchers to assess the generalizability of our results and help delineate the full scope of the model.

Endnote

1. Note that the word "production" here is not related to a typical meaning in L2 research (the process of producing a language) but rather a specific terminology in computer science.

2. Although it may seem that the D/P model and ACT-R deal with cognitive phenomena at different levels of analysis (memory vs. knowledge, respectively), a core assumption of the D/P model (and of ACT-R) is that declarative and procedural knowledge have their corresponding neurobiological origins in declarative and procedural memory. Moreover, more recent implementations of ACT-R are neurally based and integrate accumulating evidence from the neurocognition of human learning, including how declarative and procedural memory are employed for skill acquisition (Anderson, 2005, 2007).

3. An alternative explanation is that one's declarative memory capacity in general is correlated with one's capacity to maintain attention on the task (i.e., reflecting attentional mechanisms in working memory), and the decreasing role of declarative memory simply reflects the decreasing variability in participants' accuracy as they continue to fine-tune their performances through practice.

References

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, *89*(4), 369–406. http://dx.doi.org/10.1037/0033-295X.89.4.369

Anderson, J. R. (1983). *The architecture of cognition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Anderson, J. R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science*, *29*(3), 313–341. https://doi.org/10.1207/s15516709cog0000_22

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York, NY: Oxford University Press.

Bacri, T., Berentsen, G. D., Bulla, J., Støve, B. (2023). Computational issues in parameter estimation for hiddenMarkov models with template model builder. *Journal of Statistical Computation and Simulation*, *93*(18), 3421–3457. https://doi.org/10.1080/00949655.2023.2226788

Bajic, D., & Rickard, T. C. (2011). Toward a generalized theory of the shift to retrieval in cognitive skill learning. *Memory & Cognition*, *39*, 1147–1161. https://doi.org/10.3758/s13421-011-0114-z

Buffington, J., Demos, A. P., Morgan-Short, K. (2021). The reliability and validity of procedural memory assessments used in second language acquisition research. *Studies in Second Language Acquisition*, *43*(3), 635–662. https://doi.org/10.1017/S0272263121000127

Buffington, J., & Morgan-Short, K. (2019). Declarative and procedural memory as individual differences in second language aptitude. In Z. Wen, P. Skehan, A. Biedroń, S. Li, & R. L. Sparks (Eds.), *Language aptitude: Advancing theory, testing, research and practice* (pp. 215–237). New York, NY: Routledge.

Büerkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. http://dx.doi.org/10.18637/jss.v080.i01

DeKeyser, R. M. (1997). Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition*, *19*(2), 195–221. https://doi.org/10.1017/S0272263197002040

DeKeyser, R. M. (2001). Automaticity and automatization. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 125–151). New York, NY: Cambridge University Press.

DeKeyser, R. M. (2020). Skill acquisition theory. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition. An introduction* (3rd ed., pp. 83–104). New York, NY: Routledge.

Ellis, N. C. (1994). Vocabulary acquisition: The implicit ins and outs of explicit cognitive mediation. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 211–282). London, UK: Academic Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–511. https://doi.org/10.1214/ss/1177011136

Godfroid, A., & Kim, K. M. (2021). The contribution of implicit-statistical learning aptitude to implicit second-language knowledge. *Studies in Second Language Acquisition*, *43*(s3), 606–634. https://doi.org/10.1017/S0272263121000085

Hamrick, P. (2015). Declarative and procedural memory as individual differences in incidental language learning. *Learning and Individual Differences*, *44*, 9–15. https://doi.org/10.1016/j.lindif.2015.10.003

Hamrick, P., Lum, J. A. G., & Ullman, M. T. (2018). Child first language and adult second language are both tied to general-purpose learning systems. *Proceedings of the National Academy of Sciences*, *115*(7), 1487–1492. https://doi.org/10.1073/pnas.1713975115

Howard, J. H., & Howard, D. V. (1997). Age differences in implicit learning of higher order dependencies in serial patterns. *Psychology and Aging*, *12*(4), 634–656. https://doi.org/10.1037/0882-7974.12.4.634

Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, NJ: Lawrence Erlbaum.

Khalifa, H., & Weir, C. (2009). *Examining reading: Research and practice in assessing second language reading*. New York, NY: Cambridge University Press.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: The MIT Press.

Li, S., & DeKeyser, R. (Eds.). (2021). Implicit language aptitude: Conceptualizing the construct, validating the measures, and examining the evidence [Special issue]. *Studies in Second Language Acquisition*, *43*(3).

Logan, G. D. (1988). Towards an instance theory of automatization. *Psychological Review*, *95*(4), 492–527. https://dx.doi.org/10.1037/0033-295X.95.4.492

Logan, G. D. (2002). An instance theory of attention and memory. *Psychological Review*, *109*(2), 376–400. https://doi.org/10.1037/0033-295x.109.2.376

Maie, R. (2022). *Testing the three-stage model of second language skill acquisition* (Publication No. 29998996) [Doctoral dissertation, Michigan State University]. ProQuest Dissertations and Theses Global.

Meara, P. M., & Rogers, V. E. (2019). *The LLAMA Tests v3. LLABA-B v3.2 beta*. Cardiff, UK: Lognostics.

Morgan-Short, K., Faretta-Stutenberg, M., Brill, K. A., Carpenter, H., & Wong, P. C. M. (2014). Declarative and procedural memory as individual differences in second language acquisition. *Bilingualism: Language and Cognition*, *17*(1), 56–72. https://doi.org/10.1017/S1366728912000715

Morgan-Short, K., Deng, Z., Brill-Schuetz, K. A., Faretta-Stutenberg, M., Wong, P. C. M., Wong, F. C. K. (2015). A view of the neural representation of second language syntax through artificial language learning under implicit contexts of exposure. *Studies in Second Language Acquisition*

Mueller, J. L. (2006). L2 in a nutshell: The investigation of second language processing in the miniature language model. *Language Learning*, *56*(s1), 235–270. https://doi.org/10.1111/j.1467-9922.2006.00363.x

Newell, A., & Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Mahwah, NJ: Lawrence Erlbaum Associates.

Oliveira, C. M., Henderson, L. M., & Hayiou-Thomas, M. E. (2023). Limited evidence of an association between language, literacy, and procedural learning in typical and atypical development: A meta-analysis. *Cognitive Science*, *47*(7). e13310. https://doi.org/10.1111/cogs.13310

Pili-Moss, D., Brill-Schuetz, K., Faretta-Stutenberg, M., & Morgan-Short, K. (2020). Contributions of declarative and procedural memory to accuracy and automatization

during second language practice. *Bilingualism: Language and Cognition*, *23*, 639–651. https://doi.org/10.1017/S1366728919000543

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General*, *126*(3), 288–311. http://dx.doi.org/10.1037/0096-3445.126.3.288

Robinson, P. (1997). Generalizability of second language learning under implicit, incidental, enhanced and instructed conditions. *Studies in Second Language Acquisition*, *19*(2), 223–247. https://doi.org/10.1017/S0272263197002052

Robinson, P. J., & Ha, M. A. (1993). Instance theory and second language rule learning under explicit conditions. *Studies in Second Language Acquisition*, *15*(4), 413–438. https://doi.org/10.1017/S0272263100012365

Romberg, A. R., & Saffran, J. R. (2013). All together now: Concurrent learning of multiple structures in an artificial language. *Cognitive Science*, *37*(7), 1290–1320. https://doi.org/10.1111/cogs.12050

Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences USA*, *93*(24), 13515–13522. https://doi.org/10.1073/pnas.93.24.13515

Suzuki, Y. (2022). Automatization and practice. In A. Godfroid & H. Hopp (Eds.), *The Routledge handbook of second language acquisition and psycholinguistics* (pp. 308–321). New York, NY: Routledge.

Tenison, C., & Anderson, J. R. (2016). Modeling the distinct phases of skill acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(5), 749–767. https://doi.org/10.1037/xlm0000204

Tenison, C., Fincham, J. M., & Anderson, J. R. (2016). Phases of learning: How skill acquisition impacts cognitive processing. *Cognitive Psychology*, *87*, 1–28. https://doi.org/10.1016/j.cogpsych.2016.03.001

Trahan, D. E., & Larrabee, G. J. (1988). *Continuous Visual Memory Test*. Odessa, FL: Assessment Resources.

Ullman, M. T. (2004). Contributions of memory circuits to language: the declarative/procedural model. *Cognition*, *92*(1-2), 231–270. https://doi.org/10.1016/j.cognition.2003.10.008

Ullman, M. T. (2020) The declarative/procedural model: A neurobiologically-motivated theory of first and second language. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition. An introduction* (3rd ed., pp. 83–104). New York, NY: Routledge.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*, 192–196. https://doi.org/10.3758/BF03206482