TESTING THE THREE-STAGE MODEL OF SECOND LANGUAGE SKILL ACQUISITION

By

Ryo Maie

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Second Language Studies—Doctor of Philosophy

2022

ABSTRACT

Skill acquisition theorists conceptualize second language (L2) learning as the acquisition of a set of perceptual, cognitive, and motor skills. The dominant view in skill acquisition theory is to regard L2 skill acquisition as a three-stage process “from initial representation of knowledge through initial changes in behavior to eventual fluent, spontaneous, largely effortless, and highly skilled behavior” (DeKeyser, 2020, p. 83). While there is indirect evidence that indicates the existence of such developmental stages, the number and the nature of those stages are often assumed a priori, and whether or not these stages actually exist remains untested. My dissertation study was designed to test and validate the three-stage model of L2 skill acquisition derived from cognitive psychological research, namely, the cognitive, associative, and autonomous stage (Fitts & Posner, 1967), each of which draws on distinct cognitive processes for learning.

Sixty-five adult learners deliberately learned and practiced a miniature language based on Japanese, called *Mini-Nihongo*, for a total of 1,056 practice trials. The participants also took a battery of tests on three dimensions of cognitive abilities that are known to be active at each stage of skill acquisition: declarative memory, procedural memory, and psychomotor ability (Ackerman, 1988, 1992; Anderson, 1982). Comprehension practice took place in the form of a sentence-picture matching task, and production practice was implemented in the form a productive maze task. Accuracy, reaction time (RT), and the coefficient of variability (CV) of RT were analyzed as the dependent variables. There were six tests of cognitive abilities: the Continuous Visual Memory Task and LLAMA-B for declarative memory ability, an alternating serial reaction time task and a statistical learning task for procedural memory ability, and the alternating serial reaction time task and a two-choice RT task for psychomotor ability.

I analyzed the data from the language practice and the battery of cognitive tests in two steps. First, I fitted a series of hidden Markov models (HMMs) to the RT data that represented different hypotheses regarding the number of skill acquisition stages (i.e., one, two, or three stages). This first step of the analysis revealed that the acquisition of comprehension skills can be best conceptualized as a three-stage process, whereas the acquisition of production skills encompassed two stages.

Based on the best-fitting HMMs and the corresponding number of learning stages, I then utilized a series of generalized linear mixed models to investigate the nature of the identified skill acquisition stages. Specifically, I examined whether the three dependent variables, accuracy, RT, and the CV, could be predicted by the three dimension of cognitive abilities at each stage of skill acquisition. The results showed that different cognitive abilities variably predicted learning at each stage, with the trends largely consistent with the general skill acquisition theory (DeKeyser, 2020; Lyster & Sato, 2013; Y. Suzuki, 2022).

Overall, the findings of the study lend support to the three-stage model of L2 skill acquisition, but its proposed mechanisms may have to be revised to suit the specific cognitive processes involved in L2 learning. In addition, when applying the skill acquisition theory (or variants thereof) to L2 learning, one may have to analyze the theory not only at the level of learning mechanisms and processes (e.g., declarative learning, proceduralization, and production tuning) but also at the level of cognitive processing (e.g., lexical and grammatical de/encoding, syntactic parsing, and monitoring) that are specific to L2 learning.

This dissertation is dedicated to my parents.
Words cannot describe how lucky I am to be your son.

ACKNOWLEDGEMENTS

I could not have come this far without the guidance, support, and friendship I received from many individuals. First and foremost, I would like to express my sincere gratitude to my Ph.D. supervisor, Professor Aline Godfroid. Aline has been an unrivaled inspiration for me throughout my study at Michigan State University (MSU). She is my role model as a researcher in second language acquisition (SLA), and I cannot believe how lucky I am to be her student. This dissertation would have never been possible without her patient guidance and caring support. Professor Michael Long, the founder of SLA, once named Aline as someone who would lead the field of SLA for the coming decades. Now, receiving a Ph.D. and graduating from MSU as her advisee, I have never believed that more.

I would also like to thank my committee members, Professors Shawn Loewen, Koen Van Gorp, Paula Winke, and Phillip Hamrick. Dr. Shawn Loewen was the director when I entered the Second Language Studies (SLS) program and continued to support me throughout my study at MSU. He is also the only birdwatcher I could talk to in Michigan, and I cherish every moment I had with him, both serious and comical. I am also grateful to Shawn for giving Kiyo and me the name: *Dynamic Duo*. Dr. Koen Van Gorp is someone who I talked to when it came to task-based language teaching (TBLT), even though I probably did not work on it enough to call myself a TBLT researcher. Koen was also the instructor of my first course at MSU (LLT807: Language Teaching Methods), and it is still my best class at MSU. Dr. Paula Winke is the current director of the SLS program, and she is the one who helped me through many of my professional and administrative issues. I first met/contacted Paula when I was a master’s student at the University of Maryland. I was looking for a working memory task I could use for my master’s thesis. I still remember the exact moment when I got a reply from Paula. I was astonished because someone

like her, established in SLA, could be so kind and approachable. Paula's students always talk highly of her, and there is no wonder why that is the case. Lastly, Dr. Phillip Hamrick is a respected psycholinguist and cognitive scientist who kindly agreed to serve on my committee. I first found Dr. Hamrick's name when I read his doctoral dissertation and later his article in *Language Learning* (Hamrick, 2014) for my master's thesis study. I later met him at the Second Language Research Forum 2018 in Montréal. He kindly asked me a question at my presentation. I followed him after the talk, and I am glad that I gathered the courage to talk to him because he is now a member of my dissertation committee.

I am also indebted to many individuals in the SLS program and other parts of the world for the professional and emotional support they provided me as my colleagues and friends (in alphabetical order): Masaki Eguchi, Curtis Green-Eneix, Bronson Hui, Robert Randez, and Kiyotaka Suga. I also want to thank Professors Robert DeKeyser and Yuichi Suzuki for their inspiration and help throughout my graduate study at the University of Maryland and MSU.

Finally, my *arigato* goes to my family, Masaru, Yumiko, and Yusuke, for their endless support and care. I am forever indebted to them.

This dissertation study was funded by the Nation Science Foundation Doctoral Dissertation Improvement Grant (Award Number: 2140704), *NFMLTA-MLJ* Dissertation Writing Support Grants, the Dissertation Completion Fellowship from the College of Arts and Letters at MSU, and the SLS Dissertation Participant funding. Completing this dissertation would have been difficult without their financial support. In addition, I would like to thank Dr. Caitlin Tenison of Educational Testing Service for providing me with the Python codes to run hidden Markov modeling analysis and answering my queries regarding the analysis.

TABLE OF CONTENTS

INTRODUCTION

Second language (L2) learning is often conceptualized as the acquisition of a set of perceptual, cognitive, and motor skills (see DeKeyser, 2020; Lyster & Sato, 2013; Y. Suzuki, 2022 for reviews). In this view, mastering L2 skills is equated with acquiring skills in other non-linguistic domains, such as typing, driving a car, or solving arithmetic problems. The current research base in the field of second language acquisition (SLA) has provided ample evidence for the parallel nature of L2 learning and skill acquisition with respect to both the process and the product of learning (e.g., De Jong, 2005; DeKeyser, 1997; Ferman, Olshtain, Schechtman, & Karni, 2009; Robinson, 1997; Robinson & Ha, 1993). This collection of evidence suggests that L2 learning can and should be done by the same domain-general learning mechanisms that apply to learning of other non-linguistic skills (e.g., perceptual, motor, and cognitive skills).

L2 researchers to date have turned to neighboring fields, mainly cognitive psychology, for accounts of L2 skill acquisition processes using domain-general cognitive mechanisms (e.g., memory, attention, and problem-solving ability). The current dominant view is to regard L2 skill acquisition as a three-stage process: learning progresses from the cognitive, through the associative, to the autonomous stage (Fitts, 1964; Fitts & Posner, 1967); or the declarative, transitional, and procedural stage (Anderson, 1982, 1983b, 2007). At present, there is indirect evidence that indicates the existence of such developmental stages in L2 learning (Ferman et al., 2009; Pili-Moss, Brill-Schuetz, Faretta-Stutenberg, & Morgan-Short, 2020), yet the number and the nature of the stages are assumed a priori and not themselves the object of research. As an interdisciplinary discipline, the field of SLA has benefited from cross-fertilization with other scientific disciplines, but any theories adopted into L2 research must be tested for their validity vis-à-vis L2 learning. This is critical because the three-stage model (or part thereof) is already represented in many subdomains of SLA (e.g., language instruction: DeKeyser, 1998; 2001; Lyster & Sato, 2013; language assessment: ACTFL, 2012; Council of Europe, 2020), but the model itself is currently only theoretical and lacks empirical support. In the oft-cited review of skill acquisition research in L2 learning, DeKeyser (2020) lamented this fact: “More importantly for our purposes here, not much research in the field of second language learning has explicitly set out to gather data from second language learners to test (a specific variant of) Skill Acquisition Theory” (p. 88).

Against this backdrop of the gap in the literature, this dissertation brings together three lines of research in SLA and cognitive psychology to provide the first direct evidence for (or against) the influential three-stage model of L2 skill acquisition. The first line of research concerns a collection of SLA studies that investigated the parallel nature of L2 learning and skill acquisition by documenting how people develop accuracy and fluency in a novel language as a function of practice (e.g., De Jong, 2005; DeKeyser, 1997; Ferman et al., 2009; Robinson, 1997; Robinson & Ha, 1993). The second line of research focuses on the study of individual differences in skill acquisition to identify what cognitive abilities underlie learning during skill acquisition in general (Ackerman, 1987, 1988, 1990, 1992; Ackerman & Cianciolo, 2000; Ackerman, Kanfer, & Goff, 1995) and L2 skill acquisition in particular (Li, 2017; Maie, 2021; Pili-Moss et al., 2020; Y. Suzuki, 2018). Lastly, the third line of research involves using cognitive modeling to mathematically model skill acquisition processes to detect distinct phases of learning during skill acquisition (Tenison & Anderson, 2016; Tenison, Fincham, & Anderson, 2016). By taking conceptual and methodological insights from each of the three lines of research, this dissertation sets out to test the number and the nature of skill acquisition stages in the context of L2 learning.

Before moving forward, I would like to clarify the terminology that will be used throughout the dissertation. First, the term *L2 learning* will refer to any kind of development in L2 knowledge or performance without recourse to a specific learning process or mechanism involved. Hence, the term covers L2 development in any approach, be it formal, usage-based, or skill acquisition approaches. When discussing L2 learning in general, I will thus prefer the word *learn* over *acquire*. However, I will use the term *skill acquisition* to refer to the entire process of mastering skills because it is generally preferred over the other terms such as skill *learning* or *development*. Hence, skill acquisition in L2 learning will specifically be called *second language skill acquisition* or *L2 skill acquisition*, and the general theory of skill acquisition (without recourse to any specific models of process or mechanism) will be the *skill acquisition theory*. Lastly, I will use the term *second language acquisition* or *SLA* to refer to the scientific research on how people learn and use any additional language after one’s first language; hence, I will use the term *second language* or *L2* to mean any languages other than one’s first language.

CHAPTER 1: REVIEW OF THE LITERATURE

In this chapter, I review the literature on skill acquisition as it relates to L2 learning. I will first provide an overview of skill acquisition processes and discuss phenomena that are almost universally observed in skill acquisition research (Section 1.1). I will then define the concept of automaticity and automatization as the end product of skill acquisition and review some methods that have been proposed to operationalize automaticity/automatization (Section 1.2). Theoretical models of skill acquisition will then be introduced, including the three-stage model and its rival models, with particular attention paid to how each model accounts for the skill acquisition phenomena and the development of automaticity (Section 1.3). I will then review a recent line of research in cognitive psychology that attempted to pit the rival theoretical models against each other by modeling skill acquisition processes using cognitive modeling (Section 1.4). After clarifying the underpinning concepts and theories of skill acquisition, I will review the literature on L2 skill acquisition, focusing on evidence that shows the parallel nature of skill acquisition in general and L2 learning and research that investigates the role of cognitive individual differences in (L2) skill acquisition (Section 1.5). Finally, I will point out the fundamental problems in the L2 skill acquisition literature and discuss how the present dissertation research achieves to fill the gaps. Because the overall theme of the dissertation is to investigate the applicability of cognitive skill acquisition theory in the context of L2 skill acquisition, in Section 1.1–1.4, I will primarily draw from the cognitive psychology literature to review the basic phenomena of skill acquisition, the definition and measurement of automaticity/automatization, and theoretical models of skill acquisition.

1.1 Skill Acquisition and the Associated Phenomena

Skill acquisition is the process of learning skills to advanced proficiency “from initial representation of knowledge through initial changes in behavior to eventual fluent, spontaneous, largely effortless, and highly skilled behavior” (DeKeyser, 2020, p. 83). The process of skill acquisition has been well documented in a variety of domains, ranging from perceptual (e.g., Kolers, 1975; Neisser, Novik, & Lazar, 1963) and motor skills (e.g., Card, English, & Burr, 1978; Crossman, 1959; Snoddy, 1926) to complex cognitive routines (e.g., Anderson, 1983a; Card, Moran, & Newell, 1980; Compton & Logan, 1991; Neves & Anderson, 1981). Although the interpretation of the process varies in detail from one researcher to another, there is a general consensus that (a) extended practice is necessary to achieve full mastery of the skill and that (b) the end state of skill acquisition is automaticity (see Section 1.2 for more discussion of automaticity) (see VanLehn, 1996; DeKeyser, 2001, 2020, for a review in cognitive skill acquisition and in L2 skill acquisition, respectively). Additionally, the current stockpile of skill acquisition research has shown that two phenomena are almost universally observed in learning of any skill. These phenomena are (a) the power-law of practice and (b) the skill specificity. Due to the ubiquity of the phenomena, every theoretical model of skill acquisition is expected to account for how proposed learning mechanisms give rise to the phenomena.

1.1.1 The Power-Law of Practice

The power-law of practice is a scientific law of learning that states that the time it takes one to perform a skill decreases with the amount of practice, and the decrease follows a specific non-linear curve defined by a power function. Figure 1.1 (left panel) provides an example of a power function applied to skill acquisition data. The speedup in performance following the power law is typified by an initial short period of rapid decrease in performance times followed by a gradual and slow process of fine-tuning skill performance to reach the asymptotic level of performance. The seminal article by Newell and Rosenbloom (1981) illustrated that the same power function applies to a great variety of skills and domains, qualifying the phenomenon as a scientific law. Although the exact form of the power function is a subject of debate, its basic form can be expressed in the following formula:

$$T = \mathrm{I} + \beta \mathrm{N}^{-\alpha}$$

T = is the time to perform a skill, I is the asymptote (i.e., the psycho-physical limit of speed one can achieve after an infinite amount of practice), β is the difference between the initial trial and the asymptote, and therefore how much one can speedup (after an infinite amount of practice), N is the number of practice trials, and α is the learning rate that controls how fast one can speed up. The minus sign on the exponent (α) produces the decelerating decrease in performance times. One mathematical corollary of the power function is that plotting the logarithm of performance times (T) against the logarithm of practice trials (N) yields a straight line; $\log(T) \sim \log(\mathrm{N})$ produces $R^2 = 1.0$. Hence, the linear regression of $\log(T)$ (as the dependent variable) on $\log(\mathrm{N})$ (as the predictor) serves as a test tube of whether one's dataset conforms to the power-law of practice. This is why Newell and Rosenbloom (1981) alternatively called the phenomenon the log-log linear law. Some researchers showed that the accuracy of performance (Anderson, 1995) and the standard deviation of performance times (Logan, 1988, 1992, 2002) also follow the power-law of practice. However, evidence is currently too limited to draw any conclusions.

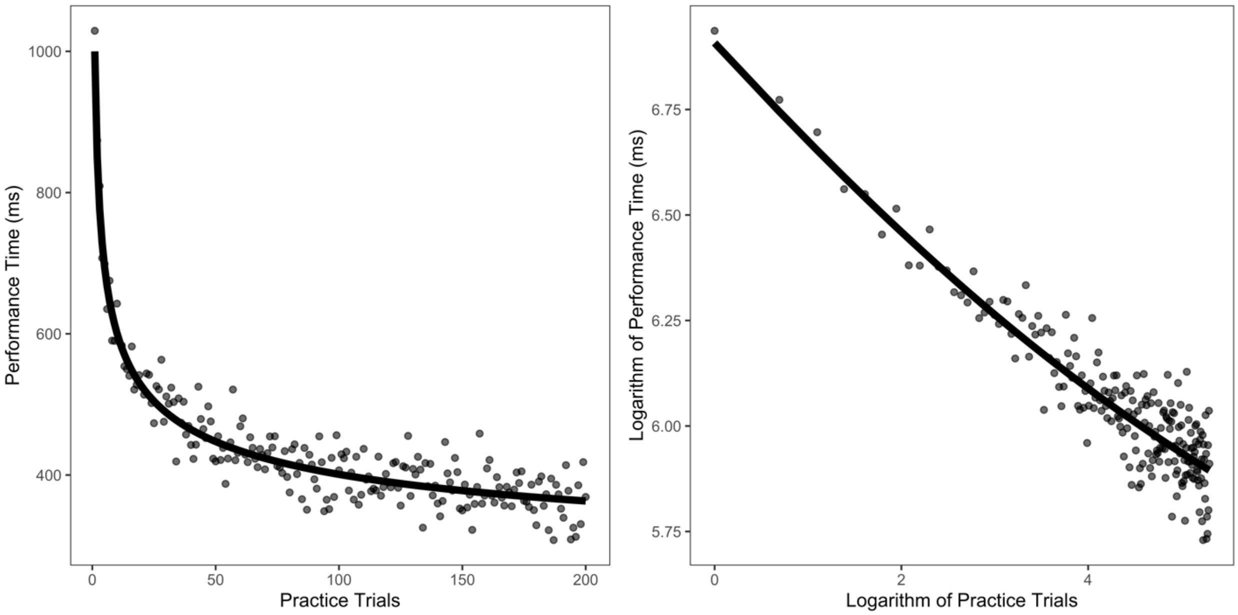


Figure 1.1. An example of the power function fit to skill acquisition data.

Note. The left panel shows data on the original scale and the right panel shows the data with performance times and practice trials transformed to their natural logarithms. The dataset was simulated from a power function: $T = 200 + 800N^{-0.3} + \mathrm{N}(0, 30)$, 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= .992.

One recurring criticism against the power-law of practice is that it often does not apply to individual data points and hence its ubiquity may be an artifact of data averaging (Adi-Japha et al., 2008; Gallistel et al., 2004; Haider & Frensch, 2002; Heathcote et al., 2000; Myung et al., 2000). For instance, Adi-Japha et al. (2008) pointed out that applying a power function to aggregated data may conceal important patterns in individuals' gains. Heathcote et al. (2000) revealed that an exponential function fit better than a power function for all 7,910 learning series (from 475 participants) of performance times. The exponential function takes the form of $T = \mathrm{I} + \beta e^{-\alpha N}$; hence, the practice trials (N) is the exponent of *e* (i.e., the Euler's number or ≈ 2.71828) rather the base of the exponent α in the power function. Heathcote et al. further proposed what they called the APEX function, which incorporated an additional pre-experimental practice parameter. One reason of the power function's inability to explain individual data is that raw data are (more) prone to inherent variability (than when they are aggregated). However, the variability not only results from systematic effects caused by the underlying learning mechanisms but also due to pure performance or sampling error that is nothing to do with the process of skill acquisition. While deviations from the power function due to the systematic effects are worth of close theoretical and empirical attention, it is unwise to question the power-law of practice if the deviations result from sampling error. Recently, some researchers showed that when sampling error is incorporated in a model (coupled with different power functions for distinct learning phases; see below), the power function and the exponential function become highly equivalent in how they fit individual and item-level data (but the APEX function shows substantially poorer fit) (e.g., Tenison & Anderson, 2016). Consequently, the compatibility of the power function to raw data is still an open question at best.

Another issue surrounding the power-law of practice is the number of power functions required to account for skill acquisition data. Initially, Newell and Rosenbloom (1981) and other trailblazing research (e.g., Anderson, 1981; Logan, 1988; MacKay, 1982) only considered using a single power function. However, Rickard (1997, 1999) later demonstrated that two power functions better account for data (see Delaney et al., 1998 for the same finding). As described in Section 1.3 (Models of Skill Acquisition), the number of power functions crucially depends on whether a given theory of skill acquisition describes the speedup in skill execution as a quantitative change in the underlying cognitive process (i.e., improvement of the same process) or a qualitative change (i.e., shifting to more efficient processes). The idea of connecting the power-law of practice to the underlying cognitive mechanisms is crucial because “[the power-law] has captured little attention, especially theoretical attention, in basic cognitive or experimental psychology, though it is sometimes used as the form for displaying data” (Newell

& Rosenbloom, 1981, p. 2). It is thus incumbent on skill acquisition researchers to theorize and investigate what its shape and ubiquity signify in terms of psychological mechanisms.

1.1.2 The Skill Specificity

The skill specificity refers to the negative correlation between the amount of practice and the generalizability of a skill; when the skill is highly practiced through a specific task or in a specific domain (e.g., visual vs. auditory), it becomes less available for transfer on another task or in a different domain. The classic experiment on the transfer of learning by Thorndike and Woodworth (1901) (see also Thorndike, 1906) was the first to document that skills may not transfer well from one task to another unless the tasks share "identical elements" (e.g., content and procedure). In skill acquisition research, a corpus of evidence has demonstrated the specificity of extensively practiced skills (see Healy & Bourne, 1995; Rogoff & Lave, 1984). Singley and Anderson (1989) provided the most impressive demonstration of the specificity of cognitive skills to date (and also an extensive literature review on the issue of transfer), showing that the skill of reading versus writing computer programs, when highly practiced, give rise to an overwhelming directional asymmetry in that performing the same skill in a reverse direction (i.e., writing computer programs when one extensively practiced reading or vice versa) leads to more errors and slower performances (see also Anderson & Fincham, 1994). As with the power-law of practice, the ubiquity of the skill specificity demands theoretical explanations.

The specificity of skill is not without counterevidence, however. Ackerman (1990) reviewed that two independent camps have investigated the issue of transfer in skill acquisition, a group of traditional transfer-of-training experiments (see Adams, 1987; Singley & Anderson, 1989 for reviews) and a group of individual-differences studies (see Marteniuk, 1974 for a review). While the former includes research reviewed above and examines whether training of a skill through a criterion task or in a specific domain influences performance on a transfer task or in a different domain, the latter group of studies investigates how individual differences in performing the criterion task can be predicted by individual differences in abilities that are theoretically related to the target skill. The underlying argument behind the individual-differences paradigm is “difficulties in discovering abilities that predict individual differences at highly skilled levels of performance” (Ackerman, 1990, p. 883); that is, if learning a skill is specific to a task or domain, abilities that are related to the performance of the skill should not predict the asymptomatic performance of the skill (see also Fleishman, 1972; Marteniuk, 1974). With this prediction, several studies show that some cognitive abilities indeed predict performances of a skill even at the asymptomatic level (Ackerman, 1987, 1988, 1990, 1992; Adams, 1957, 1987), therefore discrediting the skill specificity phenomenon. However, a caveat in interpreting these findings is that the correlational design of the individual-differences paradigm does not directly speak to transfer. In this light, findings from the traditional transfer-of-training experiments bear more empirical credibility. Nonetheless, Ackerman and colleagues provided a useful theoretical model of individual differences in skill acquisition (Ackerman, 1987, 1988, 1990, 1992; Ackerman & Cianciolo, 2000; Ackerman et al., 1995), which will be reviewed as one of the main theoretical bases for my dissertation study (Section 1.5.2).

1.2 Automaticity and Automatization

The concept of automaticity (or an automatic process) abounds in one’s everyday life. A sliding door opens automatically reading from an overhead motion sensor, a vehicle with an automatic transmission shifts gears by itself, and a modern robotic vacuum cleaner operates automatically on itself. Humans also display an amazing degree of automaticity. One does not (or is not able to) think about the process of standing or moving hands (i.e., innate automaticity), or even when carrying out complex skills such as using a smartphone, driving a car, or speaking the first language; as long as one is highly experienced, the underlying perceptual, cognitive, and motor processes do not even come to the mind (i.e., acquired automaticity). The word "automatic" originates from the Greek word (adjective) *automatos*: "self-acting" or "acting on its own". Therefore, automaticity implies that the process operates without any intervention from an active agent. Although the omnipresence of automaticity seemingly attests that its concept is well understood by scientists and the general public alike, cognitive psychologists have diverged regarding the specific features that characterize automaticity.

1.2.1 Concepts of Automaticity and Automatization

The first systematic attempt to tackle the concept of automaticity was Shiffrin and Schneider's (1977) (also Schneider & Shiffrin, 1977) dichotomy of controlled and automatic processing (i.e., the dual-process theory). The theory was innovative in that the researchers tackled the construct of automaticity as a problem of attention: a process is said to be automatic if it takes place "without the necessity of active control or attention by the subject" (Shiffrin & Schneider, 1977, pp. 155–156). An automatic process is thus a cognitive activity that does not expend (or barely requires) one's attentional resources in working memory, whereas a controlled process entails attentional processing that calls for one's cognitive resources. Learning is considered a transition from a controlled to an automatic process (see Schneider & Chein, 2003 for a more recent treatment of the dual-process theory). In L2 learning, Johnson (1996) also defined automaticity from the same perspective, as "the ability to get things right when no attention is available for getting them right" (p. 137).

The consequence of theorizing the dichotomy with respect to attention is that automatic and controlled processes can be juxtaposed with other in terms of their functional characteristics.

While an automatic process is parallel, ballistic, and effortless, a controlled process is serial, controllable, and effortful. In the context of visual search tasks (i.e., searching for a target from a display with other similar objects), Schneider and Shiffrin (1977) operationalized automatic processing as load-independent processing; if the process of visually searching a target is automatic (and hence does not require attentional resources), it should operate regardless of how much information needs to be processed at the same time. Schneider and Shiffrin found that automaticity only developed in a specific condition called "consistent-mapping" (as opposed to "variable-mapping") in which a stimulus (e.g., a visual image) always appeared as a target in the experiment but never as a distractor. As Segalowitz (2003) discussed, the role of consistent stimulus-response experiences has an important implication for language learning because a linguistic unit (stimulus) (e.g., a word or a morpheme) may not exclusively map onto a single semantic referent (response).

One caveat of Shiffrin and Schneider's (1977) dichotomy is that it presents automaticity as a binary phenomenon, automatic or not automatic. However, many cognitive psychologists now view automaticity as being on a continuum, and the transition from controlled to automatic processing is investigated as a gradual (rather than an abrupt) process. Along with domain-independent processing, researchers have also proposed many additional features of automaticity: (a) fast, (b) parallel, (c) effortless, (d) ballistic, (e) result of consistent practice, (f) unimpeded (little interference from) by a secondary task, (g) unconscious, and (h) based on memory retrieval. This large collection of the features does not mean that cognitive psychologists agree on the nature of automaticity. Moors and De Houwer (2006) (also Moors, 2016) discussed the field's inability to reach a consensus concerning which set of the features one should use to define the construct of automaticity. As a case in point, Posner and Snyder

(1975) claimed three features (ballistic, little interference from a secondary task, and unconscious), whereas MacKay (1982) proposed four features (fast, effortless, little interference, and unconscious). Schneider and Detweiler (1988) mentioned only one feature (little interference), but Anderson (1992) listed five features (fast, little attention, ballistic, consistent practice, and little interference) (see DeKeyser, 2001, p. 128 for a very similar but more extensive illustration). Given the researchers' wide disagreement, providing componential explanations of automaticity that "unpacks the components of automaticity and specifies the relations among them" (Moors, 2016, p. 264) may be an intractable problem. Rather, the aim of this dissertation is to seek a mechanistic explanation of automaticity that "specifies the low-level processes underlying automatization" (Moors, 2016, p. 264).

The word automatization is also a multi-sense term. Cognitive psychologists have largely used the term with three different levels of generality with respect to the developmental continuum of automaticity (see Figure 1.2 for illustration): (a) the entire process of developing automaticity, from initial slow(er) performance through the initial sharp drop in performance times to the gradual fine-tuning of the skill and increasing speed to reach the asymptotic performance (see Figure 1.1, left panel); (b) the initial drop in performance times only; or (c) the gradual fine-tuning of the skill only. Currently, the interpretation of the term depends on the underlying mechanisms posed to explain the development of automaticity (see Section 1.3). In the three-stage model, the initial slower performance is due to the interpretive application of declarative knowledge, and the abrupt reduction in performance times is attributed to proceduralization (Anderson, 1982, 1983b, 2007; see Section 1.3.1). In this dissertation, I will thus reserve the term automatization to refer to the flatter part of the curve where one gradually fine-tunes the performance to reach the asymptotic level (the right panel in Figure 1.2). This characterization of automatization is also consistent with how the term is defined in the SLA literature (DeKeyser, 2020; Y. Suzuki, 2022).

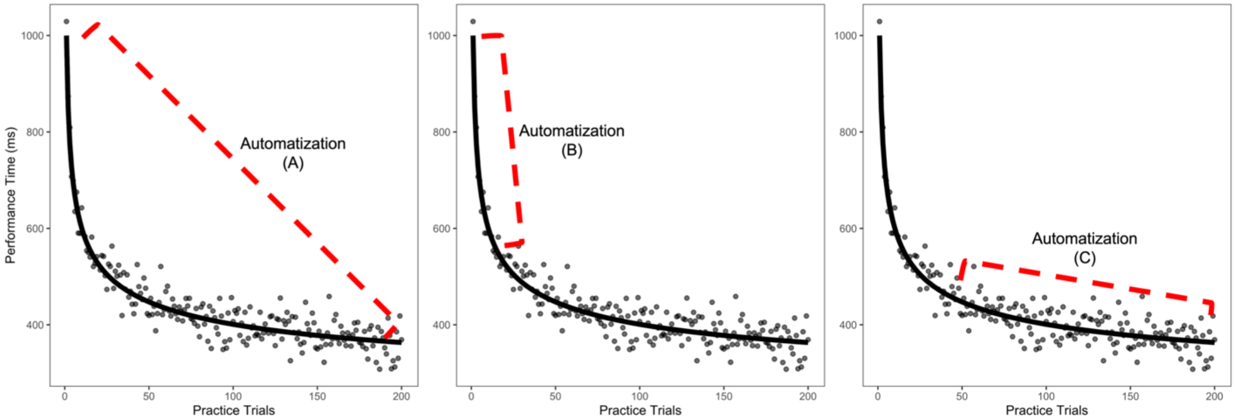


Figure 1.2. The three meanings of automatization.
Note. The left panel shows automatization as the entire process of developing automaticity, the middle panel shows automatization as the initial sharp drop in performance time, and the right panel shows automatization as the gradual fine-tuning of the skill performance.

One issue that surrounds the mechanistic explanation of automaticity is whether it arises from a quantitative change in the underlying mechanism (i.e., improvement of the same process), a qualitative change (i.e., shifting to more efficient processes), or the combination of both. Rival models of skill acquisition propose competing proposals on this point (see Section 1.3). As reviewed in the next section, theorizing the exact mechanism of developing automaticity is crucial as it has an implication for how the construct of automaticity can be measured.

1.2.2 Measuring Automaticity

Establishing reliable measurements of automaticity is crucial for understanding the process and the mechanism of developing automaticity. It is possible that one operationalizes all potential features of automaticity (see Section 1.2.1) and observes how indices operationalizing the features change as a result of practice. However, since cognitive psychologists disagree as to the best set of features that characterize automaticity (Moors, 2016; Moors & De Houwer, 2006), this method may not be an empirically realistic plan. Traditionally, skill acquisition researchers have focused on the speed aspect of skill performance (i.e., an automatic process is fast). The most common method is to track the reaction time (RT) of performance and examine how it declines as a function of practice. Given the robust findings on the power-law of practice, observing that participants have reached an asymptotic level of performance (i.e., reaching almost the end of the power-law curve; see Figure 1.1) likely indicates that the participants have achieved automaticity. Obviously, this method captures only one aspect of automaticity as it singly focuses on how fast one carries out the skill.

The most innovative development in the measurement of automaticity comes from a research program in the field of SLA led by Norman Segalowitz and colleagues (Phillips, Segalowitz, O'Brien, & Yamasaki, 2004; Segalowitz & Freed, 2004; Segalowitz & Segalowitz, 1993; Segalowitz, Segalowitz, & Wood, 1998; Segalowitz, Watson, & Segalowitz, 1995; see Segalowitz, 2010 for a review). Segalowitz and Segalowitz (1993) asserted that automaticity not only requires performance to be fast but also carried out with stability (i.e., less variability in the speed of performance). This proposition is based on the belief that automaticity results from "qualitative changes in the functioning of the underlying processes through a restructuring effect", whereby component processes underlying the skill performance "become organized differently; for selected, inefficient processes to drop out; for new, more efficient processes to replace older, less efficient ones; or for some mixture of all these possibilities to occur" (Segalowitz & Segalowitz, 1993, p. 373–374) (see also McLauglin, Rossman, & McLeod, 1983; McLeod & McLaughlin, 1986 for theoretical reviews, especially in L2 contexts). Operationally, Segalowitz and Segalowitz proposed the coefficient of variability (CV) of RT (i.e., the standard deviation of RT divided by the mean of the RT) as a measure of processing stability and suggested that a researcher must examine reductions in both RT and the CV. This is because when learners simply speeds up their performance, only RT is expected to drop (but not CV), and the standard deviation (SD) of the RT decreases at the same rate (due to the mathematical nature of smaller numbers being associated with smaller variability). However, when one's performance becomes more stable, and hence more automatic, not only RT but also the CV would decrease, and because the SD (of RT) decreases disproportionately to the rate of the RT decrease, RT and CV will positively correlate. Therefore, the evidence of automaticity in this paradigm requires both (a) RT and (b) CV decrease in some meaningful way (e.g., statistical significance), and (c) the RT and CV positively correlate with each other.

Currently, the CV has been most widely used to investigate the automaticity of lexical processing, measured through such tasks as a lexical decision task (e.g., Segalowitz & Segalowitz, 1993; Segalowitz et al., 1998; Elgort, 2011) and a semantic classification task (e.g., Phillips et al., 2004; Segalowitz & Freed, 2004; Hui, 2020; Maie & Godfroid, 2022). An issue, however, is whether the CV can also be a valid measure of automaticity at the sentence-level performance. While the first validation attempt by Hulstijn, Van Gelderen, and Schoonen (2009) showed that the validity of the CV may be limited to lexical-processing tasks only, a conceptual replication by Lim and Godfroid (2015) showed that the CV can be as useful in the study of sentence processing (using sentence construction and verification tasks) as in the study of lexical processing. More recent findings seem to support the conclusion of Lim and Godfroid (see Y. Suzuki, 2018; Pili-Moss et al., 2020; but cf. McManus & Marsden, 2019).

Segalowitz and Segalowitz (1993) and subsequent research called the CV a measure of "automatization", but they neither explicated the exact mechanism of automatization nor specified what level on the continuum of developing automaticity (see Figure 1.2) the reduction in the CV corresponds to. Theoretical models of skill acquisition (Section 1.3.1 and 1.3.2) largely argue that the putative qualitative changes in the underlying processes occur at the initial level of development, causing the early abrupt reduction in performance times. In Section 1.2.1, we defined the term automatization to refer to the latter (and flatter) end of the power-law curve. Hence, the definition of automatization as used by Segalowitz and Segalowitz (1993) does not match what is considered automatization in this dissertation. Using the terminology from the three-stage model of skill acquisition (Anderson, 1982, 1983b, 2007), the CV is rather a measure of *proceduralization* than automatization.

One last issue surrounding the use of the CV concerns how it may decrease as a function of practice. Does it linearly decrease with the amount of practice, or does the decrease follow some specific mathematical function such as a power function? A common finding in skill acquisition research is that RT tends to deviate from the power-law curve at the very initial level of development (e.g., Delaney et al., 1998; Logan, 1988; Newell & Rosenbloom, 1981; Neves & Anderson, 1981; Rickard, 1997). This is primarily because initial performance is slow and hence subject to higher variability. Intuitively, it is thus possible to predict that the CV (as a measure of performance stability and variability) may not decrease smoothly in the early phase of learning. In a similar vein, Solovyeva and DeKeyser (2018) focused on cases where the CV initially increases rather than decreases as a function of practice. They argued that the increase could be due to the addition of a new knowledge representation: “If reductions in the CV result from the elimination of (inefficient) component processes or their restructuring, increases in the CV should signal the opposite—the addition of component processes or new representations” (p. 228). Later, Hui (2020) also demonstrated that the CV can initially increase before it starts to decrease as an index of automaticity. However, a recent longitudinal study of practice (using an artificial language) by Pili-Moss et al. (2020) found that the CV simply decreased without the initial increase. At present, the issue is far from being settled. Nonetheless, the CV has the potential to index the changes in the functioning of the underlying processes as long as its use is accompanied by a detailed analysis of how it changes as a function of practice.

1.3 Models of Skill Acquisition

In cognitive psychology, several models of skill acquisition propose competing explanations of how learners achieve automaticity through repeated practice. Crucially, the models are classified according to whether they are based on a ruled-based approach or an item-based approach. In this section, I first review the three-stage model as the leading theory of the rule-based approach. Specifically, I draw on Fitts and Posner's (1967) (also see Fitts, 1964) three-stage model of learning and Anderson's Adaptive Control of Thought (ACT) theory (Anderson, 1982, 1983b, 1992, 2007; Anderson et al., 2004, 2019) to describe how the three-stage model provides a mechanistic explanation of skill acquisition and the development of automaticity (Section 1.3.1). Rival models are then introduced, and I specifically focus on the Race model (Compton & Logan, 1991; Logan, 1988, 2002) as a major theoretical model from the item-based approach, and the Component Power Laws (CMPL) theory (Bajic & Rickard, 2011; Rickard, 1997, 1999, 2004) as a model that combines the rule-based approach and the item-based approach (Section 1.3.2). The three theoretical models advance rival hypotheses regarding the number and the nature of skill acquisition stages and how they account for the phenomena widely observed in skill acquisition research (Section 1.3.3).

1.3.1 The Three-Stage Model

In theorizing the process of skill acquisition, Fitts (1964) made the first observation that learning a skill seems to consist of three phases, with transitions between stages caused by “gradual shifts in the factor structure of skills, or in the nature of processes (strategies and tactics, executive routines and subroutines) employed” (p. 261). This proposition of the “gradual shifts” assumed that each learning phase involved distinct cognitive processes. Later, Fitts and Posner (1967) summarized the three stages as the cognitive stage, the associative stage, and the autonomous stage. Although Fitts and Posner originally proposed the three stages within the context of perceptual-motor skill learning, they believed that the same theory applies to the learning of cognitive and linguistic skills as well (see Fitts, 1964, p. 243).

The cognitive stage is characterized by initial slow and controlled performance in which learners must encode the skill into some crude form that can produce the target behavior. Executing a skill in this stage is mentally taxing because the encoding process heavily relies on one’s working memory. Hence, learners are commonly observed to use working memory to verbally rehearse information required to execute the skill (i.e., verbal mediation). The necessity of this verbal mediation drops out in the associative stage because learners develop direct behavioral routines to execute the skill, and this direct route significantly reduces one’s time on task. However, the newly developed procedure can be error-prone and variable in its application, so it requires further practice to be applied more correctly and efficiently. Eventually, one achieves the autonomous stage after a long period of practice applying the direct procedure. At this point, the execution of the skill becomes automatic, which means that the skill can be performed effortlessly and simultaneously with another task that is cognitively demanding.

The three-stage model by Fitts and Posner (1967) remained agnostic on the specific psychological mechanisms responsible for learning during each stage. Anderson’s Adaptive Control of Thought (ACT) theory (Anderson, 1982, 1983b) and its computer-implemented cognitive architecture, Adaptive Control of Thought-Rational (ACT-R: Anderson, 2007;

Anderson et al., 2004, 2019; Anderson & Fincham, 1994; Anderson & Lebiere, 1998), extended the three-stage model by adding cognitive explanations of how learners come to master skills through the three stages. A cognitive architecture such as ACT-R is a theory of how human cognition learns and organizes knowledge to produce intelligent behaviors (see Anderson, 2007 for a more extensive treatment of how ACT-R simulates human cognition). In particular, ACT-R is specific to account for the acquisition of cognitive skills. In ACT-R, knowledge is represented as either declarative knowledge or procedural knowledge. Declarative knowledge is the knowledge of factual information (i.e., episodic and semantic knowledge), whereas procedural knowledge is the knowledge of how to perform a given skill. ACT-R implements declarative knowledge as "chunks" of information (Miller, 1956), and procedural knowledge is represented as sets of production rules. A production rule is a primitive rule in the form of a condition-action pair (or an IF-THEN conditional, see Table 1.1 below), which encodes a cognitive contingency such that when(ever) the condition is met, the action is performed (Anderson, 1982). As Anderson (1983b) explained, "[i]f there is one term that is most central to the ACT theory, it is 'production'". The central role of production rules makes ACT-R a rule-based approach.

Skill acquisition in ACT-R begins with learners gaining declarative knowledge about a skill through receiving instruction or observing others acting out the skill. At this initial stage, the only way to execute the skill is by engaging general skill-independent procedures (i.e., general-purpose production rules), which interpret declarative knowledge and produce the target behavior at a rudimentary level. According to Anderson (1983b), there are at least three ways in which declarative knowledge can be interpretively used by the general-purpose production rules: by (a) faithfully following declarative information that takes the form of direct instruction (e.g., a recipe); (b) using general problem-solving algorithms that will work out the skill using general knowledge within a domain (e.g., for unknown arithmetic problems, using general knowledge of mathematics); and (c) using analogy-forming procedures that map a declarative representation of a previously observed behavior onto a new behavior. The variety in which declarative knowledge can be used to guide a behavior creates the flexibility of how human cognition learns in different learning environments. The flexibility of learning at this stage is also crucial because a skill becomes hard to modify after it is extensively practiced (i.e., the skill specificity, see Section 1.1.2). This phase of learning corresponds to the cognitive stage in Fitts and Posner's model.

While the interpretation of declarative knowledge has the advantage of flexibility, it is often a slow and costly process because it requires declarative information to be retrieved from long-term memory and maintained in working memory. As a result, learners develop skill-specific procedures, or procedural knowledge, to optimize their performance. These procedures are still production rules, but they are skill-specific so that they can be applied directly, without the mediation of interpretive productions. ACT-R implements this process through *knowledge compilation*, a process subsuming two mechanisms: *composition* and *proceduralization* (Anderson, 1986; 1987; Neves & Anderson, 1981). Composition collapses sequences of production rules into one larger production, and proceduralization creates a novel procedure that is specific to the skill being practiced. The new, proceduralized production no longer requires the interpretation of declarative knowledge because the production directly incorporates declarative knowledge in its procedure. Table 1.1 illustrates the process of knowledge compilation for learning how to produce the word *broke* as the irregular past tense form of the word *break* (cf. Taatgen & Anderson, 2002). Declarative knowledge here is the fact that (a) some words take a specific irregular form for the past tense and (b) the past tense form of *break* is *broke*.

Table 1.1. Production rules to produce the past tense form of the word *break*.

	Interpretive use of declarative knowledge
P1:	IF the goal is to produce a past tense of a word AND there is a specific form for the past tense of the word THEN set the answer of the goal to retrieving the specific form.
P2:	IF the goal is to retrieve a specific form for the past tense of a word AND the word is *break* THEN set the answer of the goal to retrieving the past tense form of *break*.
P3:	IF the goal is to retrieving the past tense form of *break* AND the past tense form of *break* is *broke* THEN set the answer of the goal to retrieving *broke.*
	Macro-production created by composition
P4:	IF the goal is to produce a past tense of a word AND there is a specific form for the past tense of the word AND the word is *break* AND the past tense form of the word is *broke* THEN set the answer of the goal to retrieving *broke*.
	Skill-specific production created by proceduralization
P5:	If the goal is to produce the past tense form of the word *break* THEN set the answer of the goal to retrieving *broke.*

Initially, producing the past tense form *broke* requires three production rules that need to be executed serially (i.e., P1→P2 →P3). However, composition collapses the three productions into one larger macro-production, and this process greatly reduces one's time on task because what used to take three productions is now done by a single production. Proceduralization further takes the macro-production and develops a novel production that is skill-specific (producing *broke* as the past tense of *break*). Notice that the two pieces of declarative knowledge, (a) *there is a specific form for the past tense of the word* and (b) *the past tense form of the word is broke*, have dropped out in the new production. This is how proceduralization develops procedural knowledge based on declarative knowledge, thereby obviating the necessity of learners' reliance on working memory. Assuming that automatic behaviors are ones that do not entail attention (as one aspect of automaticity: Schneider & Chein, 2003; Shiffrin & Schneider, 1977), proceduralization is what drives automaticity. The process of knowledge compilation is a transitional stage during which one relies on both declarative and procedural knowledge. Hence, this phase of learning corresponds to the associative stage in Fitts and Posner's model.

Later, compiled productions undergo the process of *tuning*. Production tuning is a piecemeal process that increases or lowers the probability of a given production rule being chosen as the method for the skill (see also Rumelhart & Norman, 1978 for a review of tuning as part of the fundamental human learning mechanisms). Often, there are multiple ways to carry out the same skill, and learners must search for the best method to perform the task. To achieve the optimal balance between the cost and benefit of finding the solution, the search process must be both correct and fast. Anderson, Kline, and Beasley (1980) proposed three subprocesses (of tuning) to make this possible: *generalization*, *discrimination*, and *strengthening*. While generalization makes a production rule broader in its applicability, discrimination makes the scope of the production narrower. Strengthening controls the probability of selecting among different production rules such that more useful rules are strengthened and poorer rules are weakened. Although its process is more gradual than that of knowledge compilation, tuning also increases one's speed in performing the skill. Note that knowledge compilation is a qualitative change in the underlying cognitive processes, whereas tuning is a quantitative change that incrementally optimizes the same process. Additionally, ACT-R also allows *declarative strengthening*, which increases the accuracy and the speed of retrieving declarative knowledge per se. In Fitts and Posner's model, this phase corresponds to the autonomous stage.

Before concluding the review of the three-stage model, it is worth clarifying the distinction between the concept of within-stage speedup and between-stage speedup. When one increases the speed to perform a skill, an individual can become faster either within or between stages. The within-stage speedup results from improving the same psychological process; that is, by a continual quantitative change in the underlying mechanism. In contrast, the between-stage speedup is caused by qualitative changes by shifting the underlying cognitive process (or strategy) to a more efficient process. The major prediction of the three-stage model is that a majority of the performance speedup is ascribed to between-stage speedup (see Tenison & Anderson, 2016 for empirical support). This is because the between-stage speedup, such as knowledge compilation, not only increases the speed of performance but also makes its procedures more efficient.

1.3.2 Rival Models of Skill Acquisition

The Race model is an item-based approach to the mechanism of skill acquisition and the development of automaticity that has been cited as widely as the ACT theory in the cognitive psychology literature. The model was proposed by a cognitive psychologist, Gordon Logan, as part of the instance theory of automatization (with a larger sense than defined in this dissertation) (Compton & Logan, 1991; Lassaline & Logan, 1993; Logan, 1988, 1992, 2002). The Race model claims that each experience of carrying out a skill becomes encoded in memory as an “instance”. Although the model does not explicate what constitutes an instance, it is considered a type of memory trace that contains contextual and task-specific cues that are relevant to the skill. The more instances learners accrue in memory, the faster one’s performance becomes. Initially, learners perform a skill using general problem-solving algorithms (e.g., directly calculating an answer for an arithmetic problem), but they begin to rely more on retrieving past solutions as the number of instances increases with practice. The model assumes that algorithm and retrieval run in parallel; initially, algorithm strategies outpace retrieval, but the gradual accumulation of instances (with themselves racing against each other) speeds up the retrieval process and eventually takes over. A skill is thus considered automatic “when it is based on single-step direct-access retrieval of past solutions from memory” (Logan, 1988, p. 493). Note that the model does not conceive any speedup in the use of algorithms. Rather, any reductions in performance latency must be attributed to a single mechanism of accumulating instances. Because the same cognitive state (i.e., the parallel execution of algorithmic processing and memory retrieval) is maintained throughout the entire process of skill acquisition, the Race model is a one-stage model that conceptualizes the development of automaticity as a quantitative change (see also Section 1.3.3 for conceptualizing the Race model as a one-stage model).

While the Race model accounts for the practice-induced speedup using the process of accruing instances, the Component Power Laws (CMPL) Theory offers a slightly different mechanism of skill acquisition (Rickard, 1997, 1999, 2004; see also Bajic & Rickard, 2011 for a recent treatment of the theory). The CMPL theory maintains that the transition from algorithms to memory retrieval drives the development of automaticity. However, it does not assume that the two processes run in parallel. Rather, learners choose between the two from the outset, either applying algorithms or retrieving the answer from memory. More importantly, the CMPL theory suggests that the speedup not only applies to the memory retrieval process (as in the Race model) but also to the execution of algorithmic processing. In the CMPL theory, the speedup in skill execution is driven by two forces. On the one hand, learners learn to behave more efficiently by shifting the task strategy from algorithms to direct memory retrieval (between-stage speedup). On the other hand, algorithmic processing or retrieval process can themselves be accelerated (within-stage speedup). This is why Rickard (1997, 1999) proposed that two power functions, one for algorithm use and the other for memory retrieval, better account for experimental data (see Section 1.1.1). This makes the CMPL theory a two-stage model.

In summary, existing models of skill acquisition offer some insights into how skill acquisition may occur in learning of cognitive skills. ACT-R, the Race model, and the CMPL theory all share that skills are initially performed using general-purpose mechanisms, but later, skill-specific procedures (or specific memory instances) overtake as learners accumulate experience using the skills. The Race model conceives the development of automaticity as a single quantitative change (i.e., a single stage). The CMPL theory conceptualizes two stages, with the first stage characterized by performance based on general algorithms and the second stage marked by direct memory retrieval. ACT-R, on the other hand, advocates three stages, with the first stage characterized by reliance on declarative knowledge (the cognitive stage), the second marked by a mixture of both declarative and procedural knowledge (the associative stage), and the third dominated by procedural knowledge (the autonomous stage). Qualitative changes in the underlying cognitive processes drive the progression through the skill acquisition stages; for the CMPL theory, it is the shift from problem-solving algorithms to memory retrieval, and for ACT-R, the first shift is caused by knowledge compilation (by which procedural knowledge is developed) and the second is marked by the end of knowledge compilation, at which point only the long process of production tuning remains. Acceleration in skill execution is driven by two forces. On the one hand, learners accelerate within stages by gradually refining the same process. On the other hand, they can optimize the performance between stages by making qualitative changes to the underlying cognitive processes. The latter not only increases the speed of performance but also makes its procedures more efficient.

1.3.3 The Explanatory Power and Empirical Predictions of the Skill Acquisition Models

The power-law of practice and the skill specificity are two empirical phenomena that are widely observed in skill acquisition research (see Section 1.1). The three theoretical models reviewed previously provide unique explanations of how the proposed learning mechanisms give rise to the phenomena. The Race model describes that the power-law of practice results from two counteracting factors (see Logan, 1988, p. 496). As learners accumulate instances in memory, there are more opportunities to achieve extreme values of low performance times, thereby creating the speedup in performance. At the same time, the more extreme values they observe, the lower the likelihood of achieving even faster performance times, so the speedup decelerates. Since the CMPL theory is basically an extension of the Race model, it makes the same prediction for the power-law of practice. Interestingly, neither the Race model nor the CMPL theory provides an explicit description of why extensively practiced skills become less available for transfer. In this light, ACT-R offers higher explanatory power because it accounts for both the power-law of practice and the skill specificity.

Concerning the power-law of practice, Anderson and Fincham (1994) stated that the initial reduction in performance times "would reflect the compilation of the production rule [i.e., knowledge compilation], and the remaining power-raw learning would reflect the accumulation of production strength [i.e., production tuning]" (p. 1324). Knowledge compilation leads to abrupt changes in one's performance times because it fundamentally restructures the underlying mechanism of performance (see Section 1.3.1 and Table 1.1). In contrast, production tuning is a pure optimization process, so its benefit diminishes as the performance gradually approaches the optimal level. ACT-R further explains that the skill specificity results from proceduralization and production tuning (Anderson & Fincham, 1994; Anderson & Lebiere, 1998; Singley & Anderson, 1989). Once production rules are proceduralized, they become specific to the skill being practiced. Furthermore, production tuning adjusts the applicability of the production rules so that they apply in specific conditions or contexts only.

More importantly, to the aim of this dissertation, the three theoretical models propose opposing hypotheses regarding the number and the nature of skill acquisition stages. Figure 1.3 summarizes how each model accounts for the practice-induced speedup in relation to the number and the nature of skill acquisition stages. Notice that the number of stages, and hence the number of distinct cognitive states, corresponds to the number of power functions required to account for the performance speedup (see Section 1.1.1; Delaney et al., 1998; Rickard, 1997, 1999; Tenison & Anderson, 2016 for findings that each cognitive phase requires separate power functions). The Race model (the top panel) attributes any reductions in performance times to the mechanism of instance accumulation. Since there is only one mechanism improving the same cognitive state (i.e., the parallel implementation of problem-solving algorithms and memory retrieval), the model predicts that a single power function best fits experimental data. Hence, any speedup occurs within a single stage as a continuous quantitative change in the underlying mechanism (Logan 1988, 1992).

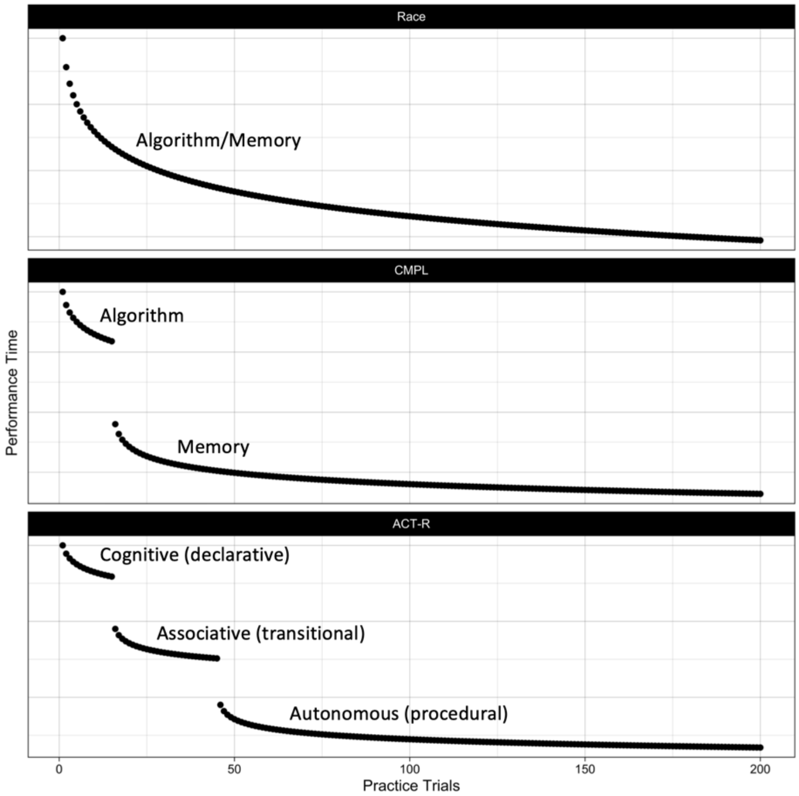


Figure 1.3. Model predictions of practice-induced speedup.
Note. The top panel shows the prediction from the Race model, the middle panel shows the CMPL theory, and the bottom panel shows ACT-R.

In contrast, the CMPL theory (the middle panel) allows for the speedup both within and between stages. Algorithmic processing and memory retrieval themselves can be accelerated (i.e., within-stage speedup), but more speedup is expected by shifting the cognitive processes (i.e., between-stage speedup). Therefore, the CMPL theory incorporates both quantitative and qualitative changes in the underlying mechanism. Two power functions, one for algorithmic processing and the other for memory retrieval, are hypothesized to best account for latency data (Rickard, 1997, 1999). Lastly, ACT-R (and Fitts and Posner's theory of learning) (the bottom panel in Figure1.3) predicts that skill acquisition is a three-stage process. Each phase (i.e., the cognitive, associative, and autonomous stage) requires separate power functions to account for the practice-induced speedup. Learners can increase the speed of performance both within and between stages, but the largest drops in performance times take place between stages by transitioning from the cognitive stage to the associative stage and from the associative to the autonomous stage (Anderson, 1982, 1983b; Tenison & Anderson, 2016).

Testing these model-based implications in empirical research is crucial for the scholarship of skill acquisition research because it allows researchers to pit rival theoretical models against each other. Recently, a line of research in cognitive psychology has attempted to test the model predictions within the framework of cognitive modeling (i.e., using computer and mathematical simulations to model human cognition). In particular, Tenison and Anderson (2016) (also Tenison et al. 2016) conducted an innovative study to investigate the tenability of the three models within cognitive skill acquisition. In the next section, I will turn to such line of research as it provides a model approach for how theoretical models of skill acquisition can be tested in L2 contexts.

1.4 Testing the Models of Skill Acquisition

Figure 1.3 in Section 1.3.3 showed that testing rival models of skill acquisition boils down to the question of whether the practice-induced speedup is caused by quantitative changes or qualitative changes in the cognitive processes underlying skill performance. The number of learning stages (or the number of distinct cognitive states) corresponds to the number of power functions required to account for latency data. However, what does it mean exactly to fit a power function to each stage of learning? Analytically, the following updated form of the power function can be used to test the number of learning stages:

$$T_{i,j} = \mathrm{I} + \beta_i j^{-\alpha}$$

In addition to the basic form of the power function discussed in Section 1.1.1 (i.e., $T = \mathrm{I} + \beta \mathrm{N}^{-\alpha}$), this formula incorporates two subscripts, i for each learning stage and j for each practice trial. Hence, $T_{i,j}$ denotes the time to perform a skill in the stage i on the trial j, and β_i indicates the slope of the power function specific to each learning stage. Hence, the number of learning stages is represented as the number of power-function slopes that are estimated. For instance, the Race model requires a single slope (β) because it assumes skill acquisition within a stage. In contrast, the CMPL theory and ACT-R claim multiple learning stages, each of which requires a different slope parameter. In (linear) regression modeling, this is analogous to having varying slopes for learning stages or treating the stages as separate dummy-coded predictor variables (e.g., $y = a + b_{\mathrm{stage1}} + b_{\mathrm{stage2}} + b_{\mathrm{stage3}} + \varepsilon$). In this light, testing the three theoretical models entails evaluating the plausibility of different power function models that incorporate varying numbers of slope parameters.

One criticism against the power-law of practice is that the power function may not apply to individual data (see Section 1.1.1). Traditionally, skill acquisition researchers have applied a power function to aggregated data after averaging raw data across individual participants and items (e.g., Anderson, 1981; Logan, 1988; MacKay, 1982; Newell & Rosenbloom, 1981). However, this practice of (only) fitting a single function over the entire sample makes an unrealistic assumption that individual learners become faster at the same rate and that every learner transitions to subsequent stages after the same amount of practice. The ideal method thus requires a type of statistical models that enables researchers to fit a power function to individual data points while at the same time evaluating the feasibility of the power-function model at the group level. Furthermore, unaggregated data are inherently subject to (more) variability, so the method must be able to detect whether the trial-by-trial speedup is due to sheer performance (or sampling) error or to systematic learning effects triggered by practice.

One candidate means to avoid modeling aggregated data is to apply a class of multilevel power-function models that incorporate varying intercepts and slopes for participants and items (e.g., mixed-effects models) (see Gelman & Hill, 2007 for how). The estimation of participant- and item-specific parameters makes it possible to test whether the form of the power function dictated by theoretical models makes sense at every level of analysis (i.e., participants, items, and the entire group); hence, it enables researchers to examine the number of skill acquisition stages at the individual and the group level within one coherent model. However, one critical limitation of the method is that it provides too little information regarding how fast individual learners move through learning stages or which stage they may reside in after a given number of practice trials. In a recent study of cognitive skill acquisition, Tenison and Anderson (2016) took advantage of a technique called *hidden Markov modeling* to overcome the limitation. The hidden Markov model (HMM) is a stochastic time-series model consisting of a Markov chain, a mathematical system that represents a sequence of states. A Markov chain makes an assumption that a given state is only dependent on the previous state (i.e., the Markov assumption). The HMM is a special type of Markov chain that treats the actual states as hidden (and hence latent) but whose probability can be estimated based on observed data. It is most commonly used as a stochastic pattern-recognition method in computational linguistics, such as research in speech recognition, where the HMM is used to segment and identify words based on temporal structures in speech as well as a database of various sounds in a variety of languages (Rabiner, 1989; see Chan, Verspoor, & Vahtrick, 2015 for using a HMM in L2 research).

Figure 1.4 provides a visual illustration of the structure of a HMM applied to skill acquisition data (see Section 2.3.2 for the mathematical rendition of the HMM used in this dissertation study). In this example, Markov states represent two learning stages. Using a vector of RTs as the dependent variable (RT_1–RT_5), the HMM provides two types of likelihood: (a) transition probability, the probability of individuals eventually moving from one state to another (i.e., from State 1 to State 2), and (b) emission probability, the probability of a data point being generated by a given state. Assuming two learning stages, the transition probability in Figure 1.4 means that learners eventually transition to the second stage with the probability of .96. In contrast, the emission probability of RT_5 in State 2, for example, is .90, which means that with the probability of .90, learners have already transitioned to the second stage at the fifth practice trial. Of particular interests to the current discussion is the emission probability because it can be used as the estimate of which learning stage participants reside in after a particular number of practice trials. The estimation of the HMM parameters considers every trajectory of learners transitioning to subsequence stages after every number of practice trials and provides the relative likelihood of the trajectories given the data.

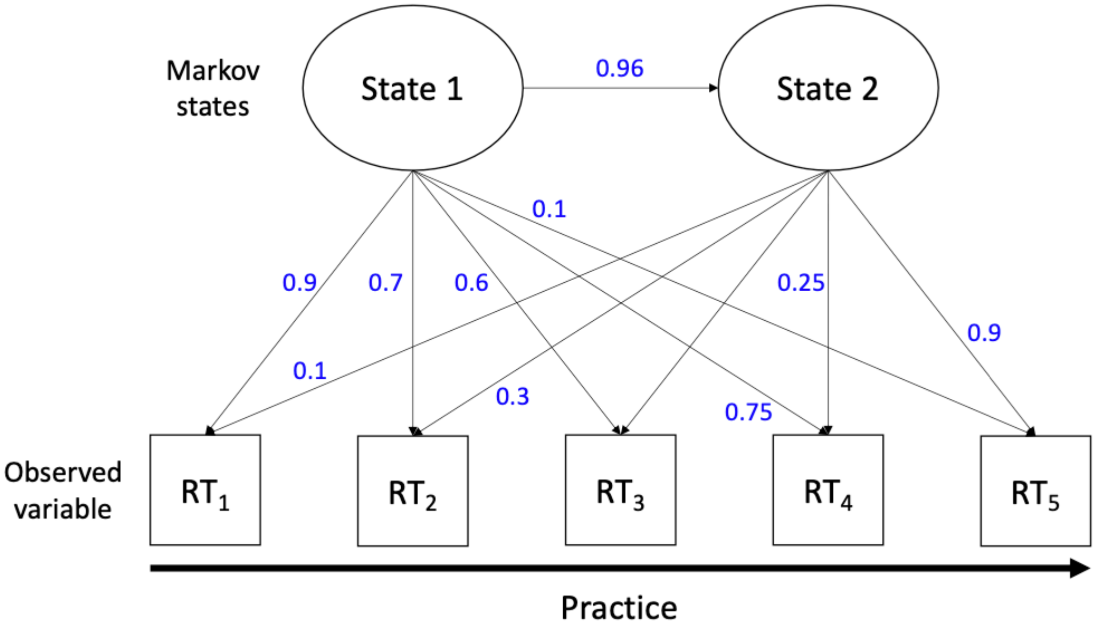


Figure 1.4. The illustration of a hidden Markov model applied to skill acquisition data.

In the study, Tenison and Anderson (2016) examined how learners developed fluency in an arithmetic task called a Pyramid problem. A typical item of the Pyramid problem takes the form of "base$height" where the base is the starting number, and the height indicates the number of terms to be summed, with each term being less than the previous one. For instance, the answer for the problem 8$3 can be found as 8 + 7 + 6 = 21. In the study, participants (*N* = 23) practiced solving 21 unique problems over six blocks of 36 trials. Of the item set, three items were chosen as practice problems that were repeated 36 times, while the remaining 18 problems were practiced only twice as novel problems. The researchers applied a HMM to the latency history of each individual participant on each practice item. As deliberated in Section 2.3.2, the HMM adopted by Tenison and Anderson (2016) was a unique HMM model that yielded the probability of each participant being in a given stage by considering how well the data are consistent with a power function that incorporated varying numbers of slope parameters (e.g., the CMPL theory requires two slopes). Since the number of slopes corresponds to the number of learning stages (see Section 1.3.3), the number of HMM states also corresponds to the number of slopes in the power function. Comparing HMMs that assumed one to five states, Tenison and Anderson found that the three-state HMM showed the best fit, while the one- and two-state HMMs were substantially less consistent with the data.

Tenison and Anderson (2016) hypothesized that the three stages identified by the HMM analysis corresponded to the cognitive, the associative, and the autonomous stage, following the three-stage theory by Fitts and Posner (1967). Specifically, they drew on the learning mechanisms proposed by ACT-R and suggested that the first stage involved a sequence of direct calculations to find the answer, the second stage involved an effortful retrieval of the past solution from memory, and the third stage involved an automatic recognition of the problem such that the solution is produced as a reflex. To validate the prediction, Tenison and Anderson collected neural signatures from an functional magnetic resonance imaging (fMRI) and examined how brain activation patterns changed as the participants transitioned from the first to the second stage and from the second to the third stage. The researchers conducted a region of interest analysis on specific regions that were hypothesized to be engaged in skill acquisition of the Pyramid problem: (a) the left horizontal intraparietal sulcus (HIPS) for numerical processing in the first stage, (b) the left lateral inferior prefrontal cortex (LIPFC) for effortful retrieval in the second stage, and (c) the left fusiform gyrus for visual recognition of stimuli and the left motor cortex for motor responding in the third stage. A regression analysis of the fMRI data showed that the left HIPS was more strongly activated when the participants were in the first stage than in the second stage, the left LIPFC was more actively involved in the second stage than in the third stage, and the left fusiform gyrus and the left motor cortex were most strongly activated in the third stage. These findings, combined with the results of the HMM analysis, provided convincing evidence that cognitive skill acquisition (at least of the Pyramid problem) is a three-stage process and that the cognitive process involved in each of these three stages is consistent with the predictions from Fitts and Posner (1967) and ACT-R (Anderson, 2007; Anderson et al., 2004, 2019; Anderson & Lebiere, 1998).

The study by Tenison and Anderson (2016) is one of the rarest studies in cognitive psychology that attempted to formally test rival theoretical models of skill acquisition using an appropriate analytical method to overcome the issues that pervaded in previous research (e.g., applying a power function to individual data points and identifying the speedup of performance due to mere performance error or systematic learning effects). In SLA, the skill acquisition theory has been one of the major theoretical approaches to explaining the process of L2 learning (DeKeyser, 2020; Lyster & Sato, 2013; Y. Suzuki, 2022). While researchers often cite and adopt various theoretical models of skill acquisition to account for L2 phenomena, the validity of the models is rarely tested vis-à-vis L2 learning. This gap in the literature thus calls for an empirical study that adopts the research design and the analytical method of Tenison and Anderson (2016) in L2 learning contexts. The aim of my dissertation study is to do just that. However, before introducing the study, it is worthwhile to review the available literature on skill acquisition in L2 learning. The literature on L2 skill acquisition mostly concerns (a) providing evidence for the parallel nature of L2 learning and skill acquisition in general and (b) investigating individual differences in L2 skill acquisition and their cognitive correlates.

1.5 Second Language Skill Acquisition

It is clear from the outset that L2 learning can be more complex than typical skill acquisition studied in cognitive psychology (e.g., arithmetic tasks such as the Pyramid problem) because successful performance in L2 requires the coordination of multiple levels of linguistic knowledge (e.g., phonology, vocabulary, morphosyntax, and pragmatics). Furthermore, L2 learning can take place in varying conditions (e.g., classroom learning through instruction vs. naturalistic learning from mere exposure or usage), which can engage different learning processes (e.g., intentional vs. incidental learning). It is thus no surprise that different approaches to L2 learning espouse different models of skill acquisition. For instance, from a usage-based perspective, N. Ellis (2002) accounted for the relationship between automaticity and frequency effects in language processing using Logan's (1988) instance theory (of which the Race model is a part). Presumably, this is due to the theory being an instance-based approach, which is compatible with the primary role of item-based learning in usage-based linguistics. In contrast, DeKeyser's (1997) seminal study on automatization of L2 morphosyntax and his accounts of L2 skill acquisition are based on the concepts in ACT-R (e.g., production rules and declarative/procedural knowledge) (see DeKeyser, 2020). Individual models (or parts thereof), therefore, should be useful to different extents, depending on different conditions and linguistic targets of learning (DeKeyser, 2001). The instance theory, for instance, may be useful to explaining incidental learning of vocabulary items, whereas ACT-R may be informative to account for learning of morphosyntax under instructed settings.

To date, there has been considerable evidence in SLA that points to the parallel nature of L2 learning and (cognitive) skill acquisition (e.g., De Jong, 2005; DeKeyser, 1997; Ferman et al., 2009; Li & DeKeyser, 2017; Robinson, 1997; Robinson & Ha, 1993; Y. Suzuki & Sunada, 2020). In this section, I review and discuss the existing literature on L2 skill acquisition. In Section 1.5.1, I focus on evidence for conceptualizing L2 learning as a type of skill acquisition. I achieve this aim by specifically discussing available evidence for the power-law of practice and the skill specificity in L2 contexts. In Section 1.5.2, I focus on individual differences in L2 skill acquisition and what kinds of cognitive correlates predict the individual differences. Such research has the potential of revealing what cognitive mechanisms underlie L2 skill acquisition.

1.5.1 Evidence for Second Language Skill Acquisition

The most extensive evidence for L2 skill acquisition comes from the seminal study by DeKeyser (1997). In the study, participants ($N = 61$) longitudinally learned and practiced how to comprehend and produce an artificial language called *Autopractan* for over eight weeks. DeKeyser specifically posed three hypotheses to find evidence for L2 skill acquisition: (a) RT and error rates (of performance) would decrease as the result of practice; (b) the decrease in RT and error rates would follow a power function; (c) resulting competence would become skill-specific such that performing in a reverse condition (i.e., comprehending the language when one extensively practiced production or vice versa) leads to more errors and slower performances. Prior to practice, participants were taught vocabulary and grammar rules of the language by means of a direct presentation, accompanied by an explicit explanation of how sentences in Autopractan can be constructed using the learned words and four case markers. Subsequently, the participants engaged in 15 sessions of comprehension and production practice (1,440 trials in total, with 720 trials for comprehension and production practice).

An observation of RT and error data showed that the decrease in RT clearly followed the power-law of practice ($R^2_{\log(y)\log(x)}$ = .974 and .966 for comprehension and production, respectively), even though the improvement in error rates was less consistent with the power function ($R^2_{\log(y)\log(x)}$ = .613 and .651 for comprehension and production, respectively). Additionally, practice led to linguistic competence that was skill-specific in that when learners were tested during the final practice session, they made more errors and performed noticeably slower in the opposite mode of language use (i.e., comprehension vs. production). Interestingly, DeKeyser also implemented a within-participant manipulation of single-task and dual-task conditions in the practice tasks so that in half of the practice trials, the participants engaged in a secondary task while simultaneously carrying out the primary task of language practice. As reviewed in Section 1.2.1, one indication of automaticity is that learners get little interference from performing a secondary (cognitively demanding) task. The results showed that participants initially performed slower and made more errors in the dual-task condition, but the difference between the two conditions disappeared in the final practice session. This finding, together with the evidence for the power-law of practice and the skill specificity, led DeKeyser to conclude that "the learning of second language grammar rules can proceed very much in the same way that learning in other cognitive domains, from geometry to computer programming, has been shown to take place" (p. 214).

Since DeKeyser (1997), many L2 studies have attested the power-law of practice and the skill specificity, but there is imbalance in the supporting evidence. Much fewer studies have explicitly tested the power-law of practice than the skill specificity (e.g., N. Ellis & Schmidt, 1998; Ferman et al., 2009; Robinson, 1997) probably because the power-law, as it is a scientific law, is already well accepted in SLA, or the exact form of how learners increase the speed or the accuracy of performance is not sometimes of primary interests to L2 researchers. Despite the limitation, Table 1.2 summarizes the available evidence for the power-law of practice in L2 learning, especially concerning the speed of performance (i.e., $\log(\mathrm{RT}) \sim \log(\mathrm{practice})$). Although the exact fit of a given power function varies from one study to another, it is clear that when learners are provided with a sufficient amount of practice (i.e., excluding Robinson, 1997, who only had 55 practice trials, which is usually not enough for automatization to take place), the applicability of the power function is robust and consistent. Furthermore, the evidence seems to hold across different types of linguistic target, whether it is the entire language, morphosyntactic structures, or vocabulary items.

Table 1.2. Summary of previous L2 research on the power-law of practice.

Study	*N*	R^2	Trials	Target
DeKeyser (1997)	61	.96-.97	720	Language
Robinson (1997)	60	.12	55	Grammar
N. Ellis & Schmidt (1998)	7	.74-.97	344	Language
Ferman et al. (2009)	8	.80-.95	1904	Language
Cornillie et al. (2017)	23	.97-.98*	264	Grammar
Hui (2020)	35	.98*	160	Vocabulary
Pili-Moss et al. (2020)	14	.92*	720	Language
Maie (2021)	40	.94	320	Vocabulary

Note. *N* indicates the sample size in the study. * indicates a reanalysis of open-access data or descriptive statistics reported in the article. When the study had multiple target grammatical structures or modes of language use (i.e., comprehension and production), the number of practice trials was divided by the number of targets; however, this was not applied to Hui (2020) and Maie (2021), who studied skill acquisition in learning of 16 vocabulary words. “Language” in the rightmost column indicates that the participants did not possess any knowledge of the language and hence learned and practiced the entire language from scratch.

Conversely, the skill specificity has been captured in many L2 studies, mostly in terms of how the learned product of practice transfers to performance in another task or in another domain (e.g., Allen, 2000; De Jong, 2005; DeKeyser & Sokalski, 1996; Keating & Farley, 2008; Li & DeKeyser, 2017; Morgan-Short & Bowden, 2006; Y. Suzuki & Sunada, 2020; Toth, 2006; VanPatten & Cadierno, 1993a, b). This line of research can be traced back to the classic debate on the superiority of comprehension versus production practice in facilitating L2 learning. VanPatten and Cadierno (1993a, b) and Cadierno (1995) initially showed that comprehension practice (labeled as processing instruction in the original studies) led to increase in both comprehension and production skills, while production practice (traditional instruction in the original studies) only led to development in production skills (see also VanPatten, 2020 for a discussion). Later, DeKeyser and Sokalski (1996) pointed out methodological limitations in the previous studies, including an instructional design that unfairly favored input practice groups and used a highly narrow operationalization of production practice that was implemented as a type of mechanical practice (i.e., practice that can be completed without understanding semantically what one is saying) (see DeKeyser, Salaberry, Robinson, & Harrington, 2002 for a more theoretical discussion). DeKeyser and Sokalski showed that when such methodological limitations are taken into account, both practice types gave rise to the skill specificity: comprehension practice is most useful in developing comprehension skills and production practice is most effective in developing production skills. A later meta-analysis by Shintani, Li, and Ellis (2013) points to the same finding.

In summary, there is copious evidence for the power-law of practice and the skill specificity in L2 learning, hence showing the plausibility of conceptualizing L2 learning as a type of skill acquisition. What is lacking, however, is empirical research that tests specific models of skill acquisition in L2 contexts. In his review of L2 skill acquisition research, DeKeyser (2020) alluded to this fact: “More importantly for our purposes here, not much research in the field of second language learning has explicitly set out to gather data from second language learners to test (a specific variant of) Skill Acquisition Theory” (p. 88). In the seminal study, DeKeyser (1997) held the ACT theory as “[t]he most widely accepted theory on how automaticity is brought about” (p. 196). This view is consistent with a contemporary review of L2 skill acquisition research (DeKeyser, 2020; Lyster & Sato, 2013; Y. Suzuki, 2022). In discussing the external validity of the ACT theory, Anderson (1983b) believed that “all higher-level cognitive functions are achieved by the same underlying architecture” (p. 261). This claim makes language no exception. In fact, Anderson (1983b, Chapter 7) provided examples of how the learning principles in ACT-R can be used to explain the process of first language (L1) acquisition, especially focusing on how children come to learn and produce syntactic structures. However, as DeKeyser (1997, 2001) pointed out, the application of the ACT theory to L1 acquisition can be controversial because the theory maintains that all learning must start out from declarative knowledge (see Anderson & Fincham, 1994, however, for relenting that not all knowledge need to start out in declarative forms). In contrast, L2 learning is (far) more likely to involve declarative learning at the initial levels of development (see DeKeyser, 1994, 2017, for discussion) especially when learners are adults, and targets of learning are explicitly taught through instruction. In this light, L2 learning affords greater compatibility with ACT-R and serves as a better place to test its learning theory. Recently, L2 researchers have investigated individual differences in L2 skill acquisition and their cognitive correlates (e.g., Li, 2017; Maie, 2021; Pili-Moss et al., 2020; Y. Suzuki, 2018). Of typical interests to this dissertation study is a group of studies that investigated the role of declarative and procedural memory ability in L2 skill acquisition. Studying the relationship between individual differences in L2 skill acquisition and declarative and procedural memory ability has the potential of revealing whether those cognitive abilities are involved in L2 skill acquisition and hence testing the applicability of the ACT-R learning mechanisms in L2 learning (see DeKeyser, 2012 for a rationale).

1.5.2 Individual Differences in Second Language Skill Acquisition

The different learning mechanisms proposed by the Race model, the CMPL theory, and ACT-R entail different predictions for cognitive abilities that are active at each stage of skill acquisition. The Race model predicts that skill acquisition is a one-stage process, and the development of automaticity results from the accumulation of instances. Although the Race model does not fully define what constitutes an "instance", it certainly assumes that an instance is "represented as a *processing episode*" (Logan, 1988, p. 495). Hence, learning must be controlled by cognitive processes that are dependent on declarative memory, given that declarative memory subsumes episodic memory (see Eichenbaum, 2017; Squire & Wixted, 2011). The same reasoning also extends to the CMPL theory in that the transition from the first to the second stage is dependent on declarative memory. On the other hand, ACT-R presupposes three stages, with the transition from the first to the second stage controlled by declarative memory and the transition from the second to the third stage controlled by procedural memory. Declarative memory is a long-term memory system that stores factual information such as episodic and semantic knowledge. On the other hand, procedural memory is one of the *nondeclarative* memory systems specializing in encoding, storing, and retrieving procedures for performing various types of skills. While declarative learning is conceptualized as (primarily) conscious, rapid, and categorical, procedural learning is known to be unconscious, incremental, and probabilistic (see Eichenbaum, 2017; Squire & Wixted, 2011).

Drawing on Fitts and Posner's three-stage theory of learning, Ackerman (1988, 1992) made the first attempt in the cognitive psychological literature to theorize and empirically test the role of (cognitive) individual differences in skill acquisition. Ackerman posited that three sets of cognitive abilities underlie each stage of skill acquisition:

1. The cognitive state is associated with demands on general intellectual abilities.
2. The associative stage is associated with demands on perceptual speed abilities.
3. The autonomous stage is associated with demands on psychomotor abilities.

General intellectual abilities are cognitive abilities that pertain to declarative, effortful, and attentional learning, such as general intelligence, declarative memory, and working memory. Perceptual speed abilities are used when one must develop a simple procedure for skill performance, including procedural memory, statistical learning ability, and classical conditioning. Finally, psychomotor abilities represent “individual differences predominantly in the speed of responding to test items with little or no cognitive processing demands” (Ackerman, 1988, p. 291). The psychomotor abilities differ from the perceptual speed abilities in that they only concern psychophysical limitations in the subject’s motor programming, which are largely independent of information processing. In a series of experiments, Ackerman and his colleagues examined the degree of correlations between the latency of skill execution and the three sets of abilities at each block of practice. They found evidence for a dynamic interplay of the cognitive abilities and how each becomes more or less dominant depending on the learning stages (e.g., Ackerman, 1987, 1988, 1992; Ackerman & Cianciolo, 2000; Ackerman et al., 1995).

In the field of SLA, only a few studies of individual differences explicitly adopted the skill acquisition theory as the primary theoretical framework (see Li, 2017; Maie, 2021; Pili-Moss et al., 2020; Y. Suzuki, 2018). However, there is an independent line of research that has investigated the role of two long-term memory systems, declarative and procedural memory, in accounting for individual differences in L2 achievement (e.g., Faretta-Stutenberg & Morgan-Short, 2018; Hamrick, 2015; Morgan-Short et al., 2014; Pili-Moss et al., 2020; see Buffington & Morgan-Short, 2019 for a review). Recent SLA research has collectively shown that L2 learning (especially for adults) is initially supported by declarative memory and later dominated by procedural memory (see Hamrick et al., 2018 for a meta-analysis). Although this is consistent with what ACT-R would predict, these studies were conducted independently of the L2 skill acquisition research. Rather, Ullman’s (2004, 2016, 2020) declarative and procedural (D/P) model served as the guiding framework. The D/P model is a neurobiologically motivated model of language, which claims that declarative and procedural memory underlie the learning, representation, and retrieval of different types of linguistic knowledge. Specifically, declarative memory supports learning and using L2 lexical items across all levels of proficiency, but for grammar, it is only responsible for the initial stages of learning, and procedural memory takes over as learners practice and develop proficiency. There is recent evidence in SLA that the use of vocabulary knowledge also involves procedural memory (Maie, 2021), but the general idea of initial reliance on declarative memory and the later involvement of procedural memory is consistent with the learning mechanisms in ACT-R.

While many have already examined the role of declarative and procedural memory in L2 achievement, only a handful investigated their role in developing L2 automaticity (Li, 2017; Maie, 2021; Pili-Moss et al., 2020; Y. Suzuki, 2018). The available evidence, however, seems to favor the prediction from ACT-R and the D/P model. For example, Pili-Moss et al. (2020) investigated how the two memory systems predicted the accuracy and automaticity of comprehending and producing a newly learned artificial language over extensive practice sessions (see also Morgan-Short et al., 2014 for details). Concerning automaticity (operationalized as the CV of RT), the researchers found a three-way interaction effect of practice sessions, declarative memory, and procedural memory, such that when the practice sessions were divided into three stages, procedural memory was predictive of automaticity at the two later stages, but this was contingent on whether learners had a higher declarative learning ability. Given the declarative-procedural transition proposed in ACT-R and the D/P model, this result seems reasonable as the transition becomes impossible if learners lack any declarative knowledge to proceduralize. However, this study did not collect RT data for production practice and only addressed skill acquisition specific to comprehension skills. The researchers, furthermore, divided practice blocks into three stages on an arbitrary basis, which precluded them from contrasting claims about the number and the nature of skill acquisition stages.

The current dominant view in SLA is to regard L2 skill acquisition as a three-stage process. However, the theory must be formally tested in L2 contexts before one can put trust in the skill acquisition theory (and variants thereof) as applied to L2 learning. This is critical because the three-stage model (or part thereof) is already represented in many subdomains of SLA (e.g., language instruction: DeKeyser, 1998; 2001; Lyster & Sato, 2013; language assessment: ACTFL, 2012; Council of Europe, 2020), but the model itself is currently only theoretical and lacks empirical support. In this dissertation research, I investigated the number and the nature of skill acquisition stages in L2 learning. The number of skill acquisition stages was tested by adopting the analysis methodology of Tenison and Anderson (2016) in an empirical study of L2 learning that longitudinally examined how L2 learners developed accuracy and fluency in a novel language as a result of practice. The nature of stages, or distinct cognitive states active at each stage of learning, was tested by adopting the research design of those L2 studies that investigated the role of cognitive individual differences in L2 skill acquisition (e.g., Maie, 2021; Pili-Moss et al., 2020; Y. Suzuki, 2018). Following the theoretical model of individual differences in cognitive skill acquisition by Ackerman (1988, 1992) and the previous research on the role of declarative and procedural memory in L2 learning, I elected declarative memory, procedural memory, and psychomotor ability as the candidate cognitive abilities. Hence, by bringing together the three lines of research in SLA and cognitive psychology, the study attempted to provide the first direct evidence for (or against) the influential three-stage model of L2 skill acquisition. In the next chapter, I describe the research and analytical design of the study, including the research questions and hypotheses, methods, procedure, and analysis.

CHAPTER 2: THE CURRENT STUDY

2.1 Research Questions and Hypotheses

The overarching goal of the study was to test the validity of the three-stage model of skill acquisition in the context of L2 learning. If, as Anderson (1983b) claimed, “language is cut from the same cloth as the other cognitive processes” (p. 261), the process and the mechanisms of L2 learning must be explained by the cognitive models of skill acquisition reviewed in the last chapter, including the three-stage model. To investigate this claim, I conducted a language learning experiment in which participants deliberately learned and practiced a novel foreign language for an extended period of time. I asked the following questions for the study:

Research Question 1: How many stages of skill acquisition, each characterized by distinct cognitive states for learning and consolidation, do L2 learners go through while learning and practicing a novel miniature language for an extended period of time?

Research Question 2: Which cognitive abilities, declarative memory, procedural memory, and psychomotor ability, are implicated at each stage of skill acquisition?

Research Question 3: Do the results for the number (RQ1) and the nature (RQ2) of skill acquisition stages in L2 learning differ between comprehension and production of the target language?

I hypothesized that L2 learning consists of three stages, given the dominant view of L2 skill acquisition as a three-stage process (RQ1) (DeKeyser, 2020; Lyster & Sato, 2013; Y. Suzuki, 2022). Assuming that the three-stage model best encapsulates L2 skill acquisition, I predicted that individual differences in declarative memory, procedural memory, and psychomotor ability would manifest themselves at the first, second, and third stage of learning, respectively (RQ2). This is to follow Ackerman’s (1988, 1992) theory and predictions based on

ACT-R and the D/P model. Furthermore, because learners in this study were trained on an entire language, beginning with explicit-deductive instruction on vocabulary items and morphosyntactic rules, I expected that the learning mechanisms proposed in ACT-R would show the highest consistency with how learners learn to comprehend and produce the target language (DeKeyser, 1997, 2001). Although there is good evidence that procedural knowledge for comprehension skills does not transfer well to production skills (or vice versa) (De Jong, 2005; Li & DeKeyser, 2017; Y. Suzuki & Sunada, 2020), the mechanisms underlying L2 skill acquisition must be identical regardless of the mode of language use. Hence, the same (or at least similar) results should be observed for comprehension and production of the language (RQ3).

2.2 Methods

2.2.1 Participants

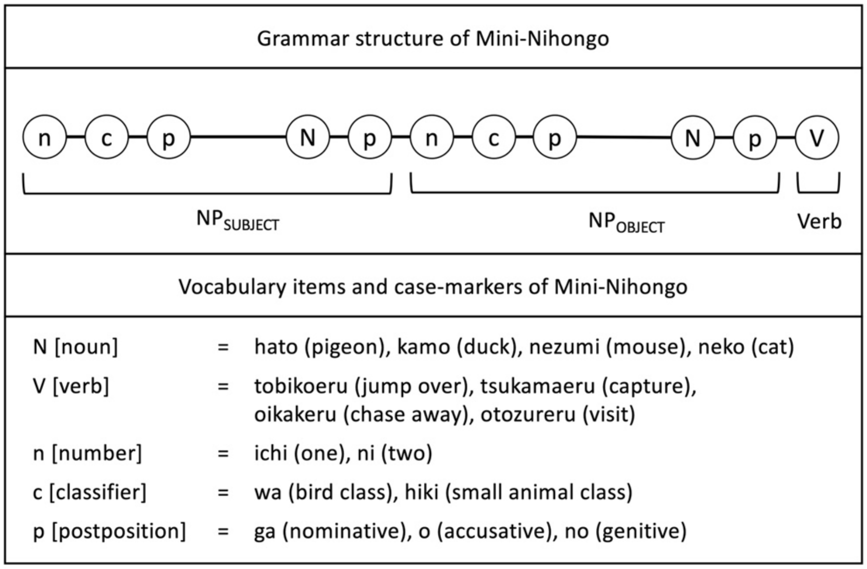
Seventy-three participants whose L1 was English were recruited to participate in the study. In total, eight participants were excluded from the analysis (i.e., 10.75% attrition) because they either did not complete the entire study (i.e., six data collection sessions) or provided responses that were psychophysically implausible for the experimental tasks at hand. For instance, one participant produced responses with RT lower than 300 milliseconds (ms) in 71.30% (1,004/1,408 trials) of the final session of production practice. With the mean accuracy of 54.75%, I deemed that this participant did not pay close attention to the task because L1 speakers, whose processing is highly automatized, take at least 300 ms to recognize a word and then manually produce a response (Jiang, 2012). After discarding data from such participants, 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, *Min* = 18, *Max* = 30). Additionally, five participants were excluded for production data because they did not properly perform production tasks in terms of the accuracy and the speed of performance (while their performance on comprehension tasks did not pose a problem). Hence, the dataset for production practice only consisted of 60 participants. The sample size is comparable to that of L2 skill acquisition research in general (e.g., De Jong, 2005 with *N* = 59; DeKeyser, 1997 with *N* = 61; Robinson, 1997 with *N* = 60)

Due to the nature of the target language, I only invited those participants who had not learned or studied any case-marking languages, such as German, Greek, Korean, Russian, and Turkish. On average, the participants knew 1.2 additional languages (*SD* = 0.79, *Min* = 0, *Max* = 4), including Spanish (*n* = 46), French (*n* = 11), Mandarin Chinese (*n* = 4), American Sign Language (*n* = 4), Arabic (*n* = 3), Italian (*n* = 3), Portuguese (*n* = 2), Albanian (*n* = 1), Latin (*n* = 1), Punjabi (*n* = 1), and Yiddish (*n* = 1). All participants were recruited through the Registrar Data Request System (https://reg.msu.edu/Forms/DataRequest/DataRequest.aspx) at Michigan State University, which distributed recruitment emails to eligible participants. The participants contacted the researcher to indicate their interest in participating in the study, and after confirming that they were indeed eligible, I sent them a URL link to access the experiment (because the study was held online: see Section 2.2.3). The participants received 60 U.S. dollars upon the completion of the entire study.

2.2.2 Language

Due to the nature of the study facing limited budget and resources, it was necessary to find a linguistic target that could be trained to automaticity within a reasonable duration of time. I chose a miniature version of Japanese, called *Mini-Nihongo* (translated as "Mini-Japanese"), that was originally constructed by Mueller and colleagues (Mueller, 2006; Mueller et al., 2005, 2007). The researchers have shown consistent evidence that grammatical and semantic violations in Mini-Nihongo elicit ERP (i.e., even-related potential) signatures that are identical to those that L1 Japanese speakers show when comprehending Japanese. The language thus preserves an appropriate level of complexity and naturalness in its grammatical and semantic system while at the same time allowing L2 learners to be trained to advanced proficiency within a relatively brief period of time. The decision to use part of a natural language (i.e., Japanese) rather than an artificial language such as Autopractan (DeKeyser, 1997) or Broncanto2 (e.g., Pili-Moss et al., 2022) was that learning a natural language (or part thereof) affords some practical value, which in turn may motivate the participants to engage in the study.

Figure 2.1 illustrates the entire structure of Mini-Nihongo adopted in the study. Most features of the language in the original form were kept, except that (a) two verbs (out of four), *tsukitobasu* (push away) and *oiharau* (chase away), were replaced with *tsukamaeru* (capture) and *otozureru* (visit) because the original verbs conveyed meanings that were hard to represent with pictures or overlapped with those of the other two verbs; (b) an adjective modifier, *akai* (red), was removed to equalize the length of the sentences across the entire item set; and (c) a temporal adverbial, *tokoro desu* (about to take place), was dropped because it added unnecessary complexity to the target language.



Vocabulary items and case-markers of Mini-Nihongo

N [noun]	=	hato (pigeon), kamo (duck), nezumi (mouse), neko (cat)
V [verb]	=	tobikoeru (jump over), tsukamaeru (capture), oikakeru (chase away), otozureru (visit)
n [number]	=	ichi (one), ni (two)
c [classifier]	=	wa (bird class), hiki (small animal class)
p [postposition]	=	ga (nominative), o (accusative), no (genitive)

Figure 2.1. The entire structure of Mini-Nihongo.

Mini-Nihongo used in this study consisted of five grammatical categories: four nouns, four verbs, two numerals, two numerical classifiers, and three postpositions (see Figure 2.1). Although Japanese in general allows scrambling of word order, I only used the Subject-Object-Verb order, which is canonical in Japanese. Hence, a sentence in Mini-Nihongo always contained two noun phrases followed by a main verb. The first noun phrase corresponded to the grammatical subject and the second to the object. A noun phrase consisted of a case-marked head noun that was modified by a numeral and a classifier. In Japanese, number is not marked morphologically and hence must be conveyed by numerical classifiers. The choice between 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 the -*no* was the genitive marker. Numerals and classifiers must be marked by the genitive marker in order to connect to the head noun. The entire structure afforded 256 unique sentences, with four examples listed below. Each sentence was matched with a colored picture that conveyed the meaning of the sentence. Because each practice session consisted of 128 trials for comprehension and production practice, I divided the stimulus list into two sets (List A and List B) and counterbalanced the order of the stimuli across the mode of language use and the number of practice sessions. In comprehension practice, the participants thus saw sentences from List A for the first and the third session of practice and from List B for the second and the fourth session. In production practice, the participants conversely saw sentences from List B for the first and the third session and from List A for the second and the fourth session. All the stimuli and the corresponding pictures can be found at: https://osf.io/x9u6h/.

1. Ichi wa no hato ga ni hiki no nezumi o tsukamaeru.
   one [bird] [gen.] pigeon [nom.] two [small-animal] [gen.] mouse [acc.] capture.
   A pigeon captures two mice.
2. Ni hiki no neko ga ichi wa no kamo o tobikoeru.
   two [small-animal] [gen.] cat [nom.] one [bird] [gen.] duck [acc.] jump over.
   Two cats jump over a duck.
3. Ni wa no kamo ga ichi hiki no nezumi o oikakeru.
   two [bird] [gen.] [duck] [nom.] one [small-animal] [gen.] mouse [acc.] visit.
   Two ducks visit a mouse.
4. Ichi hiki no nezumi ga ni wa no hato o oikakeru.
   one [small-animal] [gen.] mouse [nom.] two [bird] [gen.] red duck [acc.] chase away.
   A mouse chased away two ducks.

2.2.3 General Procedure

The general procedure of the study is summarized in Table 2.1. The study consisted of six sessions of data collection (Day 1–Day 6). In principle, the participants were required to complete the study over six consecutive days, but a two-day interval was allowed in case of emergency (see Pili-Moss et al., 2020 for the same range of intervals between study sessions). On average, the participants completed the study in 6.13 days (*SD* = 0.39). Because the entire experiment was held online on Gorilla^{TM} (https://app.gorilla.sc/), every procedural instruction within and between tasks was implemented as video instruction so that the participants fully understood what they were supposed to do (and not supposed to do).

Table 2.1. The procedure of the entire study.

Task	Min.
Day 1 (39 minutes)	
1. Background questionnaire	1
2. Two-choice response 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)	
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

Upon logging into the study, the participants were guided to the first session of the study (Day 1). They first completed an IRB-approved consent form and filled out a background questionnaire that asked about their email, age, gender, and knowledge of L1 and L2. The questionnaire is available at https://osf.io/zr9j8/. In Day 1, the participants only completed tasks of psychomotor ability and procedural memory capacity, in the order of a two-choice reaction time task (2CRT), an alternating serial reaction time task (ASRT), and a statistical learning task (SL). See Section 2.2.6 for each individual cognitive task. After completing SL, the participants were reminded that they must come back to the study on the next day for Day 2 and that they would receive a reminder email.

In Day 2, the participants first took two tasks of declarative memory, the Continuous Visual Memory Task (CVMT) and LLAMA-B, in that order. Afterward, they received explicit instruction of Mini-Nihongo by watching a 19-minute video that described and quizzed about the vocabulary and grammar structure of the language (see Section 2.2.4 for the content of the instruction). Vocabulary and grammar knowledge tests subsequently followed the instruction, which ascertained that the participants indeed developed explicit, declarative knowledge of the language. The participants were then guided to warmup practice, the purpose of which was to familiarize them with the format of comprehension and production tasks. Day 3–Day 6 had an identical structure: the vocabulary and grammar knowledge tests were administered first and then comprehension and production practice tasks. In Day 3 and Day 5, comprehension practice preceded production practice, but in Day 4 and Day 6, the order was reversed. In every beginning and end of a study session, the participants also completed a self-checklist to report that (a) they are/were in a quiet room, (b) to the best of their knowledge, they have/had a reliable internet connection, and (c) they will not/did not step away from the computer during a task. The purpose of the checklist was to remind the participants of the importance of following the criteria due to the nature of the study being offered online. The participants were informed that if they should fail to comply with the requirements (e.g., stepping away from the computer during a task for one hour), they may not be able to continue with the study. The checklist can be found at https://osf.io/69eap.

2.2.4 Language Training

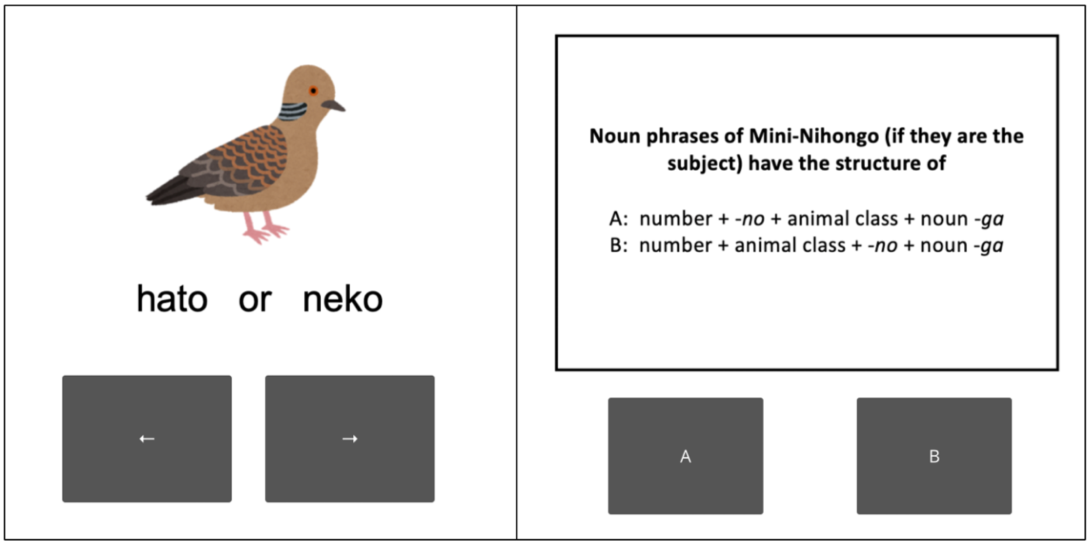
The three-stage model (represented by ACT-R) holds declarative memory as the fundamental mechanism at the initial level of learning (e.g., Anderson, 1982, 1983b, 2007). It was thus important that the participants developed declarative knowledge of the target language before engaging in comprehension and production practice. To achieve this aim, the participants received explicit instruction on vocabulary and grammar rules of Mini-Nihongo in the form of a

19-minute video (https://osf.io/vh6ap/). The instruction began with a slide showing that Mini-Nihongo is comprised of five vocabulary categories: four animal words (nouns), four action words (verbs), two words of number, two words of animal class (classifiers), and three words of grammar category (case markers). The video subsequently presented nouns of Mini-Nihongo twice by directly presenting word-picture pairs. After the first presentation, the participants were told to memorize the nouns using 30 seconds. This step was followed by a mini-exercise (four items) that asked the participants to match the words with corresponding pictures. The first presentation of the nouns was accompanied by their English equivalent (*neko* = cat), but the translation was removed in the second presentation to ensure that the participants associated the words with the pictures rather than with the L1 equivalents. The same form of presentation-exercise-presentation cycle took place for the verbs and the number words.

The instruction subsequently introduced two classifiers, -*wa* and -*hiki*, explaining that (a) the former is used to indicate a bird class and the latter to indicate a small animal class and that (b) these two words are used when combining a noun (e.g., *neko*) with a number word (e.g., *ichi*). A small exercise (four items) followed the explanation, which asked the participants to choose either -*wa* or -*hiki* depending on the picture presented. For instance, the participants chose -*wa* when they were presented with a picture of two pigeons. Lastly, three case-markers, -*ga*, -*o*, and -*no*, were presented with a description that these words (a) attach to the end of a noun and (b) indicate the subject (-*ga*) or the object (-*o*) of a sentence or the status of possession (-*no*) just like the English word "*of*" or the possessive construction "John*'s*". Afterward, all the five vocabulary categories of Mini-Nihongo were presented once again and the participants were told to review the content for one minute.

The participants then learned the phrasal and sentence structures of Mini-Nihongo. For the (noun) phrase structure, the instruction first presented all eight renditions of the noun phrase structure (i.e., *ichi*/*ni* + *wa*/*hiki* + *no* + *hato*/*kamo*/*nezumi*/*neko*) twice and asked the participants to figure out the ordering of the words. The participants were then presented a rule that a noun phrase (NP) in Mini-Nihongo takes the word order of [number] + [animal class] + [possessive marker] + [noun]. To deeper the participants' understanding of the rule, two examples were provided, showing that for instance, a noun phrase *ichi wa no hato* corresponds to [one] [bird] [of] [pigeon]. The participants then worked on a small exercise (four items) that asked them to reorder the provided words to match them with a picture. For instance, the participants saw a picture of a duck and reordered *wa*/*no*/*ichi*/*kamo* to *ichi wa no kamo*. Lastly, the participants learned how to create a sentence in Mini-Nihongo. From the outset, they were explicitly told that (a) the language has the strict S-O-V word order or NP + NP + Verb and that (b) the subject of the sentence is marked by *-ga* and the object with *-o*. The participants took 30 seconds to process and memorize the rule. This step was followed by an exercise (four items) that asked the participants to reorder eleven provided words to create a sentence that corresponded to the picture presented. At the end of the instruction, the participants were reminded that it is important that they review the vocabulary and grammar rules of the language because they would be tested at the beginning of every subsequent study session.

After the instruction as well as at the beginning of every subsequent session, the participants were tested on their declarative knowledge of vocabulary and grammar rules of Mini-Nihongo. The vocabulary test dealt with the nouns and verbs of the language and was implemented as a picture-word matching task. The participants saw a picture and two words presented together and chose the word that conveyed the meaning of the picture. Each noun was paired with every other noun as distractors (12 combinations) and every verb was paired with every other verb (12 combinations), which made up a total of 24 trials (i.e., each word tested three times). The grammar test was a metalinguistic knowledge test in a fill-in-the-blank format. The participants were presented with nine metalinguistic statements (randomly ordered) that described a morphosyntactic rule of Mini-Nihongo with some portion of the statement left blank. They were asked to choose an answer from two options. Figure 2.2 shows the outlook of the vocabulary and the grammar knowledge test, and Figure 2.3 summarizes the participant's scores on the two tests. Immediately after the explicit instruction (in Day 2), the participants had already developed solid declarative knowledge of vocabulary (*M* = .95, *SD* = .06, *Min* = .71, *Max* = 1.00) and grammar of the language (*M* = .89, *SD* = .11, *Min* = .56, *Max* = 1.00). Note that although the mean on the grammar test was lower than .90, the test only contained nine items; hence, missing even one item could make the participant's score lower than .90 (i.e., 8/9 = .88). One participant scored .56 on the grammar test in Day 2, but s/he showed a steady improvement through Day 3–Day 6, with the score of .66, .88, .88, and 1.00, respectively.



Noun phrases of Mini-Nihongo (if they are the subject) have the structure of

A: number + *-no* + animal class + noun *-ga*
B: number + animal class + *-no* + noun *-ga*

Figure 2.2. The outlook of the vocabulary (left) and grammar (right) test.

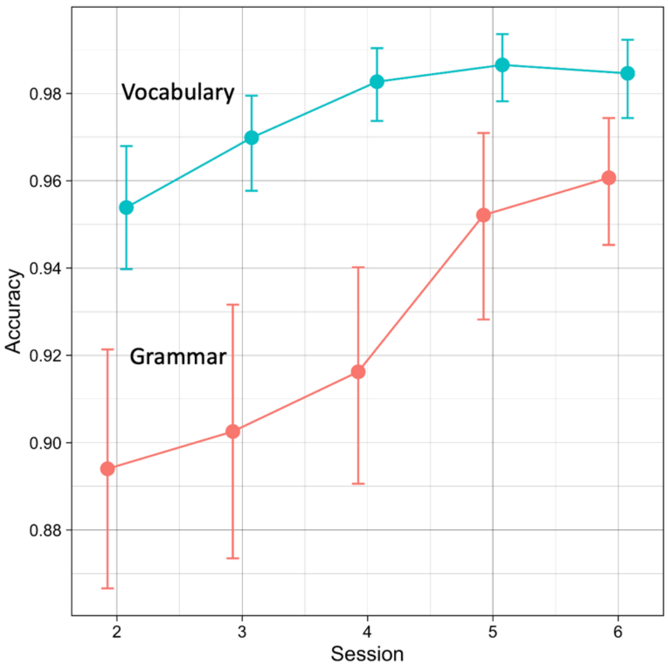
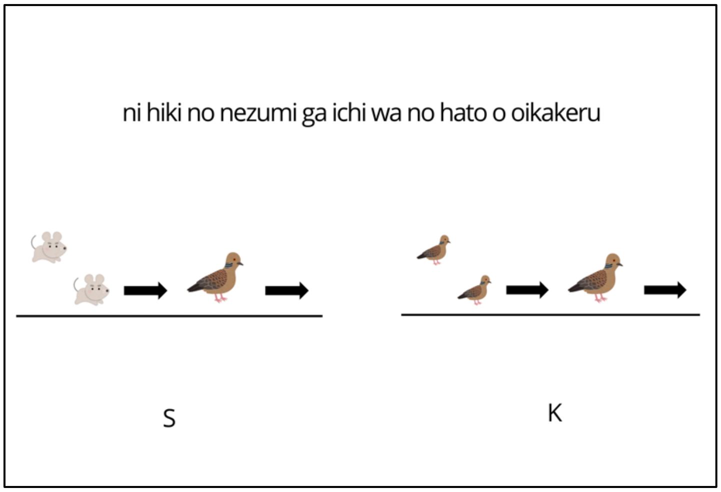


Figure 2.3. The participants' scores on the vocabulary and grammar test.
Note. The error bars show 95% confidence intervals.

2.2.5 Language Practice

In Day 3–Day 6, the participants engaged in comprehension and production practice of Mini-Nihongo after taking the vocabulary and grammar knowledge tests. The comprehension task was designed as a sentence-picture matching task in which the participants saw a sentence with two pictures and chose which picture corresponded to the sentence by pressing either the *S*-key or the *K*-key on the keyboard. Figure 2.4 shows the outlook of the task. Although DeKeyser (1997) implemented a four-picture format (instead of two) for his comprehension task to ensure that the participants fully read the sentence before making a decision, I deemed that this design was not feasible for my study because doing so for Mini-Nihongo at least required six picture options, which was likely confusing and too demanding to the participants. In this study, the two options were chosen by contrasting the pictures in terms of either (a) the subject noun, (b) the number on the subject, (c) the object noun, (d) the number on the object, or (e) the verb. Note that the word order and the case markers could not be tested directly because their positions were fixed in Mini-Nihongo. However, making a correct decision in the task required the participants to understand and process those features; for instance, the participants must recollect the word order (i.e., S-O-V) and the case-marking rules (i.e., -*wa* markers the subject and -*o* marks the object) to understand which noun phrase in the sentence corresponded to the subject or the object. The presentation of test items was randomized, with the five critical features evenly distributed throughout the task. In addition, the position of the correct answer (versus the distractor) was randomized. Each trial began with a fixation cross that was presented for 250 milliseconds (ms), and it subsequently turned into a test item. The participants received correctness feedback throughout the task.



ni hiki no nezumi ga ichi wa no hato o oikakeru

Figure 2.4. The outlook of the comprehension practice task.

The production task was implemented as a maze task (see Forster, Guerrera, & Elliot, 2009 for a review). A maze task, first introduced by Freedman and Forster (1985), is an online measure of incremental sentence processing that asks test-takers to build a sentence by choosing from a series of two alternative options as if they were going through a maze. For instance, in a five-frame trial, one sees four two-alternative options: A / bird * play / is * our / think * singing / a song * beautiful. Figure 2.5 shows the outlook of the maze task adopted in the current study. At each frame, the participants selected the continuation that best represented the pictured event. Since the numerals and classifiers had only two variants, one was always the distractor for the other (i.e., *ichi* and *ni*, *wa* and *hiki*). Each noun, verb, and case marker was paired with every other word from the same category with equal frequency. The position of the correct answer (versus the distractor) was randomized except for the number words and the classifiers because they only had two options. As in the comprehension task, trials in the production task began with a fixation cross that remained on the screen for 250 ms. Note that because Mini-Nihongo sentences consisted of eleven words, a single trial always consisted of a collection of 11 responses. For each response, the participants received correctness feedback.

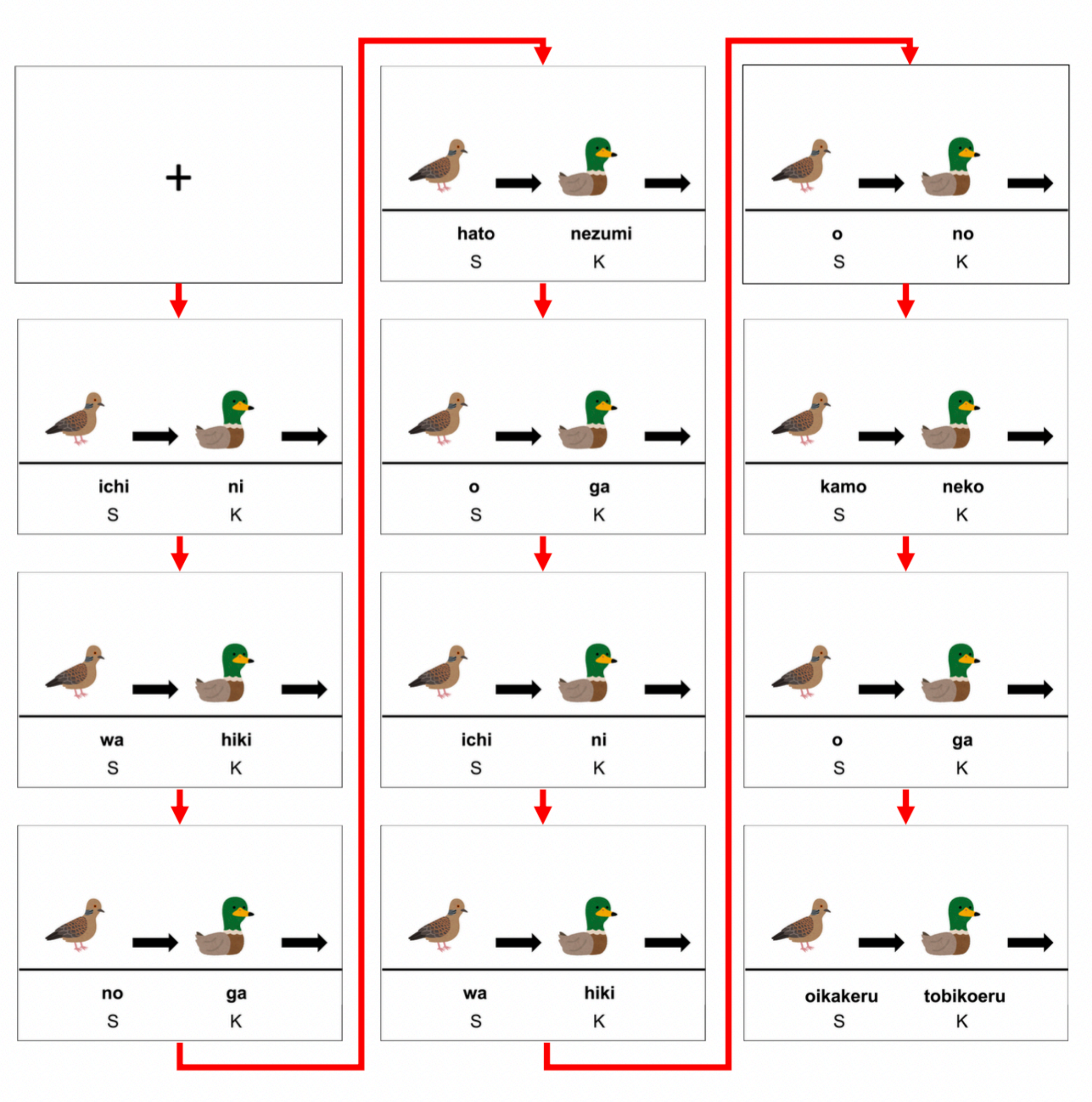


Figure 2.5. The outlook of the production practice task.

Traditional maze tasks (as used in psycholinguistic research) do not directly pertain to production skills. However, the way the task was adopted in the study rendered it (more) applicable to assessing production skills because the participants chose one of the two options to match a picture prompt rather than choosing according to whether a given option is grammatically correct or incorrect. These features are in contrast with ordinary maze tasks, for which distractors are chosen such that they are always grammatically and semantically impossible. In Figure 2.5, both *ichi* and *ni* (in the first frame) are grammatically plausible, but the latter does not match the content of the picture. Hence, the maze task in this study can be considered a type of controlled production tasks. In recent L2 research, Y. Suzuki and Sunada (2018) successfully adopted a maze task to gauge L2 speakers' sentence processing speed and automaticity. Furthermore, a large-scale study by S. Suzuki and Kormos (2022) demonstrated that RT measured by a maze task correlated well with L2 speakers' oral fluency. Through a structural equation modeling analysis, they found that RT in the maze task was the best indicator of learners' general processing speed (as part of cognitive fluency), which, in turn, was the best predictor of an aspect of (utterance) fluency termed speed fluency (indicated by articulation rate and mean length of run). Based on the findings, together with how the task was designed in the study, I deemed that the maze task was a useful measure of how the participants developed production skills. Adopting a maze task for production practice also obviated the need to consider when to start the timer of RT measurement, an issue often debated in L2 psycholinguistic research (Jiang, 2012).

Before the participants engaged in the main practice sessions in Day 3–Day 6, they were guided to initial warmup practice (in Day 2). This phase served as a familiarization period during which the participants became used to the format of the comprehension and production tasks. The warmup began with trials on individual words (i.e., nouns and verbs) and noun phrases ([number] [animal class] [possessive] [noun]), but the task then progressed to full sentences. When practicing on individual words or noun phrases for comprehension, the participants saw a word (or a noun phrase) together with two pictures and responded by choosing the picture that matched the word (or the noun phrase). For production practice, they saw a single picture and constructed a sentence (or chose a word/noun phrase) through the maze task (see Figure 2.5). For both comprehension and production warmup practice, there were 24 word-level trials (eight content words repeated three times), eight phrase-level trials, and 16 sentence-level trials. Identical items were used for both comprehension and production tasks. The entire item set for the warmup practice tasks can be found at: https://osf.io/x9u6h/.

In each main practice session (Day 3–Day 6), the participants practiced comprehending and producing Mini-Nihongo sentences for 128 trials (i.e., 256 trials in total) in 8 blocks of 16 trials. After each block, the participants were allowed to take a 3-to-5 minute break. Combining all four practice sessions in addition to the warm practice (32 trials), the participants thus practiced Mini-Nihongo for a total of 1,056 trials in 33 blocks. All participants practiced in the same list of items, but the order of presentation was randomized within the same list. In each block, the participants also encountered a surprise attention-check trial where they were asked to press the SPACE bar to continue with the task. This trial was intended to maintain the participant's focus on the task and to detect if one was mindlessly pressing the response keys (i.e., *S*-key or *K*-key) without paying attention to the task.

2.2.6 Cognitive Individual Differences

In this study, I focused on declarative memory, procedural memory, and psychomotor ability as three dimensions of cognitive abilities that have been theorized to underlie the acquisition of cognitive skills (Ackerman, 1987, 1988, 1990, 1992; Ackerman & Cianciolo, 2000; Ackerman et al., 1995). By extension, I predicted that they also play pivotal roles in L2 skill acquisition (Li, 2017; Maie, 2021; Pili-Moss et al., 2020; Y. Suzuki, 2018). There were two tasks for each ability dimension: the Continuous Visual Memory Task (CVMT) and LLAMA-B for declarative memory capacity, an alternating serial reaction time task ($ASRT_{15}$) and a statistical learning task (SL) for procedural memory capacity, and a two-choice reaction time task (2CRT) and the first block of ASRT (ASRT$_1$) for psychomotor ability. Within each dimension, I chose one domain-general (non-linguistic) and one domain-specific (linguistic) task. Table 2.2 summarizes the cognitive ability tasks.

Table 2.2. The summary of the cognitive ability tasks.

	Task	Ability	Domain
1.	CVMT	Declarative	General
2.	LLAMA-B	Declarative	Specific
3.	ASRT$_{15}$	Procedural	General
4.	SL	Procedural	Specific
5.	ASRT$_1$	Psychomotor	General
6.	2CRT	Psychomotor	Specific

2.2.6.1 The Continuous Visual Memory Task

CVMT is a test of one's ability for nonverbal declarative learning using a visual recognition paradigm (Trahan & Larrabee, 1988). The original CVMT consists of four phases (practice, acquisition, delayed recognition, and visual discrimination), but I only adopted the first two phases, which is typical of how L2 researchers have used the task (see Buffington & Morgan-Short, 2019). During the task, the participants saw a series of complex abstract designs, and they were tested on their ability to recognize seven target designs that were repeated among the other distractor designs. The task began with 11 practice trials (the practice phase) followed by 112 test trials (the acquisition phase). Of the 112 test trials, 49 were the seven target items that were presented seven times, and the remaining 63 trials were distractors that appeared only once. The order of presentation was the same as that of the original version of the task. The participants indicated whether they had seen the design (old) or not (new) in the sequence by pressing the *S*-key (old) or the *K*-key (new). Each design was only visible for two seconds, but the participants were able to respond any time later. Figure 2.6 shows the outlook of the task.

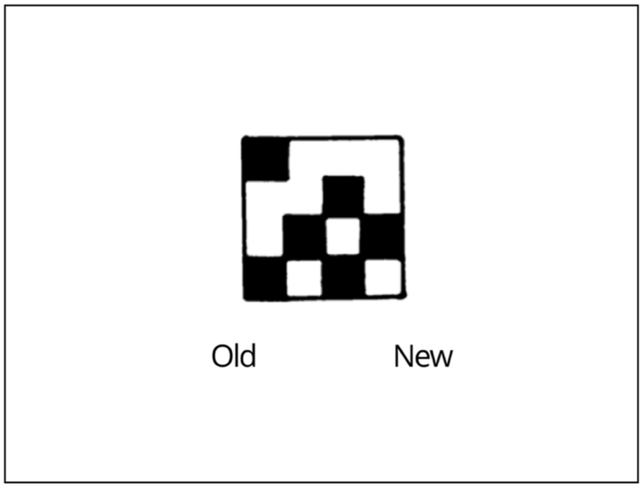


Figure 2.6. The outlook of the Continuous Visual Memory Task.

In CVMT, learning ability is quantified using *d*-prime scores. *d*-prime is a sensitivity index that operationalizes one's ability to discriminate signals from noise. In the current study, the statistic operationalized the participants' ability to discriminate old items from new items. I calculated *d*-prime scores by subtracting the *z*-score for the proportion of old 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 s new items (i.e., hits). The effective limit of *d*-prime scores is ±4.65. The internal consistency of the task based on the Kuder-Richardson Formula 20 (KR-20) was .72 [.63, .82].

2.2.6.2 LLAMA-B

LLAMA-B is a vocabulary learning subtest within the LLAMA language aptitude tests (Meara & Rogers, 2019). In L2 research, it has been used as a domain-specific (or language-based) measure of learner's declarative memory ability (e.g., Hamrick, 2015; Saito et al., 2022). The task assesses one's ability to learn the name of unfamiliar objects in two phases: the study phase and the testing phase (see Figure 2.7). In the study phase, the participants were presented with 20 unfamiliar objects with their names presented right beneath the objects. Given 2 minutes, the participants were asked to memorize the association between the names and the objects. The original version of the task implements 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, I presented an array of objects with their names together in a single screen (see Figure 2.7, the left panel). This presentation format was similar to that of Part V of the Modern Language Aptitude Test, a conceptual model of LLAMA-B (see Bokander & Bylund, 2022; Rogers et al., 2017 for reviews). In the testing phase, the participants were tested regarding how many of the associations they were able to recollect. For each trial, an object name appeared at the bottom of the screen (see Figure 2.7, the right panel). The participants indicated which object corresponded to the name by clicking the picture of the object. Instruction of the task explicitly stated that the participants were not allowed to take any notes, and if they could not find the object or did not know the answer, they could guess by clicking an object at random. The testing phase consisted of 20 items with no time limit imposed on each item. I used raw scores as the participants' declarative (and vocabulary) learning scores. The internal consistency of the task based on KR-20 was .76 [.67, .84]. LLAMA-B is publicly available at: https://www.lognostics.co.uk/tools/LLAMA_3/index.htm.

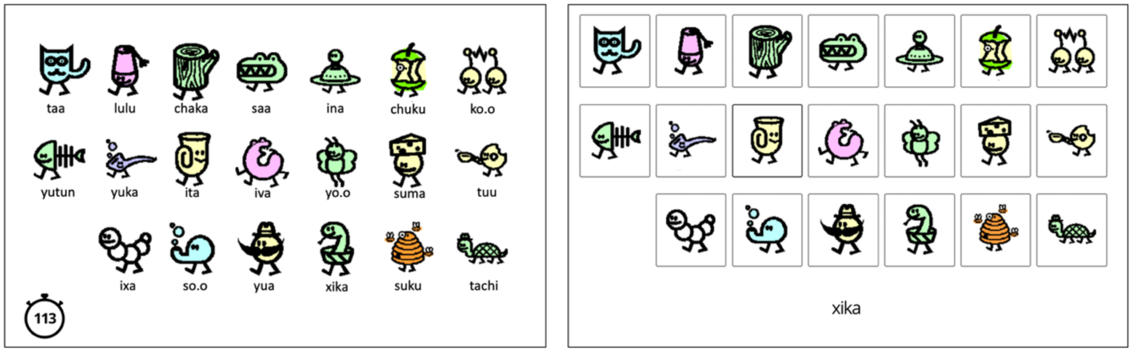


Figure 2.7. The outlook of LLAMA-B.

Note. The left panel shows the learning phase, and the right panel shows the testing phase.

2.2.6.3 Alternating Serial Reaction Time Task

The ASRT examines one's ability of implicit sequence learning (e.g., Howard & Howard, 1997; Nemeth et al., 2010). In L2 research, it is one of the most popular tasks to assess learner's procedural learning ability (e.g., Godfroid & Kim, 2021; Morgan-Short et al., 2014; Pili-Moss et al., 2020). As depicted in Figure 2.8, the participants saw an array of four circles, one of which was sequentially filled with an orange bird for each trial. The sequence in which one of the four circles was filled followed a second-order conditional rule where pattern trials were interleaved with random trials. In this study, the participants were exposed to the same sequence of `1r4r3r2`, where `r` corresponded to a random location. The participants made responses as quickly and accurately as possible by pressing the corresponding keys on the keyboard ([z] for 1, [x] for 2, [> .] for 3, and [? /] for 4, using their left middle and index fingers and the right index and middle fingers, respectively). The participants had to press a correct key to proceed to the next trial; in other words, the task did not proceed unless they produced a correct response. The task has been variably adopted by L2 researchers in terms of the amount of learning trials. For instance, Godfroid and Kim (2021) used 10 blocks of learning trials whereas Morgan-Short et al. (2014) (and also Pili-Moss et al., 2020) implemented 20 blocks. In this study, I chose a middle ground and exposed the participants to 15 blocks of learning trials. Each block consisted of 88 trials, the first eight of which were random practice trials. In total, the participants went through 600 patten trials and 720 random trials (including the practice trials).

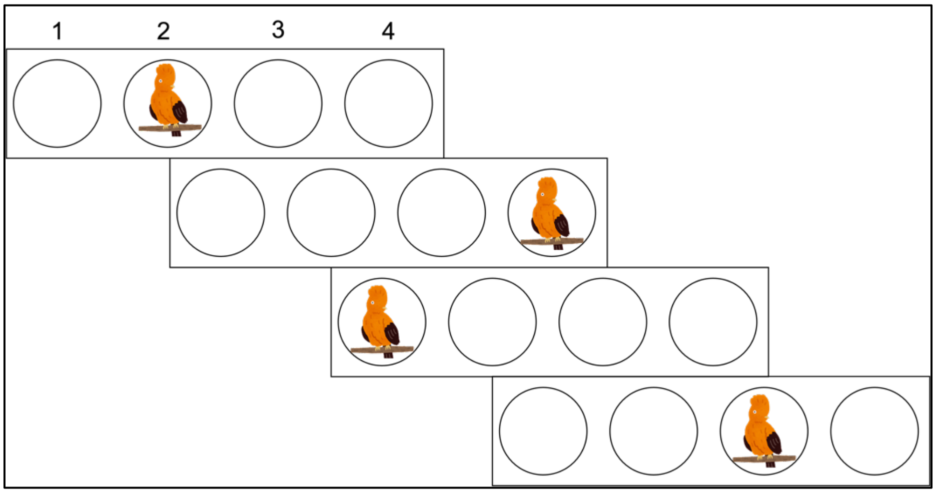


Figure 2.8. The outlook of the alternating serial reaction time task.

Learning in ASRT is often quantified by taking the difference in RT between pattern and random trials. This method of scoring yields two (overlapping) measures depending on whether one uses the entire learning blocks (e.g., Buffington et al., 2021; Morgan-Short et al., 2014; Pili-Moss et al., 2020) or the final block (e.g., Godfroid & Kim, 2021). I adopted the latter measure because Godfroid and Kim (2021) provided evidence that it is a reliable predictor of procedural (or implicit) knowledge analyzed through structural equation modeling. I took 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 responses that had RT lower than 100 ms or did not fit within the range of individual's mean ±3*SD* were removed from the analysis (1.6% of the dataset). Typically, L2 researchers estimate reliability coefficients for ASRT by splitting raw data (i.e., item-level responses) in two random halves and compute the correlation between the two halves using the Spearman-Brown prophecy formula. However, this method is not appropriate because the actual learning scores are calculated based on aggregated means rather than raw data. I thus simply took the mean of the participants' RT on the pattern and the random trials and calculated Cronbach's alpha with two items (i.e., RT on the pattern and the random trials). The internal consistency of the scores was α = .98 [.98, .99].

Finally, I calculated the mean of the participants' RT in Block 1 as a measure of the their psychomotor ability (ASRT_1). The internal consistency of the scores based on Cronbach's alpha (using the mean RT on the pattern and the random trials) was .98 [.98, .99].

2.2.6.4 Statistical Learning Task

A statistical learning task based on language(-like) stimuli typically examines one's ability to learn either adjacent (e.g., Aslin, Saffran, & Newport, 1998; Saffran, 2002; Thompson & Newport, 2007) or non-adjacent dependencies (e.g., Gómez, 2002; Gómez & Maye, 2005; Newport & Aslin, 2004). However, Romberg and Saffran (2013) pointed out that learning of a language often involves the simultaneous learning of both adjacent (e.g., collocation) and non-adjacent relationships (e.g., morphosyntax); therefore, investigating the learning of adjacent and non-adjacent dependencies at the same time is most conducive to examining a statistical learning ability relevant for language learning. In this study, I thus adopted a statistical learning task used in Romberg and Saffran (2013, Experiment 1). The target stimuli followed those of Gómez (2002), 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 categorial 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, as shown in Figure 2.9. Specifically, the X words were grouped into four groups (of four words): X_{ED}, X_{HP}, X_{LP}, and $X_{\text{unattested}}$. X_{ED} were evenly distributed words that occurred with the same probability for each

A_B frame. For each X word, there was one A_B frame for which it was X_{HP} (high probability words), one frame for which it was X_{LP} (low probability words), and one frame for which it was $X_{unattested}$ (unattested). In each frame, X_{HP} words occurred four times more frequently than X_{LP} words, and X_{LP} and X_{ED} words were equally frequent. $X_{unattested}$ word were not instantiated in the frame. The entire stimulus set can be found at: https://osf.io/cdy8b.

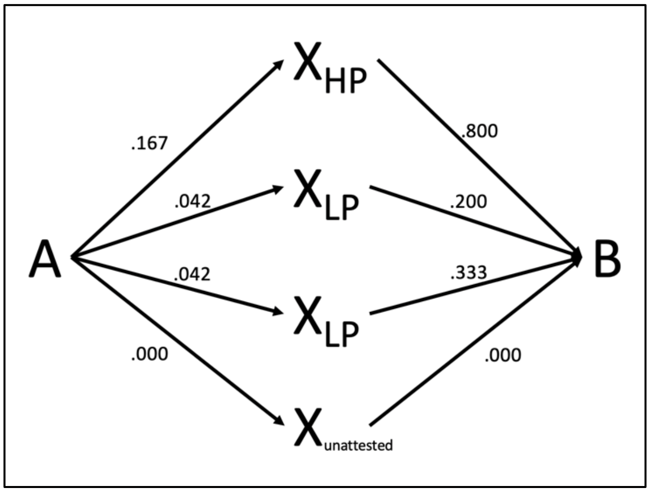


Figure 2.9. The stimulus structure in the statistical learning task adopted from Romberg and Saffran (2013).

In Romberg and Saffran (2013), the modality of the task was auditory following Gómez (2002). However, I adapted the task using visual stimuli to make the task consistent with the other tasks in the current study (including the practice tasks), which were all visually based. One issue associated with visual statistical learning tasks is that participants tend to perform better in the visual mode. For instance, Onnis, Christiansen, Chater and Gómez (2002) showed that participants, when exposed to a list of stimuli containing non-adjacent dependencies, on average scored more than 10% higher in the visual mode than in the auditory mode. This difference was likely due to the fact that the visual presentation of stimuli makes the target rule (whether hidden or not) more salient than otherwise. This feature of visual modality can be potentially problematic to statistical learning tasks because statistical learning is, by nature, an implicit process (see Monaghan, Schoetensack, & Rebuschat, 2019 for a review). To circumvent the issue, I piloted the task with four applied linguists who were familiar with some type of statistical learning tasks. The underlying logic was that if these linguists did not consciously identify the underlying rules (i.e., adjacent and non-adjacent dependencies), naïve participants in my study (who were not a linguist) would be unlikely to notice the rules.

In the task, the participants were exposed to a list of 72 three-word phrases (i.e., A-X-B) for four repetitions (288 trials in total). An interstimulus interval between each phrase was set at 750 ms, following Romberg and Saffran (2013). Subsequently, a recognition test assessed to what extent the participants developed knowledge of adjacent and non-adjacent dependencies. In the test, the participants saw two phrases in a sequence and decided which one of the two phrases sounded more familiar in that they have heard it during the familiarization phase. There were 30 items, 12 for non-adjacent dependencies, 12 for adjacent dependencies, and 6 for checking whether the participants paid attention during the familiarization phase. Both options in the non-adjacent dependency trials contained X words that were evenly distributed across the three A_B frames (i.e., X_{ED}) because the adjacent relationships between A words and X words and X words and B words could give rise to construct-irrelevant variance that is nothing to do with measuring one's ability of learning the non-adjacent dependencies. For the adjacent dependency trials, I always contrasted X_{HP} and $X_{unattested}$ to test the knowledge of the adjacent dependencies. The attention-check trials contained X words that the participants never heard during the familiarization period (*chila*, *coomo*, *nilbo*, *taspu*, *wadim*, and *wiffle*). The participants were tested in three blocks, in the order of the non-adjacent dependency trials, the adjacent dependency trials, and the attention-checking trials (see Romberg & Saffran, 2013, pp. 7–9).

However, the order of presentation was randomized within each block. The test stimuli can be found at https://osf.io/cdy8b.

After completing the recognition test, the pilot participants received a three-part questionnaire that asked them to (a) verbally report any rules they noticed during the task (i.e., retrospective verbal reports), (b) judge the familiarity of six phrases that contained non-adjacent dependencies (three correct and three incorrect A_B combinations) along with their confidence in judgment (i.e., confidence ratings), and (c) rate the familiarity of six phrases that contained adjacent dependencies (three high-frequency and three low-frequency phrase). For both confidence and familiarity ratings, the participants used a scale of 1 to 5, with 5 indicating higher confidence or familiarity. Figure 2.10 summarizes the mean confidence and familiarity ratings of the pilot participants. All four pilot participants rated correct phrases more confidently (*M* = 3.75, *SD* = 0.86) than incorrect phrases (*M* = 3.66, *SD* = 0.88) and high-frequency phrases as being more familiar (*M* = 3.91, *SD* = 1.16) than low-frequency phrases (*M* = 3.66, *SD* = 1.33). However, the difference seemed quite minimal, and considering the variance associated with each mean, it was more likely that the participants rated all phrases equivalently. Two of the four pilot participants stated in the retrospective report that the first and the third words always constituted a pair, but none of them (even after 288 trials) provided an explicit example of the A_B frames (non-adjacent) or touched upon the probabilistic relationship between X words and A/B words (adjacent). These results met the so-called zero-correlation criterion for unconscious knowledge (Dienes, Altmann, Kwan, & Goode, 1995; Dienes & Scott, 2005). However, the results also need to be interpreted with caution given that the methodological reliability of confidence ratings has been questioned (e.g., Maie & DeKeyser, 2020) and that learners tend to underreport what they consciously know in retrospective reports (e.g., Hama & Leow, 2010).

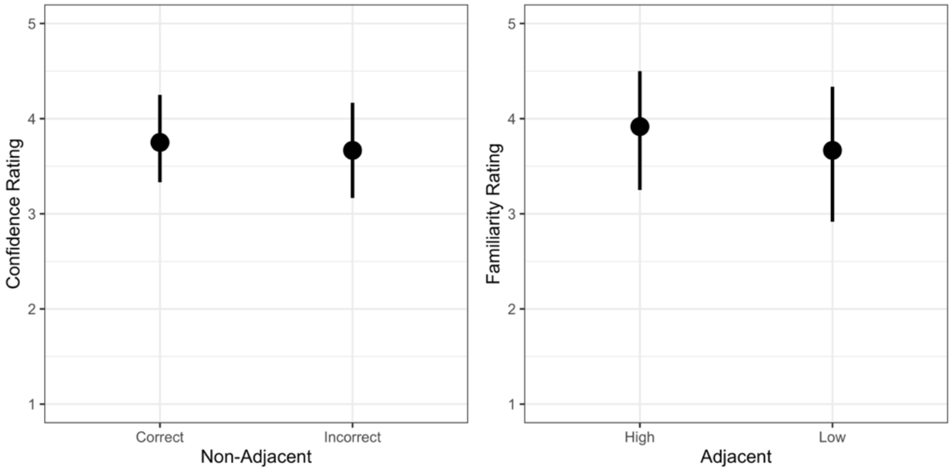


Figure 2.10. The mean confidence/familiarity ratings of the pilot participants.
Note. The error bars show 95% confidence intervals.

Romberg and Saffran (2013) did not provide an estimate of reliability for the statistical learning task. I thus calculated KR-20 for each section (non-adjacent and adjacent dependency trials) as well as for the whole test: $\mathrm{KR20}_{\mathrm{test}}$ = .70 [.60, .81], $\mathrm{KR20}_{\mathrm{NA}}$ = .70 [.60, .81], and $\mathrm{KR20}_{\mathrm{A}}$ = .44 [.25, .64]. The reliability of the adjacent-dependency section seemed low but was still on par with the reliability of procedural learning (or memory) tasks often used in L2 research (but see Perruchet, 2021 for a critical comment on the issue). I examined item-level statistics of the adjacent-dependency section to improve its psychometric quality. Specifically, I examined how removing each item changed the reliability coefficient. As a result, I excluded two items (Item 18 and 21), for removing these two items resulted in an increase in the overall internal consistency of the section. The correlation between the non-adjacent (k = 12) and adjacent sections (k = 10) was r = .32 [.09, .52], p = .007, suggesting that the abilities to learn categorical non-adjacent dependencies and probabilistic adjacent dependencies only coincide with each other to a small extent. However, scores in each section correlated strongly with scores on the entire test: r = .86 [.79, .91], p < .001 for the non-adjacent dependency section and r = .75

[.62, .84], $p < .001$ for the adjacent dependency section. In the main analysis, I thus adopted the participants' raw scores on the entire test ($k = 22$) because the total scores on the test captured the participants' scores in each section reasonably well.

2.2.6.5 Two-choice Reaction Time Task

Choice reaction time tasks are primarily used to investigate psychomotor functioning in humans and animals (see Smith, 1968; Trueman, Brooks, & Dunnett, 2021 for reviews). In contrast to simple reaction time tasks, during which one reacts to a single stimulus associated with only one response type (e.g., clicking as soon as one detects the word *bird*), choice reaction time tasks involve multiple stimuli each requiring a separate response (i.e., pressing the *S*-key when one sees the word *apple* and the *K*-key when *orange*). The task thus entails not only rapid identification of target stimuli but also accurate categorization of them depending on the responses assigned to each. Ackerman (1988, 1992) used choice reaction time tasks as measures of learner's psychomotor processing speed. In this study, I adopted a two-choice RT task (2CRT), which is the most common format of the task. The participants were randomly presented with either the word *falcon* or *eagle* and asked to press the *S*-key as soon as they recognize *falcon* and the *K*-key whenever they saw *eagle*. There were 50 experimental trials (25 trials for each word) with 10 practice trials with the words *cat* and *dog*. Each trial began with a fixation cross, which subsequently turned into the stimulus. I took an individual participant' mean RT for correctly responded trials as their individual difference score. The mean accuracy of the participants' responses was .95, $SD = .03$, 95% CI [.88, 1.00]. Responses with RT shorter or longer than an individual mean RT $\pm 3SD$ were removed from the analysis (5.63% of the dataset). I took an individual participant's mean RT for correctly responded trials as the individual difference score. The split-half reliability of the scores (divided into items for *falcon* versus *eagle*) was *r* = .94 [.90, .96].

2.3 Analysis

In this section, I describe the statistical analysis conducted to answer the research questions of the study (see Section 2.1 for the research questions and hypotheses). I will first describe how dependent and independent variables were defined and how the dataset was processed to identify and replace outlying data points. I will then review the conceptual and mathematical backgrounds of hidden Markov modeling, especially with reference to the specific hidden Markov model (HMM) adopted from Tenison and Anderson (2016). The HMM analysis was used to test the number of skill acquisition stages the participants underwent while practicing Mini-Nihongo (RQ1). Finally, I will lay out the details of regression models that were specified to investigate which cognitive individual difference variables predicted learning at each stage of skill acquisition identified by the HMM analysis (RQ2).

2.3.1 Measurement

In this study, I focused on three dependent variables: accuracy, RT, and coefficient of variability (CV) of RT. Table 2.3 presents the operational definitions of accuracy and RT as observed in each practice trial. CV was computed for each participant at the block level by dividing the standard deviation of RT by the corresponding mean. According to the three-stage model of skill acquisition, the first stage of learning (i.e., the cognitive stage) involves the reliance on declarative memory and general cognitive abilities such as problem-solving skills. In this stage of learning, it is important that learners develop accuracy of performance so they can proceduralize a correct set of behaviors (DeKeyser, 2015; Lyster & Sato, 2013; Y. Suzuki, 2022). I predicted that accuracy shows the earliest sign of learning and hence would be correlated with the participant's declarative memory capacity. CV, on the other hand, was used to operationalize the degree of proceduralization (Segalowitz & Segalowitz, 1993; see also Section 1.2.2), which is known as a necessary process for learners to proceed to the second stage (i.e., the associative stage). CV should thus be predicted by one's procedural memory capacity especially during the second stage. Lastly, after reaching the asymptotic level of performance (i.e., the autonomous stage), learners can only be distinguished in terms of their psychophysical limitations in generating motor responses. Hence, they were expected to only differ in mere RT of performance, which would be predicted by their psychomotor ability.

Table 2.3. The operational definition of accuracy and RT.

	Accuracy	RT
Comprehension	Whether or not participants chose the correct picture out of two options (0 or 1)	The time from the onset of a stimulus to the participant's response (seconds)
Production	Whether or not participants chose the correct option out of two pictures across the entire sentence. Each trial consisted of 11 word-level responses, and hence the accuracy was calculated as $\frac{\sum_{i=1}^{n=11}(0 \text{ or } 1)}{11}$ (0–1.0)	The sum of RT across 11 word-level decisions, with each response involving choosing an option out of two words (seconds)

Prior to the statistical analysis, I processed the RT data in two steps. At the trial level, I first winsorized 2.5% of the data from either end of the distribution and replaced the excluded values with the trial mean (aggregated over the participants). This step was necessary to filter out extremely fast or slow responses, given the structure of the comprehension and production practice tasks. Mean replacement is often criticized for its potential to overly inflate confidence in the mean as a summary of data (because it increases data points on the mean). I avoided the issue by using the trial mean over the entire sample rather than the mean of each participant. Figures 2.11 and 2.12 show the histogram of RT for each participant on comprehension and production practice after the winsorization was applied. Note that RT was transformed to its logarithm because it only takes positive values, which often makes its distribution positively skewed. Mean replacement was preferred over simply removing the data points because the hidden Markov modeling analysis (see Section 2.3.3) required a complete dataset; it was also preferred over replacing with the corresponding boundary value (2.5% and 97.5% point of the distribution) because the analysis tends to be sensitive to extreme values. I applied the winsorization method because there is no consensus regarding the lower boundary of how long L1 or L2 speakers take to comprehend or produce a sentence in general.

Subsequently, I computed the mean and standard deviation of RT for each participant in each block and removed any data points that were outside the range of the individual mean ±3SD. The trimmed values were replaced with the respective block mean of the participant. Finally, I removed data points on the first trial of each practice session (Trial 17, 145, 273, and 401) because the participants tended to perform unusually slow (or slower than expected) on those trials. See Figure 2.13 for the issue at hand (especially for comprehension data). When learners show this kind of regression between study sessions, it is unclear whether this is due to forgetting of the target skill or because they simply become less familiar with the experimental task at hand. Examining Figure 2.13, the initial slowness was only observed in the first trials, and the participants recovered their speed from the second trial and onwards. This suggested that the regressions were most likely due to the fact that the participants simply needed some time to briefly refamiliarize themselves with the experimental tasks. In skill acquisition research, this phenomenon is referred to as a *warmup decrement* (Adams, 1961; see also Anderson & Fincham, 1994, Experiment 2).

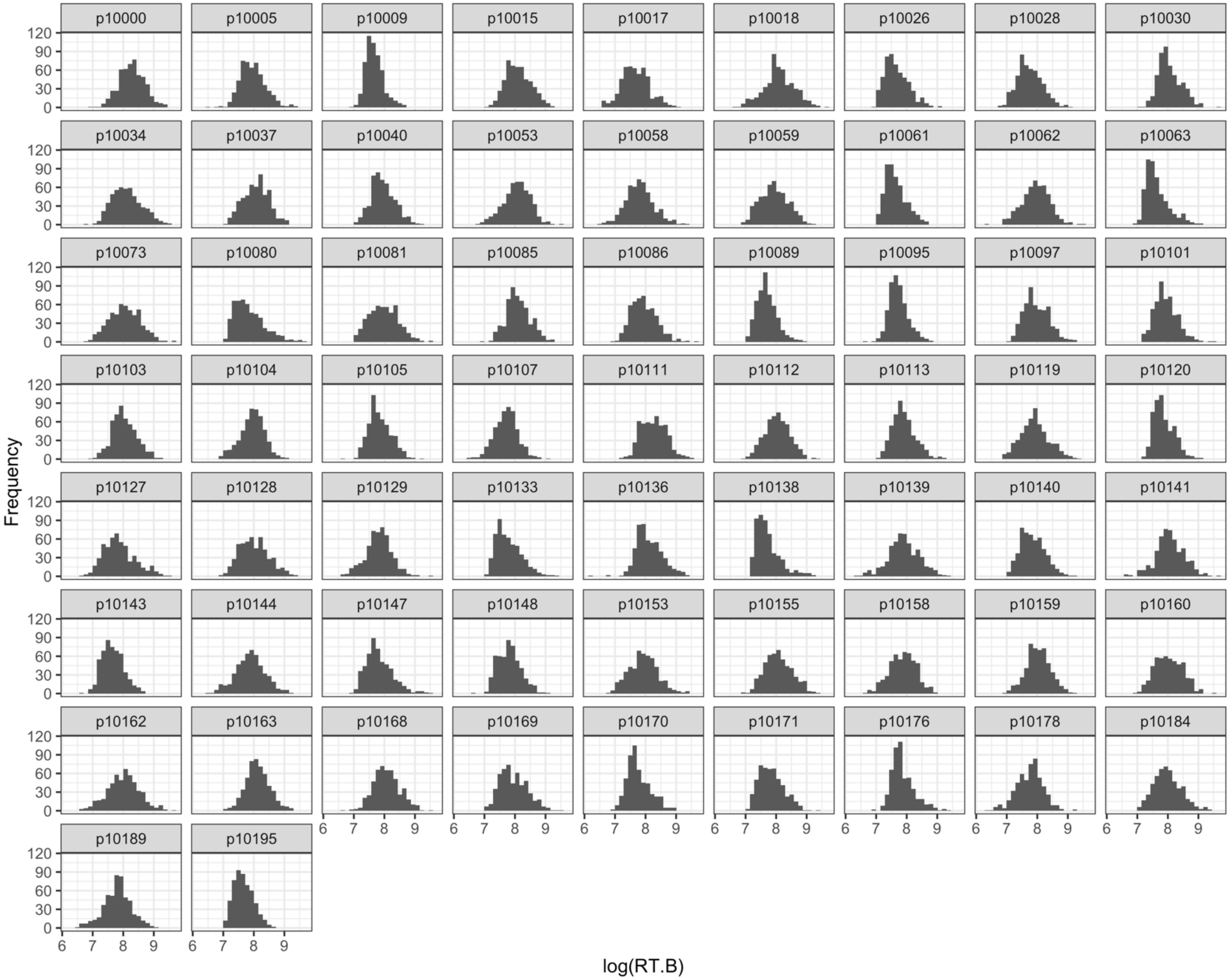


Figure 2.11. Histogram of reaction time data for comprehension practice.

Note. RT was transformed to its logarithm.

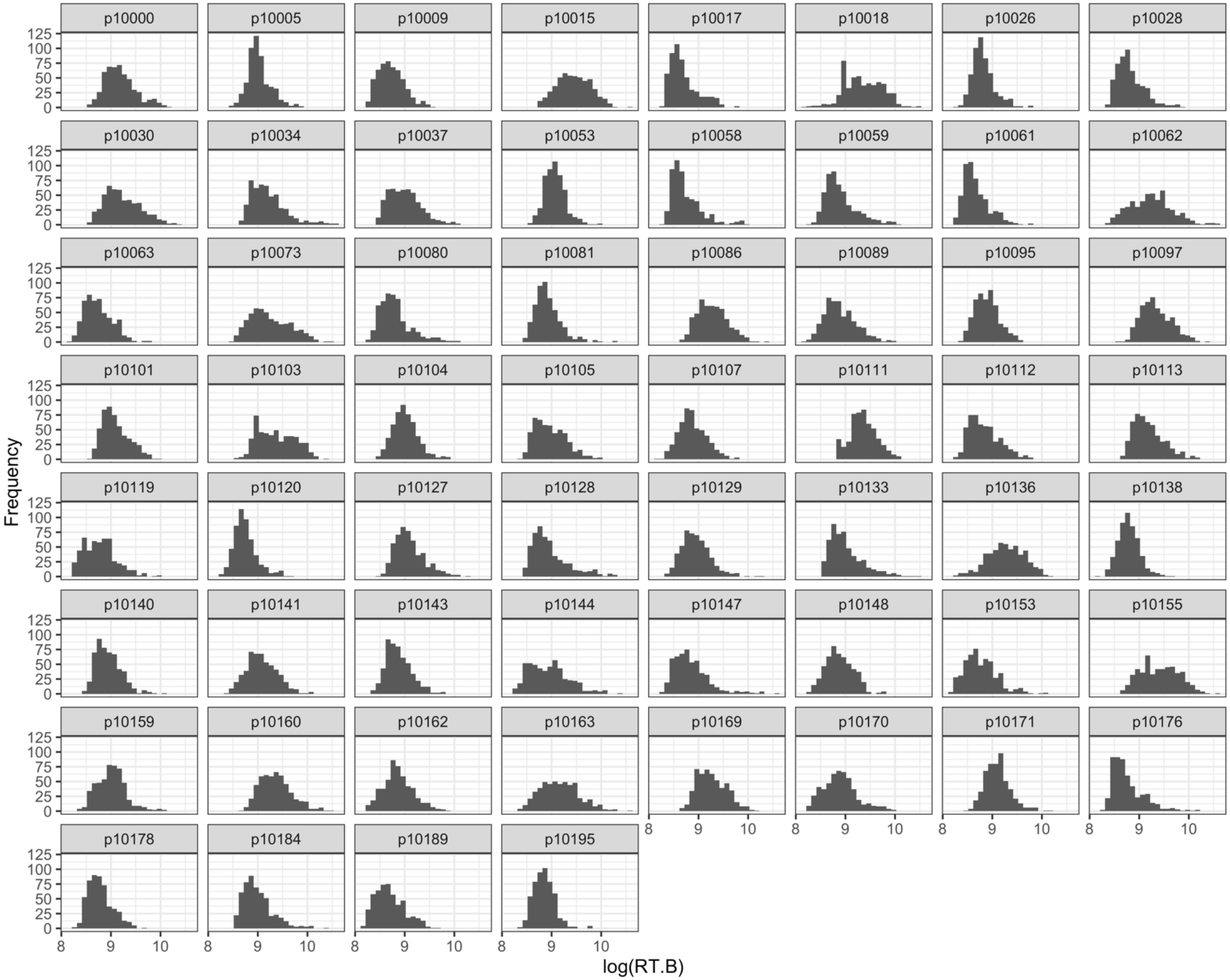


Figure 2.12. Histogram of reaction time data for production practice.

Note. RT was transformed to its logarithm.

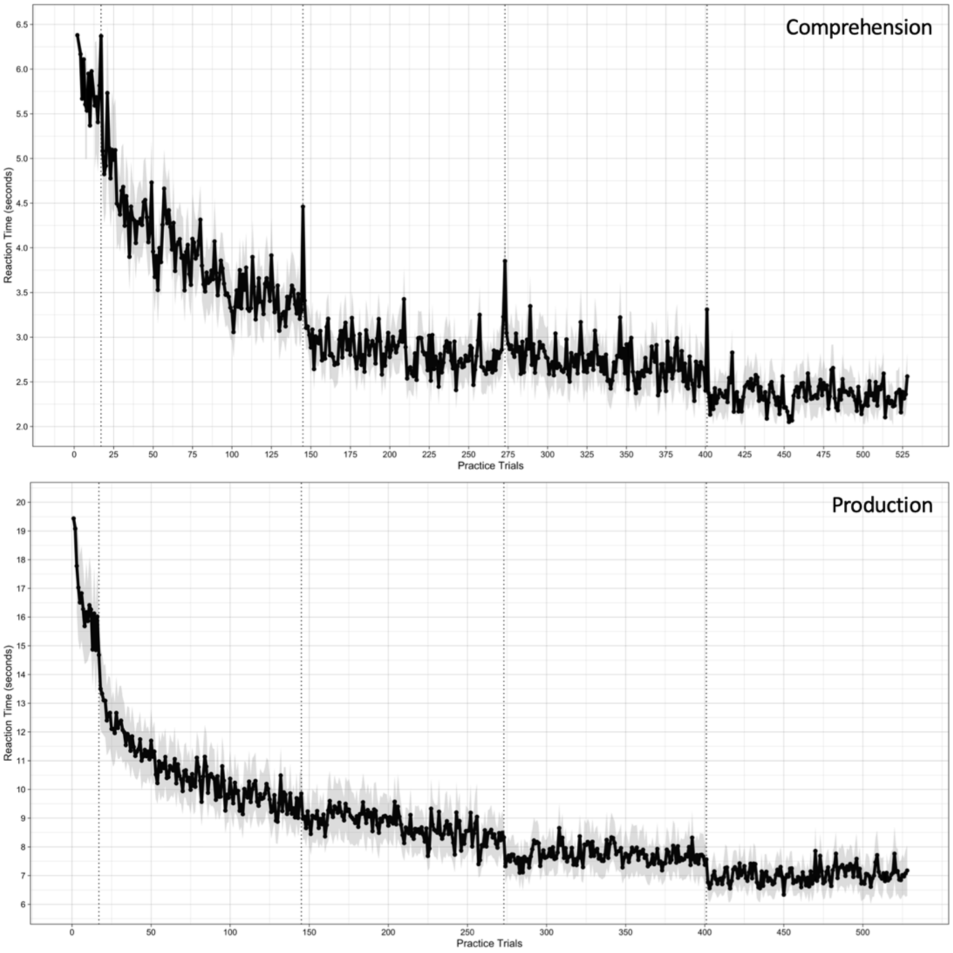


Figure 2.13. RT data before the removal of the first trials.

Note. The dotted lines show the first trial of each practice session.

The first and the second step of data cleaning in total affected 6.54% and 5.60% of the dataset for comprehension and production practice, respectively. Figure 2.14 shows the changes from the original dataset (RT) after the first step (RT.A) and the second step (RT.B). Those data points that did not coincide with their values in the previous step are marked in red. The denser concentrations of data points in the left panels (than in the right panels) show that the first step affected more data (6.15% and 5.01%) than the second step (0.39% and 0.58%). Furthermore, the fact that the discrepancies between RT.A and RT.B (the right panels) were mostly below the straight line indicated that the second step mostly affected data points that were beyond the individual mean +3SD (rather than −3SD).

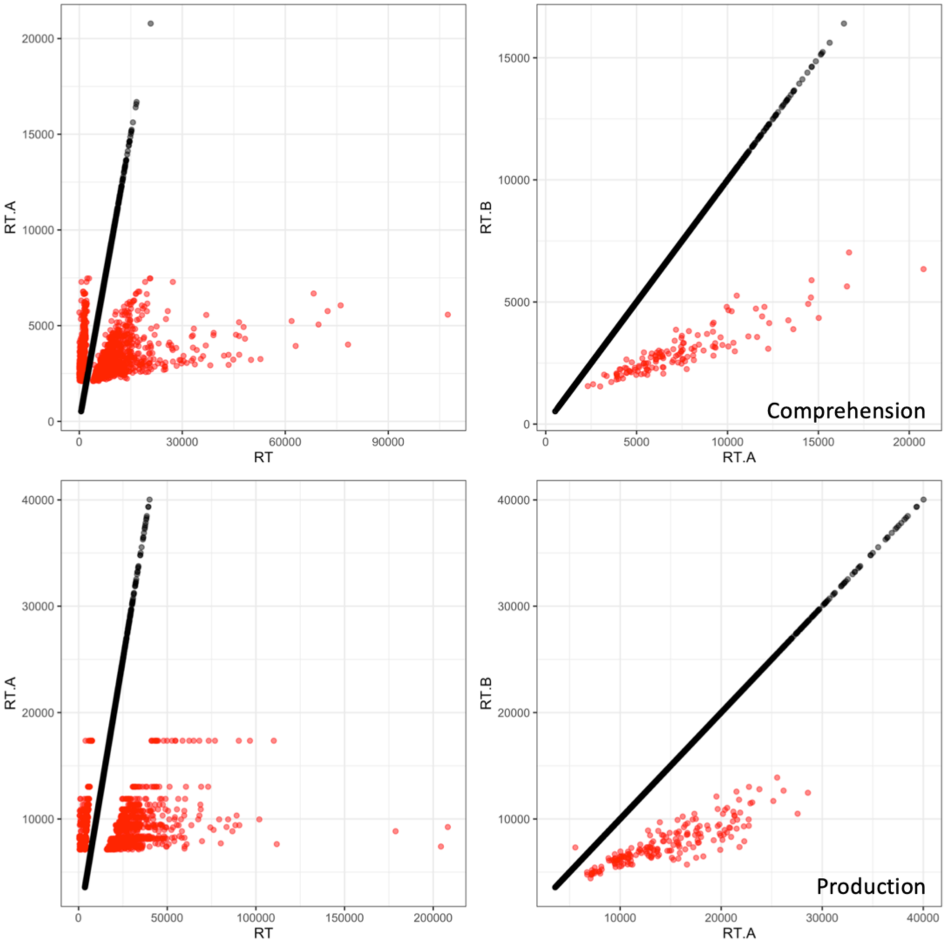


Figure 2.14. The summary of changes in the dataset following data processing. *Note*. Those data points that do not coincide with their previous step are marked in red.

Table 2.4 summarizes variables that were used as independent (or predictor) variables in the study. In the regression analysis (Section 2.3.3), I subtracted 1 from Trial and Block so that the first trial or block corresponded to the intercept of the model. All variables from the cognitive tests were transformed to their *z*-scores to make the intercept and their coefficients (easily) interpretable. The state occupancy (Stage) was dummy-coded with the first stage as the baseline. Lastly, scores from the declarative memory measures (CVMT and LLAMA-B) and the psychomotor ability measures were combined into corresponding factor scores. See 3.1.2 for (a) the convergent validity evidence of CVMT and LLAMA-B to operationalize one's declarative memory capacity and 2CRT and ASRT_1 as a measure to quantify one's psychomotor ability and (b) the procedure of exploratory factor analysis to extract the factor scores.

Table 2.4. The operational definition of independent variables.

Variable	Definition
Trial	The number of practice trials (1–524)
Block	The number of practice blocks (1–33)
Stage	Which learning stage the participants resided in, identified by the HMM analysis (First, Second, and Third)
CVMT	*d*-prime score (-4.65–4.65); see Section 2.2.6.1
LLAMA-B	Percentile of accurate responses across the test (0–100)
ASRT_{15}	The difference between the mean RT on random trials and that on pattern trials in Block 15
SL	Raw accuracy score across the test (1–22)
2CRT	The mean RT across the task
ASRT_1	The mean RT in the first block of the task

2.3.3 Hidden Markov Modeling

The Hidden Markov model (HMM) is an extension of the Markov chain, a stochastic model that represents a sequence of random variables called Markov states (Rabiner, 1989). Markov chain makes an assumption that the future state is only dependent on the current state, and the past state does not influence the future state except via the current state (i.e., the Markov assumption). In the current study, the Markov states represented learning stages defined as distinct cognitive states the participants went through while practicing how to comprehend and produce the target language. In this dissertation, I will use the term *state* to refer to the hypothesized learning stage encoded in the model and *stage* to refer to the actual learning stage whose ontological reality can be assumed. The HMM treats the actual states as hidden, but their probability can be estimated based on observed data by assuming that the hidden states produced the observed data. It is a type of machine learning model often utilized in computational linguistics and natural language processing (see Jurafsky & Martin, 2021, Appendix Chapter A for a review). For instance, the HMM can be used to develop an automated speech recognition system by identifying words (states) based on acoustic information (observed data); it can be used to create a scheme of automatically tagging lexical items by categorizing words into respective parts of speech. In the current study, the HMM estimated the underlying learning stages based on an array of RT obtained through comprehension and production practice trials. In cognitive psychology, Anderson and colleagues have shown that the HMM analysis can be applied and generalized to a wide variety of cognitive tasks, and when tested with datasets where the true number of processing stages (or development stages) is known, the model recovers the stages with a reasonable degree of accuracy (e.g., Anderson, 2012; Anderson et al., 2018; Anderson & Fincham, 2014; Anderson, Zhang, Borst, & Walsh, 2016; Borst, Ghuman, & Anderson, 2016; Tenison, Fincham, & Anderson, 2016).

Following Tenison and Anderson (2016), I fitted a series of HMMs to RT data at the trial level and estimated the probability of each participant residing in a given state on each practice trial (but see below for an issue in parameter estimation). Informed by the models of skill acquisition (Section 1.3), I compared three HMMs, with one, two, and three states, by examining how well each model fit the data. The one-, two-, and three-state models corresponded to the prediction by the Race model, the CMPL theory, and ACT-R, respectively. The HMM adopted in this study was a complex model in which a three-parameter power function was embedded. Note that Tenison and Anderson (2016) also tested a three-parameter exponential function and the APEX function (Heathcote et al., 2000), but I only focused on the power function because (a) the power-law of practice is already well established in skill acquisition research, (b) Tenison and Anderson showed that the APEX function tended to provide worse fits than the power or the exponential function, and (c) as shown in Section 3.1.1, the power function was noticeably more compatible with the current dataset than the exponential function. As introduced in Section 1.4, the three-parameter power function can be denoted as

$$\mathrm{RT}_{i,j} = \mathrm{I} + \beta_i j^{-\alpha},$$

where the reaction time on practice trial j within State i ($\mathrm{RT}_{i,j}$) was modeled as a function of the intercept (I), the slope per state (β_i), and the learning rate parameter (β_i). Following Tenison and Anderson (2016), I assumed that the intercept and the exponent were constant across learning states and estimated different values of the slope for each state. Note that the number of parameters was not exactly three in the model because estimating a slope per state meant that the model required one extra parameter for each additional learning state. Additionally, the HMM estimated transition probability, the probability of individuals eventually moving from one state to another (see Section 1.4). There were $i - 1$ transition parameters for each i-state model; for instance, the three-state model required two transition parameters, one for transitioning from the first to the second state and the other from the second to the third state. In total, the entire HMM estimated $2i + 1$ parameters for an i-state model (except for the one-state model which did not expect any state transition). I estimated the value of the parameters that maximized the probability of obtaining the current dataset. Specifically, the probability of a sequence of RT for each participant (524 trials) was estimated using the following formula:

$$\Pr(j,i) = \sum_{k=1}^{N-j}\left[(1-\pi_{i,i+1})^{k-1}\pi_{i,i+1}\left(\prod_{m=1}^{k} g\left(\mathrm{RT}_{j+m-1}, 3, \frac{\widehat{\mathrm{RT}}_{im}}{3}\right)\right)\Pr(j+k, i+1)\right]$$

$$+ (1-\pi_{i,i+1})^{N-j}\left(\prod_{m=1}^{N-j+1} g\left(\mathrm{RT}_{j+m-1}, 3, \frac{\widehat{\mathrm{RT}}_{im}}{3}\right)\right)$$

$\Pr(j, i)$ denotes the probability of RT from trial j to the last trial given that the participant entered State i on trial j. The equation consists of two parts, one within the summation sign (Σ) and the other outside. The former calculates the probability of trials in which the participant transitions to the next state, and the latter concerns trials in which the participant remains in the same state until the last trial. $\pi_{i,i+1}$ is the transition probability from State i to State $i + 1$, and $(1 - \pi_{i,i+1})^{k-1}\pi_{i,i+1}$ denotes the probability of spending k trials in State i. This part of the equation allowed the model to consider every possible number of trials in each specified state and to choose the best number of trials that maximized the likelihood of obtaining the observed data. In other words, the model considered every possible way of partitioning a sequence of RT data (524 trials) into specified sets dictated by the learning states. Hence, the HMM jointly estimated the power-function parameters, the transition probability (or probabilities), and the number of practice trials within each state.

The probability of spending k trials in State i, $(1 - \pi_{i,i+1})^{k-1}\pi_{i,i+1}$, was then multiplied by $g(\mathrm{RT}_{j+m-1}, 3, \frac{\widehat{\mathrm{RT}}_{im}}{3})$, the probability of the observed RT (i.e., RT_{j+m-1}) on trial $j + m - 1$ given the predicted latency from the power function, $\widehat{\mathrm{RT}}_{im}$, for the mth trial in State i. $g(\,)$ means that the probability was computed based on a gamma distribution, which made it possible to explicitly incorporate the variability among the RT data as part of the model. The Gamma distribution (or the Gamma function) is a class of continuous probability distributions among the exponential family, and it is often used in psychology to model the distribution of RT data (see Palmer et al., 2011 for example). Following Tenison and Anderson (2016), I set the shape parameter of the distribution to 3 and the scale parameter to $\frac{1}{3}\widehat{\mathrm{RT}}_{im}$. This assumed that the variance of RT decreased in proportion to the values of RT; that is, when RT decreased due to practice, so did the variance of RT. The last part of the equation within the summation sign,

$\Pr(j + k, i + 1)$, is the probability of RT from trial $j + k$ to the last trial given that the participant enters the next state (State $i + 1$) on trial $j + k$.

Lastly, the equation outside the summation sign deals with cases in which the participant stays in the same state until the last trial. $(1 - \pi_{i,i+1})^{N-j}$ indicates subtracting the transitional probability ($\pi_{i,i+1}$) from 1, which is the probability of not transitioning between states during trial j to the last trial (N). Note that when the transition probability is 0, the entire equation reduces to what is outside the summation sign, the probability of trials where the participant remains in the same state until the last trials.

In the current analysis, applying the HMMs to trial-level RT data led to an issue of scalability; that is, computational inability to calculate the probability of each participant residing in a given state at each trial. This is due to the fact participants in this study had substantially more practice trials (524 trials) than those in the original study (36 trials). As I inspected the issue, this was not an issue of the models per se, but rather an issue of the estimation method (the expectation maximization algorithm, see below) not being able to converge on a solution. In principle, there were other approaches to estimate the parameters (e.g., naïve Bayes), but those methods have not been used in skill acquisition research and hence were not available at the time of data analysis. To resolve the computational issue, I chose to find the lowest level of data aggregation required to make the computation tractable. Obviously, the first candidate was to analyze the RT data at the block level, which involved aggregating every 16 trials of practice (33 blocks). However, DeKeyser (1997) showed that in his study of L2 skill acquisition (morphosyntax), proceduralization could have been complete as early as by the first 16 trials of practice. This meant that aggregating 16 practice trials into one data point could run the risk of missing the first (very brief) stage of skill acquisition. Instead, I engaged in an exploratory approach, in which I aggregated every 2 to 16 trials of practice and searched for the lowest level of data aggregation that allowed the HMM algorithm to run. This was at the level of four trials (524 / 4 = 131 bins) for comprehension practice and at the level of six trials (514 / 6 $\approx$ 88 bins) for production practice. Hence, I averaged RT data in every 4-trial and 6-trial bins for comprehension and production practice, respectively, and used the resulting dataset to fit the HMMs. I acknowledge that it was most ideal to analyze the raw (trial-level) data, but as discussed in Section 3.2, this data averaging was unlikely to affect the results, especially in regards to the number of HMM states most consistent with the data. Hence, I assumed that at every four or six trials of practice, participants were in the stage of skill acquisition.

Although HMM states represented leaning stages in this study, the mapping between the two sides was not completely one-to-one because the current HMM conceptualized a distinct state for each number of practice trial (or a 4- or 6-trial bin rather); that is, the model went through every trajectory of state transitions to be possible at each trial (or a bin). Figure 2.15 illustrates the difference between a typical HMM (Figure 2.15a) and the model adopted from Tenison and Anderson (2016) (Figure 2.15b). At each trial, participants in the current model had two options: they either proceed to the next trial of the current state or transition to the first trial of the next state. The sheer reason of adopting this complex model was to circumvent the fact that models with within-state speedup violates the Markov assumption when there is only one Markov state for each learning state. Due to this complexity, if there were i states, the HMMs had $131i$ states for comprehension practice and $88i$ states for production practice (if no aggregation had not been applied, there would have been $524i$ states). All parameters associated with the HMMs and the power function were estimated using the expectation maximization algorithm (Rabiner, 1989). The HMM fitting was done on Spyder (Version 5.1.5;

https://www.spyder-ide.org/), an open-source platform to program and execute the Python language. The codes to execute the analysis was provided by Dr. Caitlin Tenison, the corresponding author of Tenison and Anderson (2016).

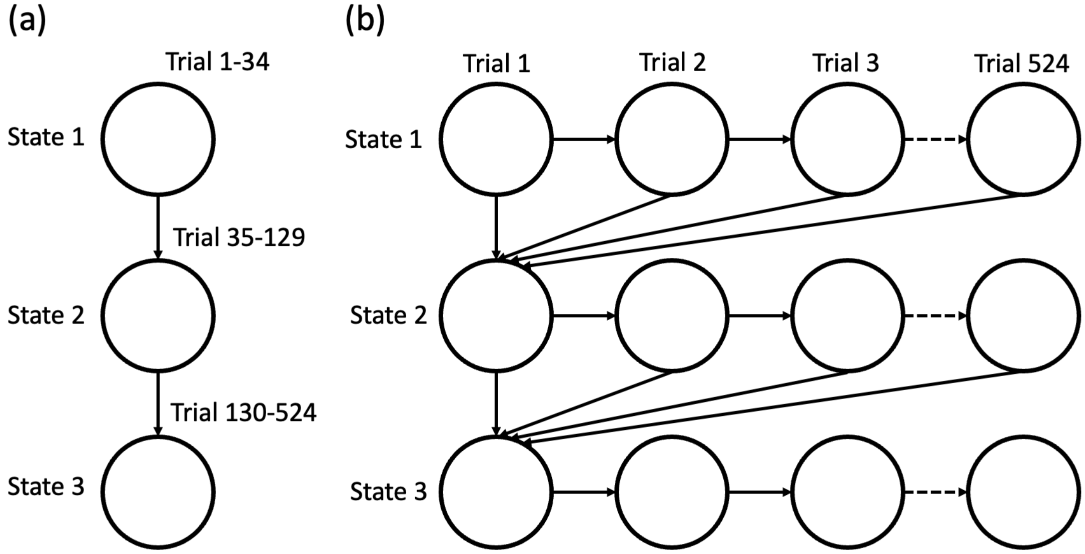


Figure 2.15. The difference between a typical HMM and the model in the current study.

After fitting the HMM with differing number of states (i.e., one to three states), I compared the models based on Akaike Information Criterion corrected for small sample sizes (AICc) and Bayesian Information Criterion (BIC). Although Tenison and Anderson (2016) solely relied on BIC to compare the competing HMMs, the use of BIC assumes that the true model exists in the set of candidate models, which is rarely true in (most) psychological research. In contrast, AIC(c) does not make such an assumption, and it only computes the information loss when a researcher's model is compared against the true model (even though AIC has a drawback of tending to prefer overly complex models, which is not the case for BIC) (see Burnham & Anderson, 2002; Kass & Raftey, 1995 for discussions advocating AIC or BIC). I chose AICc over AIC because AIC is only valid for large datasets, with a common threshold being $\frac{n}{k} < 40$, where n is the sample size and k is the number of parameters estimated (Burnham & Anderson, 2002). AICc and BIC were defined as:

$$\mathrm{AICc} = -2\log L + 2k + \frac{2k(k+1)}{n-k-1}$$

$$\mathrm{BIC} = -2\log L + k\log(n)$$

where $\log L$ is the log likelihood of obtaining 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, I calculated the so-called Akaike weight and the BIC model weight, which can be interpreted as a conditional probability of a model when compared to the other candidate models in the set (with the value between 0 and 1). I followed the formulation of the weights summarized in Wagenmakers and Farrell (2004):

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

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 for (or against) competing HMM models.

One caveat of the current HMM is that it did not conceptualize cases in which the participants regressed back to the previous learning states. However, as is clear in both cognitive psychology and SLA, the skill acquisition theory is also a theory of skill *retention* (Kim, Ritter, & Koubek, 2013; Li & DeKeyser, 2017; Y. Suzuki & Sunada, 2019). Fitting a power function to RT data thus assumed that there would be a smooth and continuous decrease in RT, which is not true particularly when a skill must be learned over a long period of time. While I was willing to make such idealization for the purpose of the current study, any studies that span over multiple sessions/days necessarily invite some degree of forgetting on the participants' side.

2.3.3 Regression Modeling

Regression modeling answered the second research question of the study by investigating which cognitive individual difference variables predicted learning at each stage of skill acquisition. I modelled three dependent variables, accuracy, CV, and RT by fitting generalized linear mixed models (GLMM) using Bayesian inference. I used the software *R* (Version 4.1.2; R Core Team, 2022) and the *R* package *brms* (Version 2.16.3; Büerkner, 2017), which was a front-end *R* package of *Stan* (Version 2.21.3; Stan Development Team), a probabilistic programming language for Bayesian inference and optimization. In Bayesian analysis, prior knowledge in the form of probability distributions is combined with observed data to create posterior distributions (see Gelman et al., 2013; Kruschke, 2015; McElreath, 2020 for general reviews). Mathematically, posterior distributions are derived as the precision-weighted average of the prior and the observed data. Maie and Godfroid (2022), Murakami and Ellis (2022), and Saito et al. (2020) provide recent examples of Bayesian data analysis in SLA.

Below, I list mathematical details of the GLMMs that were applied to the current dependent variables. To sum, accuracy from comprehension practice was modeled with a binomial GLMM, but the same variable from production practice was modeled using a zero-one inflated beta GLMM. Regardless of the mode of language practice, RT and CV were modeled with normal GLMMs. Often, these regression models are alternatively called a logistic, zero-one inflated beta, and linear mixed-effects model, respectively.

Binomial model for accuracy (comprehension)

y_i ~ Binomial(n, p_i), where $i = 1, 2, \dots, n$ and indicates participants

$$p_i = \frac{1}{1 + e^{\mathrm{X\beta}}}$$

$$\begin{aligned} X\beta_i = {} & \alpha + \alpha_{\text{subject}[i]} + \alpha_{\text{item}[j]} + \left(\beta_{\text{Trial}} + \beta_{\text{subject}[i]}\right)x_{\text{Trial}} + \beta x_{\text{Stage2}} + \beta x_{\text{Stage3}} + \beta x_{\text{Declarative}} + \beta x_{\text{ASRT15}} + \beta x_{\text{SL}} + \beta x_{\text{Psychomotor}} \\ & + \beta x_{\text{Stage2:Declarative}} + \beta x_{\text{Stage2:ASRT15}} + \beta x_{\text{Stage2:SL}} + \beta x_{\text{Stage2:Psychomotor}} + \beta x_{\text{Stage3:Declarative}} + \beta x_{\text{Stage3:ASRT15}} \\ & + \beta x_{\text{Stage3:SL}} + \beta x_{\text{Stage3:Psychomotor}} \end{aligned}$$

$$\begin{pmatrix} \alpha_{\text{subject}} \\ \beta_{\text{subject,Trial}} \end{pmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \ \mathbf{S_{subject}} \right)$$

$$\mathbf{S_{subject}} = \begin{pmatrix} \sigma^2_{\alpha_{\text{subject}}} & \sigma_{\alpha_{\text{subject}}}\sigma_{\beta_{\text{subject,Trial}}}\rho \\ \sigma_{\beta_{\text{subject,Trial}}}\sigma_{\alpha_{\text{subject}}}\rho & \sigma^2_{\beta_{\text{subject,Trial}}} \end{pmatrix}$$

$\alpha \sim \text{Normal}(1, 5)$

all βs $\sim \text{Normal}(0, 1)$

$\sigma_{\alpha_{\text{subject}}} \sim \text{HalfCauchy}(1)$

$\sigma_{\beta_{\text{subject,Trial}}} \sim \text{HalfCauchy}(1)$

$\rho \sim \text{LKJ}(1)$

Zero-one inflated beta model for accuracy (production)

$y_i \sim \mathrm{Beta}(\upsilon_i, \omega_i)$, where $i = 1, 2, \ldots, n$ and indicates participants

$$\frac{\upsilon_i}{\upsilon_i + \omega_i} = \mu_i$$

$$\mu_i = \frac{1}{1 + e^{\mathrm{X\beta}}}$$

$$X\beta_i = \alpha + \alpha_{\mathrm{subject}[i]} + \alpha_{\mathrm{item}[j]} + \beta x_{\mathrm{Stage2}} + \beta x_{\mathrm{Declarative}} + \beta x_{\mathrm{ASRT15}} + \beta x_{\mathrm{SL}} + \beta x_{\mathrm{Psychomotor}} + \beta x_{\mathrm{Stage2:Declarative}} + \beta x_{\mathrm{Stage2:ASRT15}} + \beta x_{\mathrm{Stage2:SL}} + \beta x_{\mathrm{Stage2:Psychomotor}}$$

$\alpha_{\mathrm{subject}} \sim \mathrm{Normal}(0,\ \sigma^2_{\alpha_{\mathrm{subject}}})$

$\alpha \sim \mathrm{Normal}(1, 5)$

all βs $\sim \mathrm{Normal}(0, 1)$

$\sigma_{\alpha_{\mathrm{subject}}} \sim \mathrm{HalfCauchy}(1)$

Normal model for CV (comprehension, production)

$y_i \sim \mathrm{Normal}(\mu, \sigma^2)$, where $i = 1, 2, \dots, n$ and indicates participants

$$\mu = \alpha + \alpha_{\mathrm{subject}[i]} + \alpha_{\mathrm{item}[j]} + \left(\beta_{\mathrm{Trial}} + \beta_{\mathrm{subject}[i]}\right)x_{\mathrm{Trial}} + \beta x_{\mathrm{Stage2}} + \beta x_{\mathrm{Stage3}} + \beta x_{\mathrm{Declarative}} + \beta x_{\mathrm{ASRT15}} + \beta x_{\mathrm{SL}} + \beta x_{\mathrm{Psychomotor}}$$

$$+ \beta x_{\mathrm{Stage2:Declarative}} + \beta x_{\mathrm{Stage2:ASRT15}} + \beta x_{\mathrm{Stage2:SL}} + \beta x_{\mathrm{Stage2:Psychomotor}} + \beta x_{\mathrm{Stage3:Declarative}} + \beta x_{\mathrm{Stage3:ASRT15}}$$

$$+ \beta x_{\mathrm{Stage3:SL}} + \beta x_{\mathrm{Stage3:Psychomotor}}$$

Note: $\beta x_{\mathrm{Stage3}}$, $x_{\mathrm{Stage3:Declarative}}$, $\beta x_{\mathrm{Stage3:ASRT15}}$, $\beta x_{\mathrm{Stage3:SL}}$, and $\beta x_{\mathrm{Stage3:Psychomotor}}$ were dropped for production practice.

$$\begin{pmatrix} \alpha_{\mathrm{subject}} \\ \beta_{\mathrm{subject,Trial}} \end{pmatrix} \sim \mathrm{MVNormal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \ \mathbf{S}_{\mathbf{subject}} \right)$$

$$\mathbf{S}_{\mathbf{subject}} = \begin{pmatrix} \sigma^2_{\alpha_{\mathrm{subject}}} & \sigma_{\alpha_{\mathrm{subject}}}\sigma_{\beta_{\mathrm{subject,Trial}}}\rho \\ \sigma_{\beta_{\mathrm{subject,Trial}}}\sigma_{\alpha_{\mathrm{subject}}}\rho & \sigma^2_{\beta_{\mathrm{subject,Trial}}} \end{pmatrix}$$

$\alpha \sim \mathrm{Normal}(0, 1)$ for comprehension and production

all βs $\sim \mathrm{Normal}(0, 1)$ for comprehension and production

$\sigma_{\alpha_{\mathrm{subject}}} \sim \mathrm{HalfCauchy}(1)$ for comprehension and production

$\sigma_{\beta_{\mathrm{subject,Trial}}} \sim \mathrm{HalfCauchy}(1)$ for comprehension and production

$\rho \sim \mathrm{LKJ}(1)$ for comprehension and production

Normal model for RT (comprehension, production)

$\log(y_i) \sim \text{Normal}(\mu, \sigma^2)$, where $i = 1, 2, \ldots, n$ and indicates participants

$$\alpha + \alpha_{\text{subject}[i]} + \alpha_{\text{item}[j]} + \left(\beta_{\text{Trial}} + \beta_{\text{subject}[i]}\right)x_{\text{Trial}} + \beta x_{\text{Stage2}} + \beta x_{\text{Stage3}} + \beta x_{\text{Declarative}} + \beta x_{\text{ASRT15}} + \beta x_{\text{SL}} + \beta x_{\text{Psychomotor}}$$

$$+ \beta x_{\text{Stage2:Declarative}} + \beta x_{\text{Stage2:ASRT15}} + \beta x_{\text{Stage2:SL}} + \beta x_{\text{Stage2:Psychomotor}} + \beta x_{\text{Stage3:Declarative}} + \beta x_{\text{Stage3:ASRT15}}$$

$$+ \beta x_{\text{Stage3:SL}} + \beta x_{\text{Stage3:Psychomotor}}$$

Note: βx_{Stage3}, $x_{\text{Stage3:Declarative}}$, $\beta x_{\text{Stage3:ASRT15}}$, $\beta x_{\text{Stage3:SL}}$, and $\beta x_{\text{Stage3:Psychomotor}}$ were dropped for production practice.

$$\begin{pmatrix} \alpha_{\text{subject}} \\ \beta_{\text{subject,Trial}} \end{pmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \ \mathbf{S_{subject}} \right)$$

$$\mathbf{S_{subject}} = \begin{pmatrix} \sigma^2_{\alpha_{\text{subject}}} & \sigma_{\alpha_{\text{subject}}}\sigma_{\beta_{\text{subject,Trial}}}\rho \\ \sigma_{\beta_{\text{subject,Trial}}}\sigma_{\alpha_{\text{subject}}}\rho & \sigma^2_{\beta_{\text{subject,Trial}}} \end{pmatrix}$$

$\alpha \sim \text{Normal}(0, 2)$ for comprehension and all $\alpha \sim \text{Normal}(0, 3)$ for production

all βs $\sim \text{Normal}(0, 1)$ for comprehension and production

$\sigma_{\alpha_{\text{subject}}} \sim \text{HalfCauchy}(1)$ for comprehension and production

$\sigma_{\beta_{\text{subject,Trial}}} \sim \text{HalfCauchy}(1)$ for comprehension and production

$\rho \sim \text{LKJ}(1)$ for comprehension and production

2.3.3.1 Accuracy (comprehension)

The probability of choosing an correct answer, that is, $p_i(y_i = 1)$, was modeled using a binomial distribution with the number of trials, n, set at 524 trials. A probability is by nature bound between 0 and 1 and hence was transformed to its logit (i.e., log odds) (which is bound between $-\infty$ and $+\infty$) by the logit link function so that the independent variables corresponded to the logit of the probability in a linear manner. The predictor variables included a main effect of residing in a given skill acquisition stage, that is, $\beta x_{\mathrm{Stage2}}$ and $\beta x_{\mathrm{Stage3}}$ (Stage 1 corresponded to the intercept) and a main effect of declarative memory capacity ($\beta x_{\mathrm{Declarative}}$), ASRT$_{15}$ ($\beta x_{\mathrm{ASRT15}}$), SL ($\beta x_{\mathrm{SL}}$), and psychomotor ability ($\beta x_{\mathrm{Psychomotor}}$). I let each stage and cognitive variable interact with each other, so there was a two-way interaction of Stage 2 with declarative memory capacity ($\beta x_{\mathrm{Stage2:Declarative}}$), ASRT$_{15}$ ($\beta x_{\mathrm{Stage2:ASRT15}}$), SL ($\beta x_{\mathrm{Stage2:SL}}$), and psychomotor ability ($\beta x_{\mathrm{Stage2:Psychomotor}}$), and a two-way interaction of Stage 3 with declarative memory capacity ($\beta x_{\mathrm{Stage2:Declarative}}$), ASRT$_{15}$ ($\beta x_{\mathrm{Stage2:ASRT15}}$), SL ($\beta x_{\mathrm{Stage2:SL}}$), and psychomotor ability ($\beta x_{\mathrm{Stage2:Psychomotor}}$). Although I was primarily interested in how each cognitive variable predicted the logit-transformed accuracy rate at each learning stage, I added a main effect of Trial to account for the changes in the accuracy rate as a function of practice trials.

Random effects consisted of the maximal structure allowed by the experimental design. Hence, I estimated varying intercepts and slopes of Trial for individual participants. Note that random effects for items could not be incorporated because there were only two repetitions of the same items across the four practice sessions. Random effects were estimated as if they were drawn from a multivariate normal distribution with the mean of 0 and with the standard deviation indicated by the variance-covariance matrix $\mathbf{S_{subject}}$. See above for the content of the variance-covariance matrix. A multivariate normal distribution is a multi-dimensional extension of a normal distribution. This specification is the same as that of GLMMs implemented in *lme4* package (Bates et al., 2015).

Because there was no a priori information regarding how the participants would increase accuracy (in the specific experimental tasks adopted in the current study) as a function of practice and how the cognitive individual difference variables predicted the participant's accuracy rate, I used weakly informative priors so that the data would overwhelm the priors when there was (at least some) meaningful information in the dataset. Weakly informative prior can be defined as a class of prior distributions "that are explicitly designed to encode information that applies to a general class of problems without taking full advantage of problem-specific knowledge" (Gelman, Simpson, & Betancourt, 2017, p. 3). A weakly informative prior is made intentionally weak so that it does not affect the posterior distribution but still allows for the regularization of extreme values in the posterior samples. For the intercept (α), I used a prior distribution in the form of a normal distribution with the mean of 1 and the standard deviation of 5. This prior expected that aggregating over the cognitive variables, the mean accuracy rate of the participants at Stage 1 (and Trial 1) would be $\mathrm{logit}(1) = .731$, but it can be between .017 and .997 (= $\mathrm{logit}[1 - 5]$ and $\mathrm{logit}[1 + 5]$) within ± 1*SD* of uncertainty. For the regression coefficients, I set the prior distribution as a normal distribution with the mean of 0 and the standard deviation of 1. This prior expected that when the mean accuracy rate was .731, one standard deviation increase in CVMT, for instance, was associated with a 15% increase in accuracy ($\mathrm{logit}[1 + 1] - \mathrm{logit}[1] = .149$).

For the random-effects parameters, that is, $\sigma_{\alpha_{\mathrm{subject}}}$ and $\sigma_{\beta_{\mathrm{subject,Trial}}}$, I used a half-Cauchy distribution with its scale parameter set at 1. The half-Cauchy distribution is a positive half of a Cauchy distribution, which can be derived by holding the degree of freedom parameter of a student-*t* distribution to 1 (note, as *df* → ∞, *t* distribution → normal distribution). I chose a Cauchy distribution because Gelman (2006) showed that with complex models such as GLMMs, standard deviation parameters can often be better approximated by a half-Cauchy distribution than by an inverse-gamma or non-informative prior distribution (which are traditionally used to model error parameters). In addition, I used the Lewandowski-Kurowicka-Joe correlation (LKJ) distribution (Lewandowski, Kurowicka, & Joe, 2009) as the prior distribution for the correlation between the random-effects parameters. I set its scale parameter to 1 so that the distribution becomes (almost) uniform to allow for any value of the correlation between the varying intercepts and slopes for individual participants.

I estimated the posterior distribution of the model parameters through a Markov chain Monte Carlo (MCMC) simulation consisting of four chains of 5,000 iterations each (with 1,000 warmup trials). *Stan* implements a No-U-Turn Sampler as a MCMC algorithm, which is an extension of Hamiltonian Monte Carlo (Hoffman & Gelman, 2014). To check whether each MCMC chain converged on model parameters with a stationary distribution, I monitored whether the value of $\hat{R}$ associated with each parameter (as a convergence index) was within the range of $1 \leq \hat{R} \leq 1.1$ (Gelman & Rubin, 1992). I adopted expected a posteriori (i.e., the mean of the posterior distribution) and 95% credible intervals (CrI: i.e., highest posterior desity intervals) as the point and interval estimates of the model coefficients, respectively. The procedure of parameter estimation was the same throughout the regression analysis and hence it will not be described further in this dissertation unless any changes were made.

2.3.3.2 Accuracy (production)

The zero-one inflated beta regression is an extension of the beta regression, which, in addition to handling variables that are bound between 0 and 1 (as the beta regression does), deals with data that have many data points near 0 or 1 (see Ospina & Ferrari, 2012 for a review of a general class of zero-or-one inflated beta models). It models a dependent variable with a beta distribution, a continuous probability distribution defined on the interval between 0 and 1. It takes two shape parameters, υ and ω, whose relative values determine the shape of the distribution. The mean of the distribution, μ, can be found as $\mu = \frac{\upsilon}{\upsilon+\omega}$, and it is this mean that is predicted by independent variables in a beta-regression model. The fixed effects contained the same predictor variables as those in the binomial GLMM, except that (a) parameters that pertained to Stage 3 were dropped because the two-state HMM was the best model for production practice data and that (a) the main effect of Trial was removed because estimating its effect, that is, the trial-by-trial change in accuracy rates (that were already quite high), proved quite computationally challenging such that it was unable to draw samples from the posterior distribution to estimate the parameter. As shown in Figure 3.1 and 3.2 and Table 3.1, this change (of removing Trial) was unlikely to be an issue because the participants did not increase their accuracy over the course of practice trials (or blocks). Note that because Trial was not entered into a model as a covariate, the model also did not contain varying slopes of Trial for individual participants. Lastly, because I expected that the participants would have the same level of accuracy for comprehension and production practice, I assigned the same prior distributions for the intercept and slope parameters (i.e., $\mathtt{Normal}(1, 5)$ for α and $\mathtt{Normal}(0, 1)$ for βs).

2.3.3.3 CV (comprehension, production)

The CV of RT was modeled with a normal GLMM. The fixed and random effects were identical to those in the models for accuracy rates. I used a prior distribution of $\mathtt{Normal}(0, 1)$ for the intercept, which expected that the mean CV at Stage 1 would equal 0, which can be between -1 and 1 within 1*SD* of uncertainty. The same prior distribution was used for the slope parameters as well, which assumed that when the mean CV is 0, one standard deviation increase in CVMT, for instance, would be associated with 1 increase in CV. This prior did not assume any direction of the change in CV and only expected the degree of change so vaguely that it did not affect the resulting posterior estimate.

2.3.3.4 RT (comprehension, production)

As with CV, RT (log-transformed) was modeled with a normal GLMM. The only difference in the fixed effects was that Trial was transformed to its logit to expect that RT decreases as a power function of practice trials. Because CV and RT were on different scales, I assigned different prior distributions. For the intercept, I assigned a normal distribution of $\mathrm{Normal}(0,2)$ for comprehension practice and $\mathrm{Normal}(0,3)$ for production practice. These priors predicted that for comprehension of the target language, the participants' mean RT at the first stage was 1 second (=log$[0]$) with uncertainty over the values of 0.13 to 7.38 seconds within $\pm 1SD$ (log$[-2]$ to log $[2]$ seconds); for production practice, the prior expected the mean of 1 second with uncertainty over values between 0.04 and 20.08 seconds within $\pm 1SD$. Similarly, for the slope parameters, I used $\mathrm{Normal}(0,1)$, which predicted that one standard deviation increase in CVMT, for instance, is associated with an increase of 3.71 seconds in RT.

Taken together, there were six GLMMs used to analyze the current comprehension and production practice datasets: a binomial GLMM for accuracy in comprehension practice, a zero-one inflated beta GLMM for accuracy in production practice, and four normal GLMMs for CV and RT in both comprehension and production practice. Throughout the analysis, I used weakly informative priors so that when there were informative patterns in the data, the priors were overwhelmed by the data and hence bore no (or very little) influence on the statistical inference drawn from the results.

CHAPTER 3: RESULTS

In this chapter, I report the results of the analysis conducted to investigate the number and the nature of skill acquisition stages in L2 learning. I will first present the results from a preliminary analysis of the data, including descriptive statistics of the dependent variables and cognitive individual difference variables. The results of the HMM analysis will then be presented with reference to the estimated parameter values (i.e., the power-function parameters, the transition probability, and the number of trials within each state) and the results of the model comparison based on AICc and BIC. Lastly, I will present the results of regression modeling to describe how the participants developed accuracy and fluency as a function of practice and the learning stages estimated by the HMM analysis. Furthermore, it will also reveal which cognitive variables predict the development at each stage of learning.

3.1 Preliminary Analysis

3.1.1 Language Practice

3.1.1.1 Accuracy

Figure 3.1 and 3.2 show the descriptive summary of the participants' accuracy rate at the trial and block level of analysis for comprehension and production practice, respectively. Descriptive statistics of the accurate rates at the block level are summarized in Appendix A. Overall, the participants maintained a high level of accuracy throughout the study, and this applied to both comprehension and production practice. It seemed that the participants were less accurate and provided more variable responses during the first 25 trials or so, but the overall mean consistently revolved around or above 90% thereafter, reinforcing the earlier finding from the vocabulary and grammar knowledge test that the participants had developed solid declarative knowledge about the language before engaging in comprehension and production practice.

Although the variability in the accuracy rates was noticeably smaller in production practice than in comprehension practice, this was due to the fact that data points in production practice were an amalgamation of eleven responses, and hence their scale was graded (i.e., 0–1) rather than categorical in comprehension practice (i.e., 0 or 1).

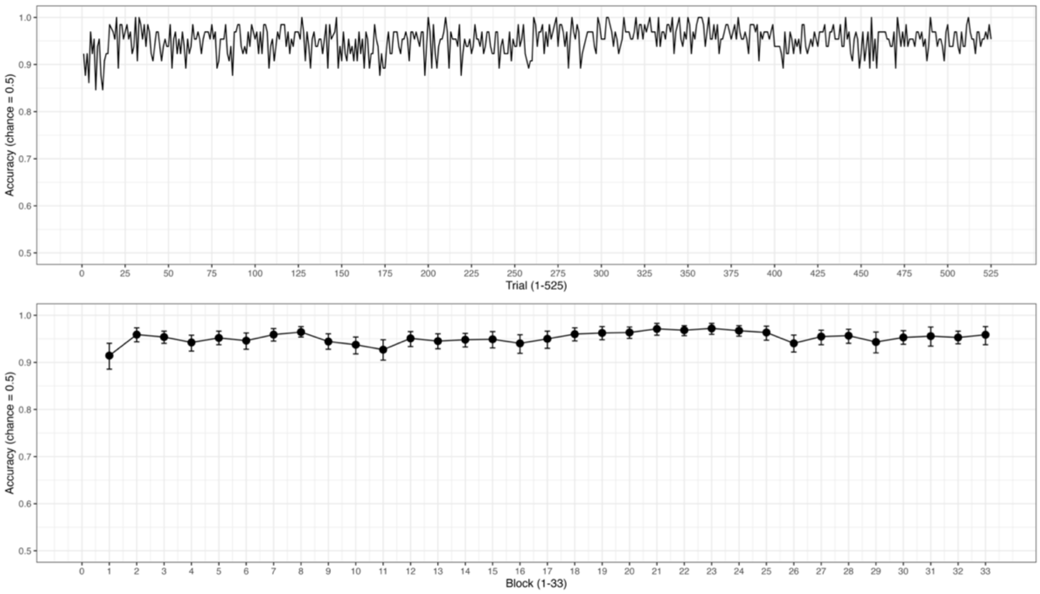


Figure 3.1. The mean accuracy rate over the practice trials (upper) and blocks (bottom) for comprehension practice.
Note. The change performance is 0.5 (or 50%). The error bars in the lower panel show 95% CIs.

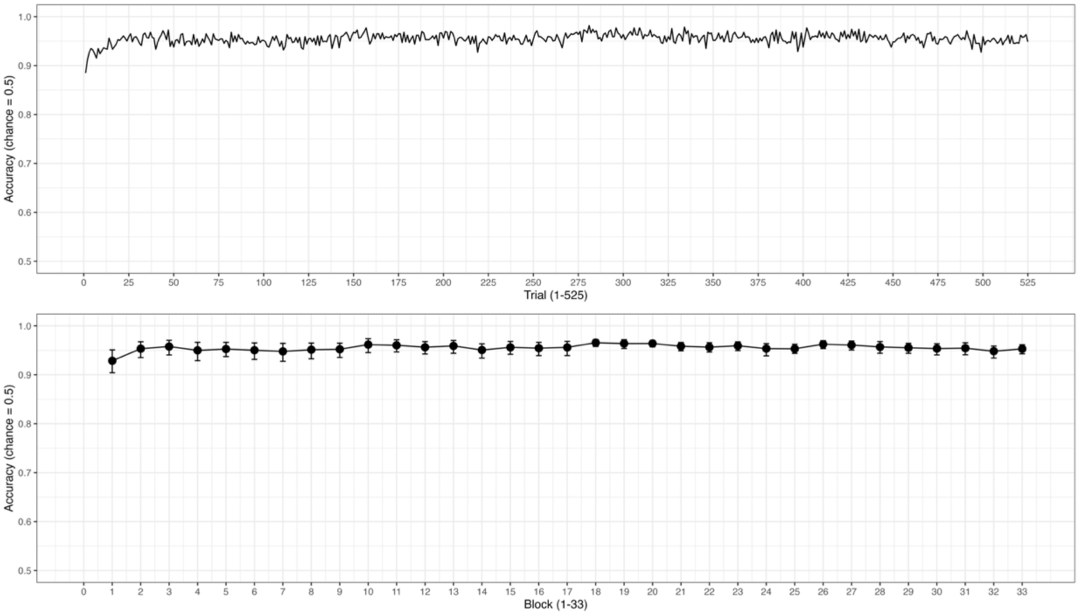


Figure 3.2. The mean accuracy rate over the practice trials (upper) and blocks (bottom) for production practice.

Note. The change performance is 0.5 (or 50%). The error bars in the lower panel show 95% CIs.

In order to validate the observations, I modeled the accuracy data with a generalized linear mixed model (GLMM). Note that because the accuracy rates for comprehension and production practice were on a different scale at the trial level (i.e., 0 or 1 for comprehension practice and the proportion of correct responses out of 11 responses for production practice), the comparison could be made only at the block level. Hence, the dependent variable was the proportion of correct responses per practice block; that is, the individual's mean accuracy rate over 16 trials. Due to the scale of the dependent variable (i.e., a proportion) and the fact that the data were highly concentrated near 1, I modeled the accuracy rates using a zero-one inflated beta GLMM. See Section 2.3.3 for the description of the zero-one inflated beta regression. The fixed effects included the mode of language practice (comprehension = 0 vs. production = 1) and practice blocks (1–33); the random effects only included participant-specific varying (or random) intercepts. I estimated the parameters using Bayesian inference with four MCMC chains of 5,000 iterations each (1,000 warmup iterations).

Table 3.1 shows the results. Although the participants did not seem to improve as they increased the amount of practice over 33 blocks (b = 0.002), they performed more accurately in production practice (b = 2.353+0.666) than in comprehension practice (b = 2.353). However, the difference could be considered minimal in effect size because the participants only scored 4% higher in production practice ($\text{logit}[2.352 + 0.666] = .953$) than in comprehension practice ($\text{logit}[2.353] = .913$). This is understandable as the model was dealing with accuracy data that were consistently high across the practice blocks, even from Block 1. Because the participants were already accurate from the first block of practice, it came as no surprise that the change in accuracy did not follow a power function of practice blocks, $\log(\text{Accuracy}) \sim \log(\text{Block})$: R^2 = .22 and .29 for comprehension and production practice, respectively.

Table 3.1. The posterior estimates from the accuracy model.

	Fixed Effects				
	Estimate	SE_b	95% CrI		$\hat{R}$
Intercept	2.353	0.057	2.242	2.466	1.002
Mode	0.666	0.023	0.621	0.710	1.001
Block	0.002	0.001	-0.000	0.004	1.000
	Random Effects				
	SD	SE_{SD}	95% CrI		$\hat{R}$
Participant	0.405	0.039	0.339	0.491	1.001

$R^2 = .212$

Note. For Mode, comprehension practice was coded as the baseline (and hence corresponds to the intercept).

3.1.1.2 Reaction Time

Figure 3.3 and 3.4 present the mean reaction time of the entire sample and of individual participants as a function of comprehension and production practice trials, respectively. Descriptive statistics are provided in Appendix A. At the sample level, the participants displayed an impressive rate of learning; after 524 trials, the participants' mean RT decreased from 7.15 seconds (*SD* = 3.06, 95% CI [6.40, 7.90]) to 2.56 seconds (*SD* = 1.04, 95% CI [2.30, 2.81]) for comprehension practice and from 19.43 seconds (*SD* = 6.58, 95% CI [17.75, 21.12]) to 7.15 seconds (*SD* = 3.44, 95% CI [6.29, 8.05]) for production practice. In the end, the participants thus required only about one-third of the time they initially needed to perform the task. The decrease in RT also seemed to follow a power function as the curve exhibited a initial sharp drop followed by a gradual speedup. To confirm the observation, I fitted a normal GLMM with an identity link function. The model regressed the logarithm of RT on the logarithm of practice trials (i.e., a log-log model). I tested the power-function model against an exponential-function model in which the logarithm of RT was predicted from practice trials on the original scale. Random effects included varying intercepts and slopes of practice trials for individual participants. Again, the models were Bayesian models estimated based on four MCMC chains of 5,000 iterations each (with 1,000 warmup trials).

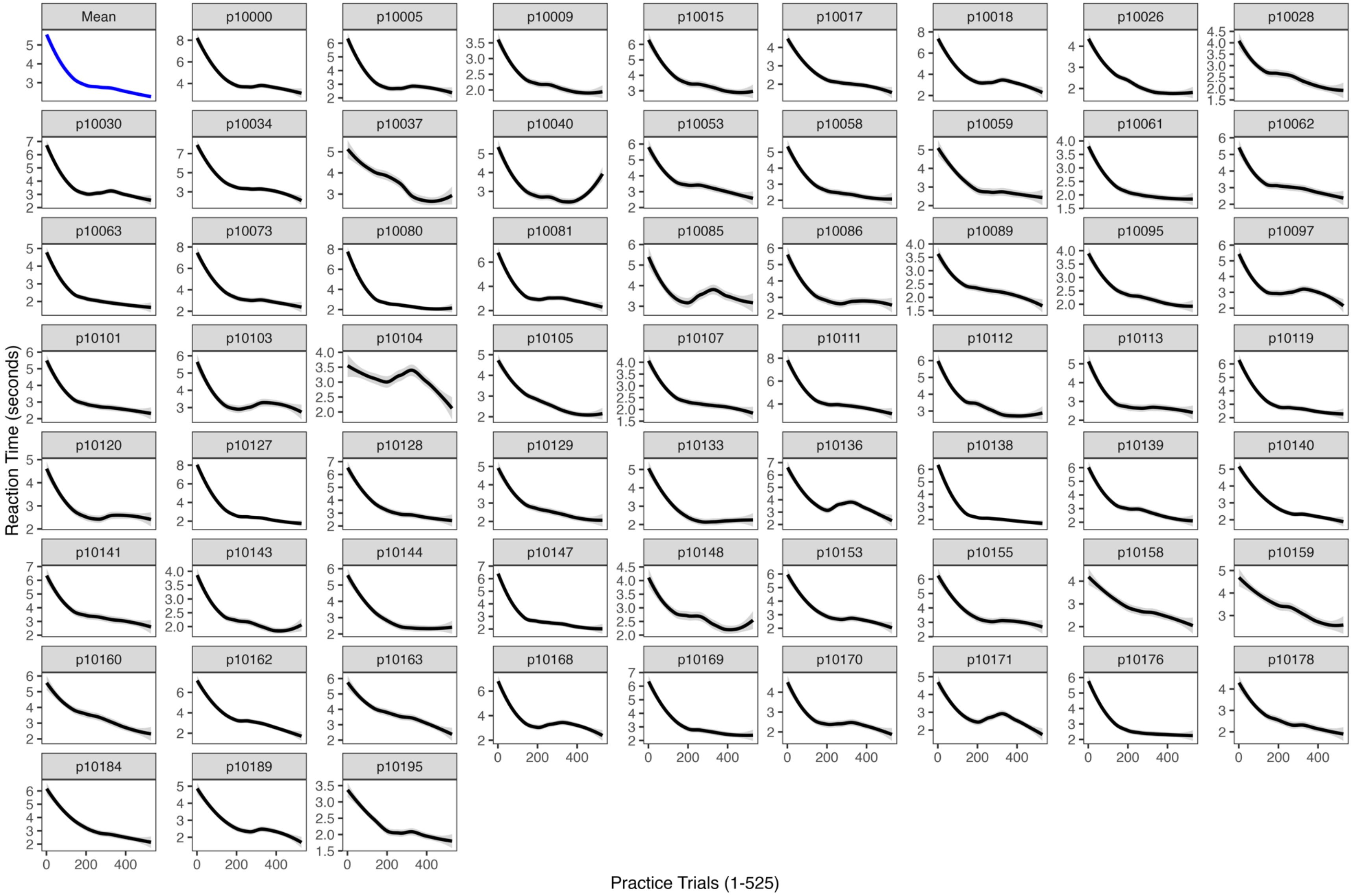


Figure 3.3. Individual reaction time as a function of practice trials (comprehension practice).

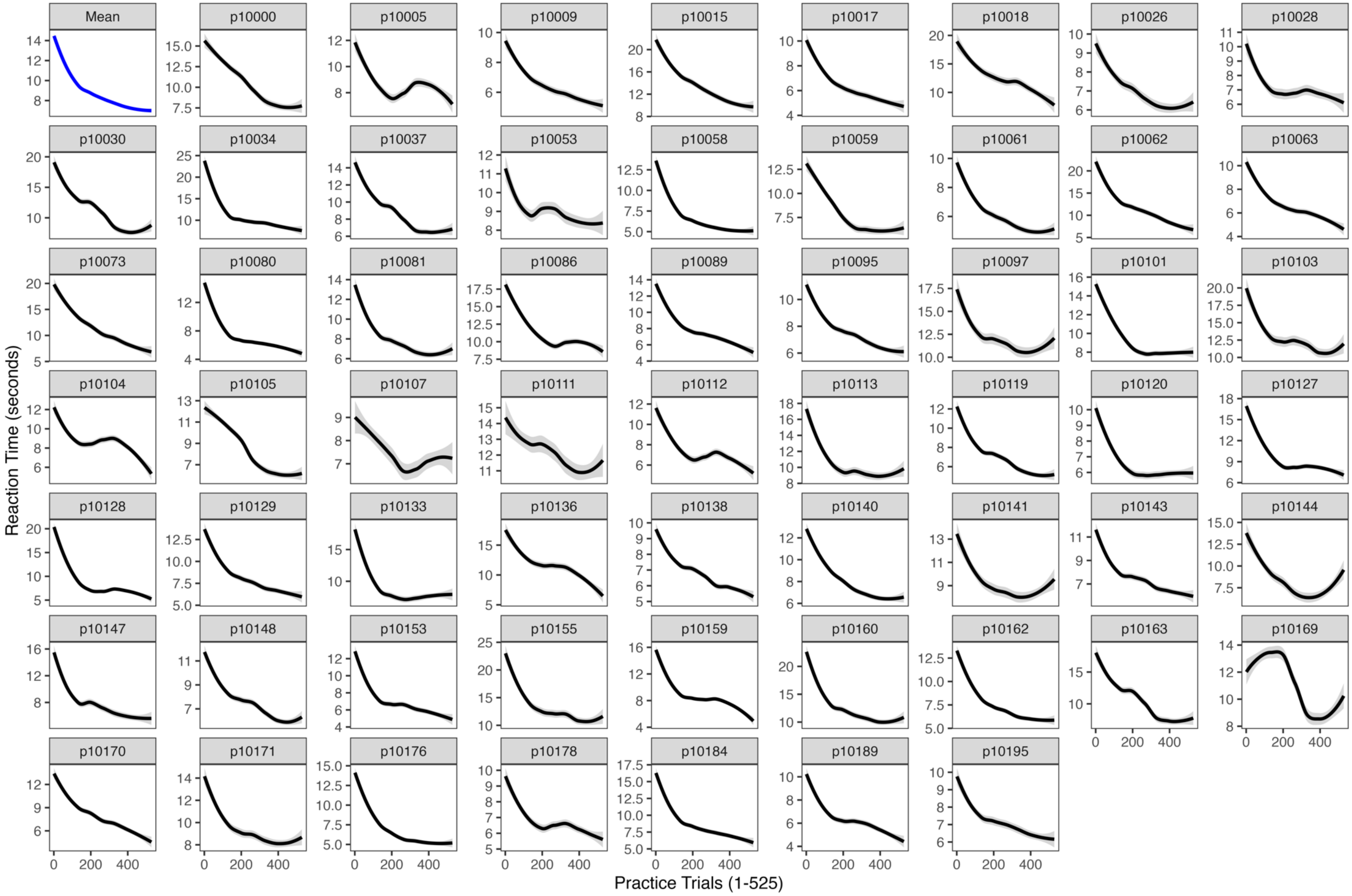


Figure 3.4. Individual reaction time as a function of practice trials (production practice).

Despite the previous finding that an exponential function in general provides a better fit to unaggregated data (Heathcote et al., 2000), the power function was more consistent with the current dataset for both comprehension (R^2 = .36 vs. .32) and production practice (R^2 = .59 vs. .55). This result can also be seen in Figure 3.5, which summarizes the relationship between RT and practice trials in terms of the power function and the exponential function. In the figure, the more linear the relationship between RT and practice trials, the better the fit of a give function becomes. However, examining the value of the model fit (R^2), neither the power function nor the exponential function seemed to account for the RT-practice relationship well. This is likely due to the fact the power-law of practice tends not to apply to individual raw data.

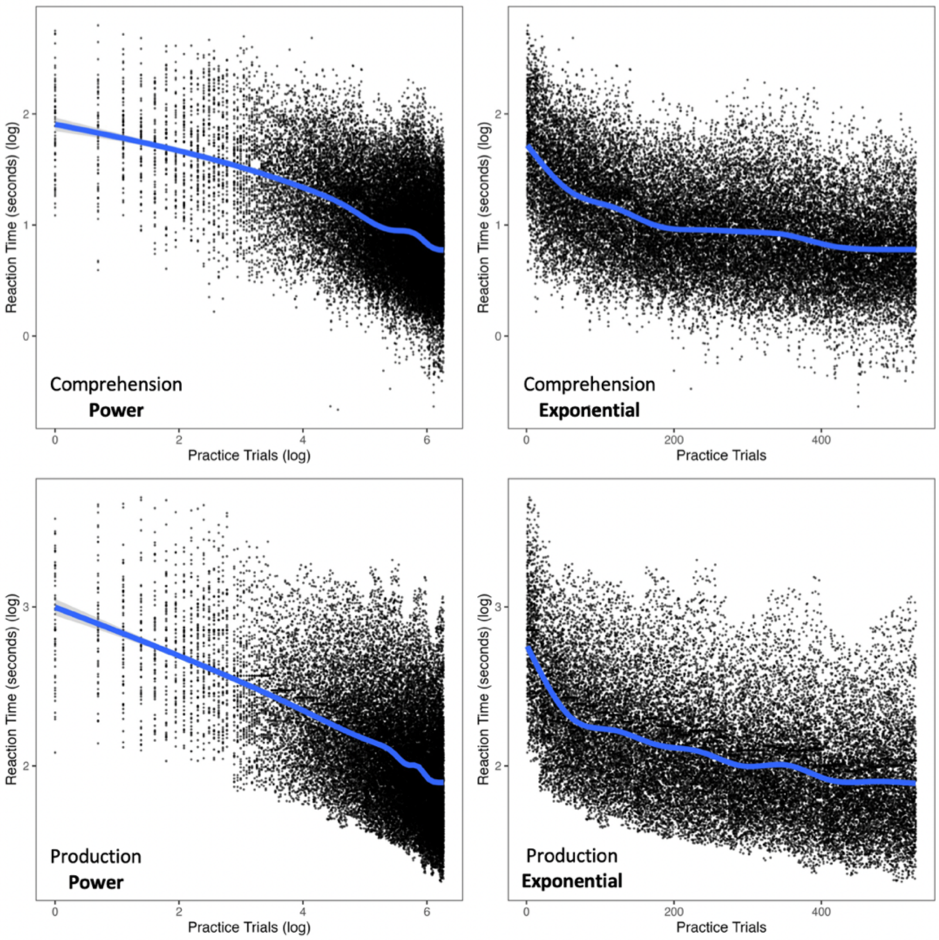


Figure 3.5. The power and exponential function on comprehension and production data.

To compare the results of the current study with those of previous (L2) skill acquisition research, I also tested the power and the exponential functions on RT data that were aggregated over participants and practice blocks. Figure 3.6 and 3.7 summarize the aggregated RT as a function of practice blocks for comprehension and production practice, respectively. The power function fit the RT data very well (R^2 = .97 and .96), but the exponential function (R^2 = .79 and .81) provided a fit that was subpar, compared to the results of previous L2 skill acquisition research (see Table 1.2). Though roughly, not only RT but the standard deviation of RT also seemed to decrease as a power function of blocks: R^2 = .88 and .79 for comprehension and production practice. See Figure S1 and S2 in Appendix A for visual summaries of the standard deviation data. The current dataset is thus well consistent with the power-law of practice and justifies the use of a power function in the HMM analysis (see Section 2.3.3).

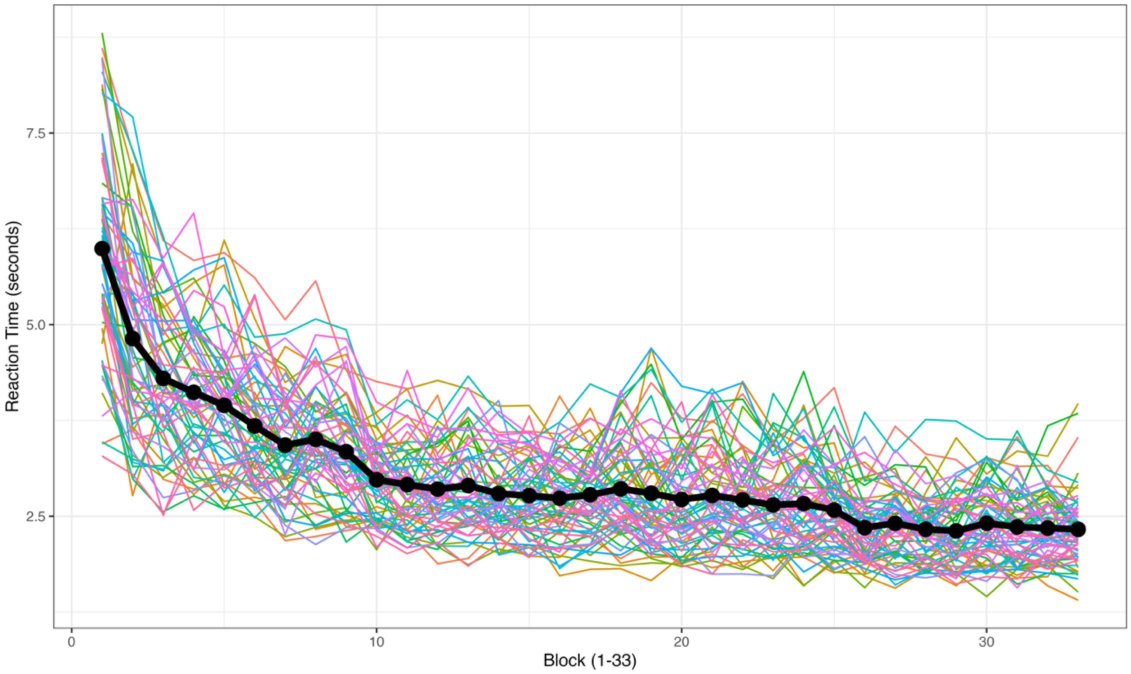


Figure 3.6. Aggregated RT data at the block level for comprehension practice.
Note. The black dots and line show the means over the participants and colored lines display the individual means.

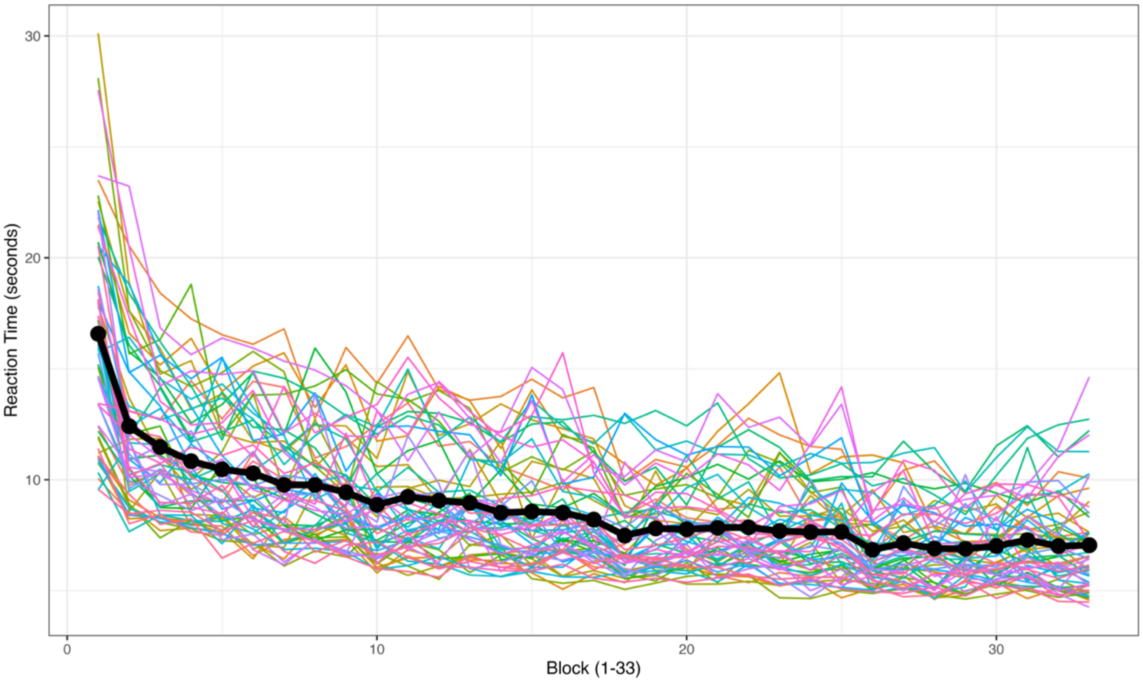


Figure 3.7. Aggregated RT data at the block level for production practice.
Note. The black dots and line show the means over the participants and colored lines display the individual means.

3.1.1.3 Coefficient of Variability

Figure 3.8 and Figure 3.9 summarize the changes in the coefficient of variability (CV) of RT as a function of practice blocks for comprehension and production of Mini-Nihongo, respectively. From Block 1 to Block 33, the CV decreased for both comprehension (from 0.37 [.35, .39] to 0.33 [0.31, 0.35]) and production practice (from 0.22 [0.20, 0.24] to 0.20 [.18, .23]), but there was extreme variability in how the CV decreased for each individual participant. Previously, Hui (2020) demonstrated that at the group level, CV can initially increase before it begins to decrease for proceduralization, following an inverted U-shaped curve. In the current study, this seemed like to be the case for production practice, but in comprehension practice, the CV rather decreased before it started to increase. This observation, however, did not generalize to the participant-level observations as there were high individual differences in how the CV changed for individual participants. For some, the CV decreased rather smoothly, but for others, it had a period of increase before decreasing. In some cases, there was some increase in the CV but it never experienced any decrease. Hence, the group-level observations in Figure 3.8 and 3.9 (i.e., blue lines) are very likely to be an artifact of data averaging. Nonetheless, at the sample level, the participants showed a decrease in the CV as a function of practice even though the decrease was minimal.

The decrease in the CV serves an index of proceduralization (or automatization, as discussed in Segalowitz & Segalowitz, 1993) if and only if the CV positively correlates with the corresponding mean RT. At Block 33, the correlation between the CV and mean RT was *r* = .18 [-.05, .41] for comprehension practice and *r* = .56 [.36, .71] for production practice. As far as the CV is concerned, this meant that proceduralization only took place for production skills. However, this was at odds with the results on RT, in which the participants' speed of performance was found to have reached an asymptote for both comprehension and production skills (see Figure 3.3, 3.4, 3.6, and 3.7). In Chapter 4, I discuss the validity and reliability of using the CV as an index of proceduralization based on these contrasting results.

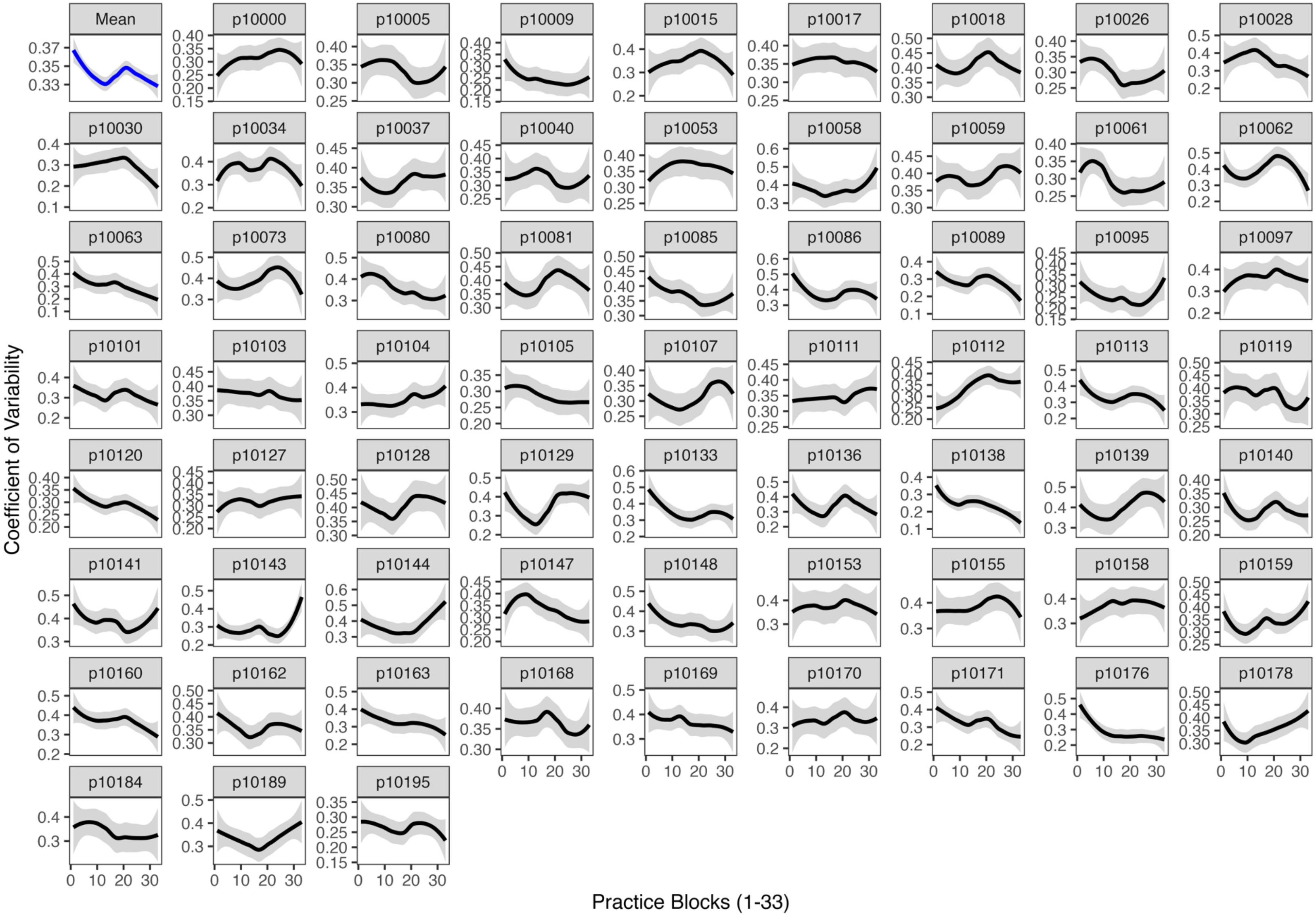


Figure 3.8. The coefficient of variability as a function of comprehension practice blocks.

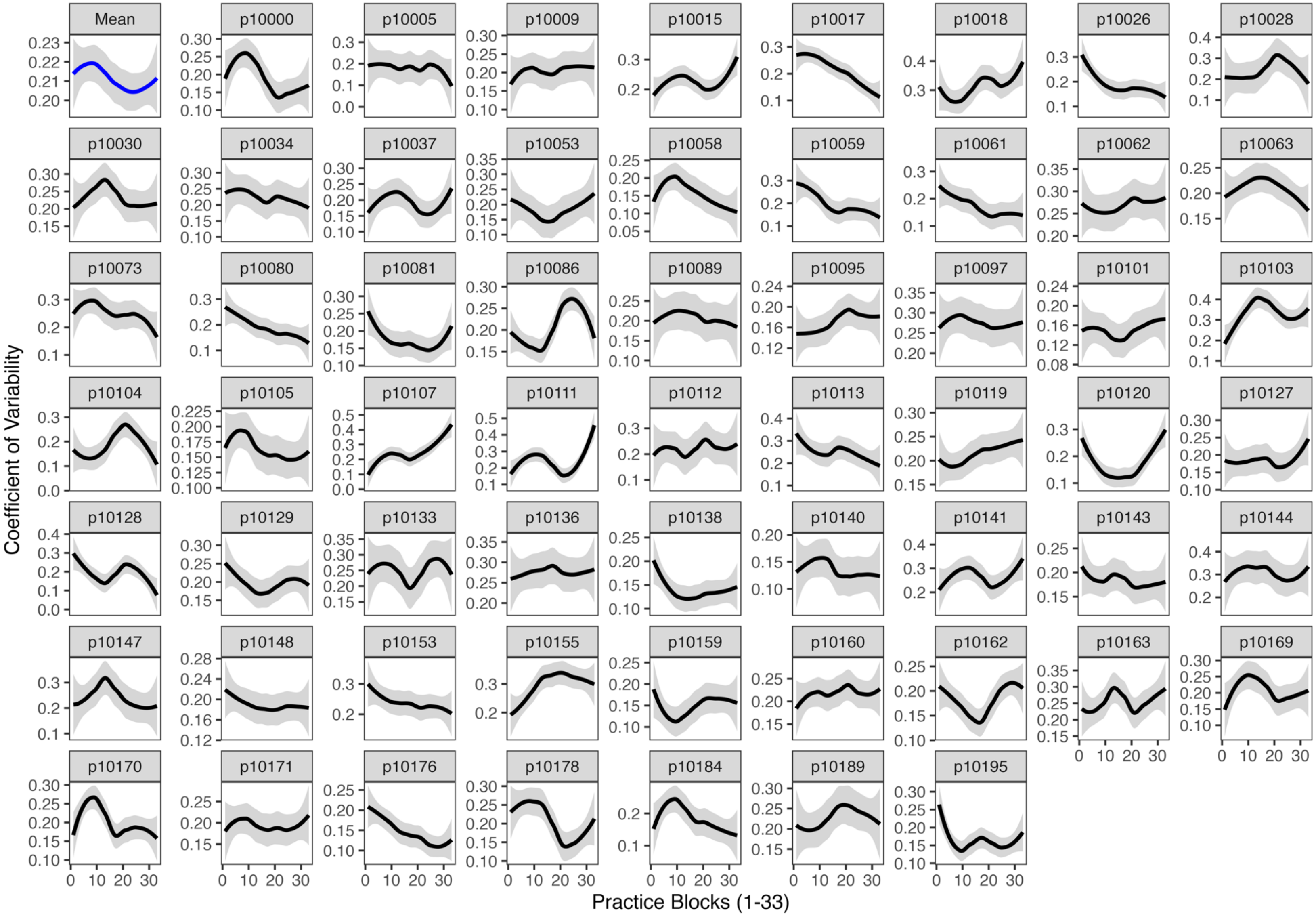


Figure 3.9. The coefficient of variability as a function of production practice blocks.

3.1.2 Cognitive Individual Differences

Figure 3.10 shows the correlation matrix of cognitive individual difference variables. Declarative memory measures (CVMT and LLAMA-B) and psychomotor ability measures (ASRT_1 and CRT) correlated positively within their respective ability dimension (r = .41 [.19, .60] for CVMT and LLAMA-B and r = .61 [.44, .74] for ASRT_1 and CRT), showing the convergent validity of CVMT and LLAMA-B to operationalize the participant's declarative memory capacity on the one hand and of ASRT_1 and CRT to operationalize one's psychomotor ability on the other. However, ASRT_{15} and SL did not seem to relate to each other, which was unexpected from a theoretical standpoint but still consistent with the results of previous research on the convergent validity of procedural memory tasks (e.g., Buffington et al., 2021; Godfroid & Kim, 2021). In particular, SL positively correlated with LLAMA-B (r = .24 [.00, .46]), which may indicate that SL may tap into declarative memory rather than procedural memory capacity.

Because the measures of declarative memory and psychomotor ability showed the evidence of convergent validity, I reduced their dimensions (i.e., four measures into two ability dimensions) through exploratory factor analysis. I used *fa*( ) function in the R-package *psych* (Version 2.1.9: Revelle, 2022) to model two latent factors based on CVMT, LLAMA-B, ASRT_1, and 2CRT. I estimated a set of factor loadings that provided minimum residuals given the four indicator variables and two latent factors. I rotated the factor matrix using the oblimin method to allow for the correlation between the factors. Table 3.2 summarizes the results.

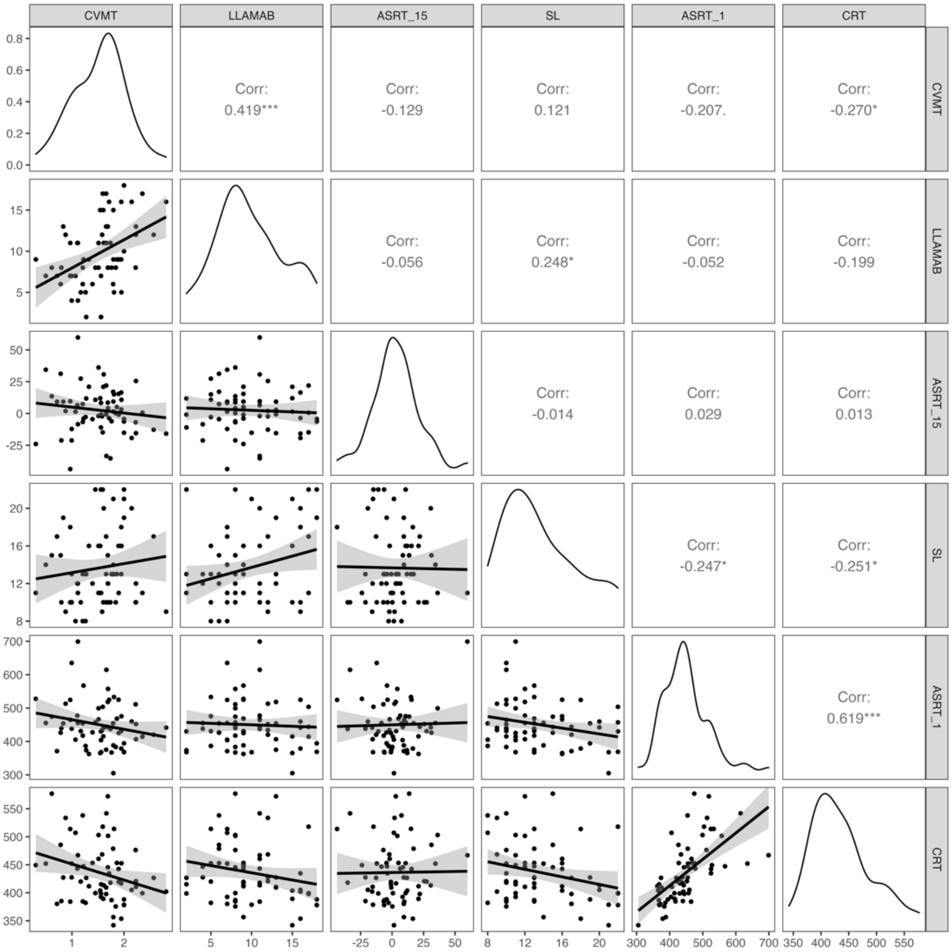


Figure 3.10. The correlation matrix of cognitive individual difference variables.

Table 3.2. Results of the exploratory factor analysis.

	Factor loading			
	1	2	h^2	KMO
CVMT	-0.17	0.51	.32	.62
LLAMA-B	0.04	0.79	.62	.53
ASRT$_1$	0.89	0.07	.77	.52
2CRT	0.69	-0.15	.54	.55
Variance explained	.33	.23		

Note. Factor correlation: *r* = - .199.

As predicted, the four measures clustered into expected groups. Factor 1 was highly loaded by ASRT$_1$ (λ = 0.89) and 2CRT (λ = 0.69) (compared to CVMT: λ = -0.17 and LLAMA-B: λ = 0.04) and hence was labeled as the psychomotor ability, and Factor 2 was consistent with

CVMT ($\lambda = 0.51$) and LLAMA-B ($\lambda = 0.79$) (compared to ASRT$_1$: $\lambda = 0.07$ and SL: $\lambda = -0.15$) and hence interpreted as the declarative memory capacity. I extracted participants' scores on the two latent variables by retrieving the corresponding factor scores. Together, Factor 1 and 2 explained 56% of the total variance among the indicator variables. This meant that there was still 44% of variance that could be explained by other indicators that were not measured and/or sampling error within each indicator variable. This fact is also conveyed in relatively low values of the Kaiser-Meyer-Olkin statistic of sampling adequacy (overall KMO = .55) (which quantifies the amount of common variance among indicator variables). However, the Bartlett's test of sphericity showed that the correlation matrix of the four indicators was significantly different from its identity matrix (i.e., a matrix with no correlations) ($\chi^2(6) = 48.21$, $p < .001$), showing that although the indicators may not share a large amount of variance, there was certainly redundancy that could be combined.

Table 3.3 summarizes the descriptive statistics of the cognitive individual difference variables including the factor scores on the declarative memory capacity and psychomotor ability. There were three observations. First, the participants' scores on CVMT were comparatively lower than those of participants in previous L2 research (e.g., Morgan-Short et al., 2014: $M = 2.11$ and Faretta-Stutenberg & Morgan-Short, 2018: $M = 1.75$–2.09; Walker et al., 2020: $M = 1.80$), which suggested that at the sample level, the participants in the current study may have had weaker declarative memory capacity than that of L2 learners often recruited for laboratory research in SLA. The same observation held true for LLAMA-B (e.g., Hamrick, 2015: $M = .52$; Saito et al., 2022: $M = .64$; Suzuki, 2021: $M = .68$). Second, the mean score on ASRT$_{15}$, that is, the mean difference between RT on the pattern and random trials in Block 15, was -2.54 milliseconds, with its 95% CI covering both negative and positive values almost equally. This suggested that at the group-level, the participants may not have learned the second-order conditional rule underlying the training stimuli. This result may question the validity of the current ASRT task as a measure to operationalize the participant's procedural memory capacity. Nonetheless, this number was very similar to what has been reported in previous SLA research (e.g., Buffington et al., 2021: *M* = 2.99; Faretta-Stutenberg & Morgan-Short, 2018: *M* = 1.26–2.53; Godfroid & Kim, 2021; *M* = -1.15 ms), and this line of research still shows the positive relationship between L2 learning and the procedural learning ability that ASRT is assumed to tap into. Lastly, the mean accuracy rate on SL was .62 [.26, .97], showing that although the participants as a group showed a learning effect on the task, there were still individual differences in how well they learned the adjacent and non-adjacent dependencies based on the training exposure to exemplar stimuli. This figure is similar to that of Romberg and Saffran (2013, Experiment 1) (*M* = .64) from which the current SL task was adopted. Note that the mean of declarative memory capacity and psychomotor ability was 0 with the standard deviation of 1 because factor scores are standardized variables often computed in the form of *z*-scores.

Table 3.3. Descriptive statistics of cognitive individual difference variables.

	Mean	*SD*	*Min*	*Max*	95% CI
CVMT	1.52	0.50	0.30	2.80	0.53, 2.52
LLAMA-B	0.49	0.21	0.10	0.90	0.08, 0.89
ASRT$_{15}$	-2.54	17.91	-59.76	43.73	-38.00, 32.92
SL	0.62	0.18	0.36	1.00	0.26, 0.97
ASRT$_{1}$	450.13	70.05	305.08	699.11	311.43, 588.81
CRT	436.29	53.18	342.98	576.98	330.99, 541.57
Declarative	0.00	1.00	-2.15	2.95	-1.98, 1.98
Psychomotor	0.00	1.00	-1.76	2.05	-1.98, 1.98

3.2 The Number of Skill Acquisition Stages

In this section, I present the results of the HMM analysis used to identify the number of learning stages the participants were likely to have gone through while learning how to comprehend and produce the target language. I discuss the results on comprehension practice first and then move onto those on production practice. In the regression analysis that follows, I assumed the number of stages identified by the HMM analysis (see Section 3.3).

3.2.1 Comprehension

Table 3.4 summarizes the results of the model comparison among one-, two-, and three-state HMM models applied to the RT data from comprehension practice. The table presents the log-likelihood of the models given the current dataset (log (L_i)), the value of AICc, the difference between AICc of a model and that of the best-fitting model (Δ_i(AICc)), Akaike weight (w_i(AICc)), as well as the equivalent statistics based on BIC. As far as the log-likelihoods were concerned, the three-state model was most consistent with the data. The three-state model also showed the lowest value of AICc and BIC, and the corresponding model weight indicated that it was far more likely to be consistent with the data than the other models (w_i(AICc) = .999 vs. $\approx$ 0; w_i(BIC) = .998 vs. $\approx$ 0 and .001). At minimum, the three-state model was 998 times more likely to be true than the one- or two-state models.

Table 3.4. Results of the HMM model comparison for comprehension practice.

	Model comparison						
	log (L_i)	AICc$_i$	Δ_i(AICc)	w_i(AICc)	BIC$_i$	Δ_i(BIC)	w_i(BIC)
1-state	-13881.46	27768.92	1699.09	≈ 0	27790.07	1670.90	≈ 0
2-state	-13043.39	26096.79	26.96	≈ 0	26132.03	12.86	.001
3-state	-13027.91	26069.83	0	.999	26119.17	0	.998
	Power-function parameters						
	I	β_{State1}	β_{State2}	β_{State3}	α		
1-state	0.07	7.17	–	–	-0.23		
2-state	0.67	6.19	3.78	–	-0.20		
3-state	0.00	6.48	4.71	3.49	-0.12		

Table 3.4 additionally contains the best-fitting parameters of a power function given the number of learning stages. The power function provides predicted values of RT at each trial of

practice in a learning given stage. For the three-state model, the fact that the intercept was estimated to be zero meant that each slope parameter corresponded to the mean of RT at the first trial of each stage: hence, the model-based average RT of the first trial in Stage 1 was 6.48 seconds, in Stage 2 was 4.71 seconds, and in Stage 3 was 3.49 seconds. Note that in descriptive statistics, the participants' mean RT at the first and the last trial of practuce was 7.15 seconds (*SD* = 3.06, 95% CI [6.40, 7.90]) and was 2.15 seconds (*SD* = 0.78, 95% CI [1.96, 2.35]), respectively. Hence, if there had been only one learning stage, that is, if the speedup had only been due to within-stage speedup due to some quantitative change in skill performance, the RT would have only decreased to 3.98 seconds ($7.15 * 131^{-0.12} = \mathrm{RT}_{\mathrm{Trial1}} * \mathrm{bin}^{\alpha}$). This shows that the remaining 1.83 seconds ($3.98 - 2.15$) had to be due to between-stage speedup, possibly by shifting the underlying mechanism of skill performance to a more efficient process.

Next, Figure 3.11 displays the probability of stage occupancy across the practice trials. The probability is based on the proportion of the number of the participants residing in a given stage. Hence, although I will interpret the probability as the probability of the entire group residing in Stage 1, 2, and 3, it alternatively shows the proportion of the number of the participants within the sample that occupied a given learning stage.

It was clear that all participants started from Stage 1 (which was a constraint of the HMM model), but they eventually moved to Stage 2 and Stage 3 as a function of practice. Stage 2 became the majority state for the participants (i.e., the probability larger than .50; or more than 50% of the participants) at Trial 38, and this was Trial 288 for Stage 3. To investigate the relationship between the participants' cognitive individual differences and how fast they moved through learning stages, I calculated the number of trials the participants required to reach Stage 2 and Stage 3 and regressed the variable on the four indices of cognitive individual differences (declarative memory capacity, ASRT$_{15}$, SL, and psychomotor ability, all standardized). I used a Poisson model (with the logarithm link function) because the number of trials was a count variable that only took positive values. Note that I also included trials to reach Stage 2 as a covariate when modeling trials to reach Stage 3. The models were Bayesian models with parameters estimated through four MCMC chains each with 5,000 iterations (minus 1,000 warmup trials). Table 3.5 summarizes the results.

All independent variables predicted the number of trials toward Stage 2 and Stage 3. First, declarative memory capacity was negatively associated with the number of trials to reach Stage 2 and Stage 3, indicating that those participants who had higher declarative memory capacity required fewer trials to transition to Stage 2 and to Stage 3. This was also true for psychomotor ability, but the relationship was statistically inverse because the scale of psychomotor ability was based on the speed of performance (or RT), meaning that those participants who had higher psychomotor ability required fewer practice trials to move to Stage 2 and to Stage 3. Lastly, while ASRT$_{15}$ and SL were positively associated with the number of trials to reach Stage 2, but there was an inverse relationship when it came to trials to reach Stage 3. When interpreted with the assumption that "a treatment [or learning] variable interacts with an ID [i.e., individual difference] variable because the treatment variable requires a mental process that is facilitated/hampered by the value of the ID variable" (DeKeyser, 2012, p. 190), it was declarative memory (and psychomotor ability) that facilitated the faster transition from Stage 1 to Stage 2, whereas those participants who relied more on procedural memory tended to take longer time to move to Stage 2. However, this relationship was reversed for procedural memory in relation to the transition from Stage 2 to Stage 3 such that those participants with higher procedural memory capacity required fewer trials to transition to Stage 3.

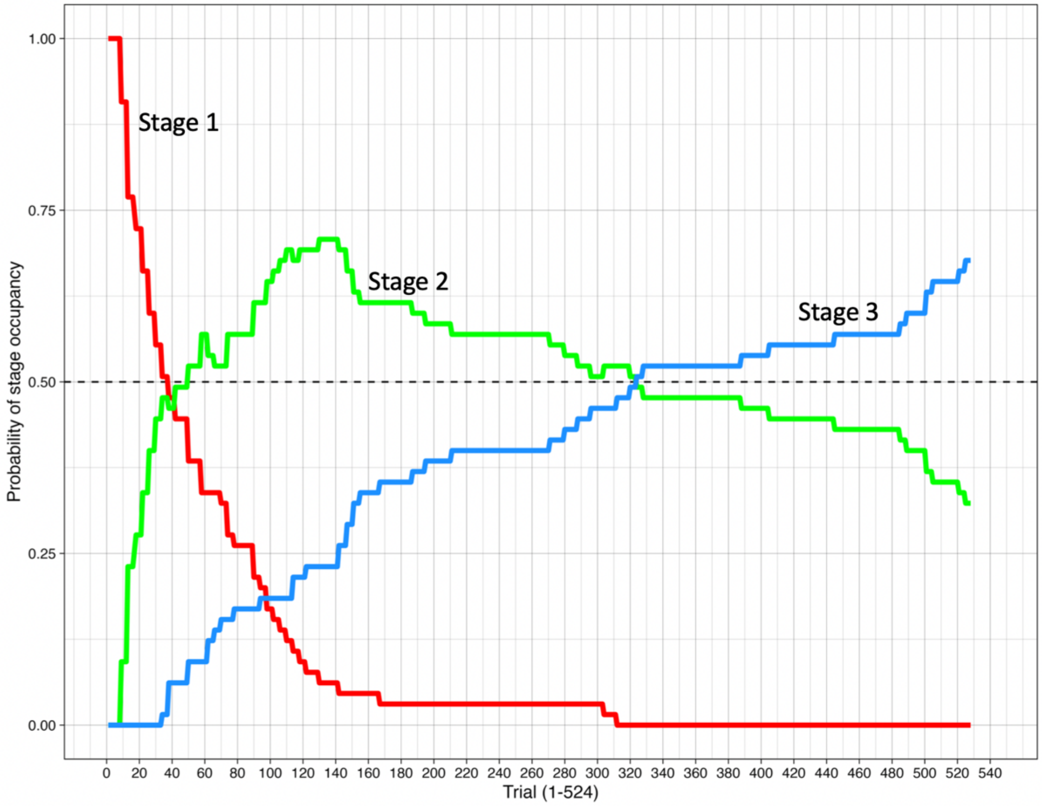


Figure 3.11. Probability of state occupancy in comprehension practice.

Table 3.5. Summary of Poisson regression model for trials to reach Stage 2 and Stage 3.

	Stage 2				
	b	SE_b	95% CrI		$\hat{R}$
Intercept	3.976	0.018	3.942	4.011	1.000
Declarative	-0.272	0.018	-0.308	-0.236	1.000
ASRT_{15}	0.084	0.015	0.055	0.112	1.000
SL	0.136	0.018	0.101	0.170	1.000
Psychomotor	0.255	0.016	0.222	0.286	1.000
	Stage 3				
	b	SE_b	95% CrI		$\hat{R}$
Intercept	5.496	0.010	5.476	5.516	1.000
$\text{Trials}_{\text{Stage2}}$	0.004	0.000	0.003	0.004	1.000
Declarative	-0.158	0.008	-0.173	-0.143	1.000
ASRT_{15}	-0.081	0.006	-0.094	-0.069	1.000
SL	-0.053	0.008	-0.068	-0.038	1.000
Psychomotor	0.151	0.007	0.137	0.164	1.001

3.2.2 Production

Table 3.6 summarizes the results of the model comparison among one-, two-, and three-state HMM models applied to the RT data from production practice. In contrast to comprehension practice, the two-state model was most consistent with the data and at minimum 498.5 times more likely to be true than the other models. The best-fitting parameters for the two-state power function showed that the mean RT of the first trial in Stage 1 was 17.49 seconds and in Stage 2 was 11.46 seconds. In the descriptive statistics, the mean RT at the first and the last trial of practice was 19.43 seconds (*SD* = 6.58, 95% CI [17.75, 21.12]) and 6.84 seconds (*SD* = 2.62, 95% CI [6.17, 7.51]), respectively. If there had only been a single state, the mean RT would have only decreased to 10.85 seconds ($19.43 * 88^{-0.12}$). Hence, the remaining 4.01 seconds must be due to a speedup between stages. Note that the slope for Stage 3 in the three-state model was estimated to be extremely small (β_{State3} = 0.38 seconds), which indicated that imposing three slopes was likely more than what the dataset could have handled and hence unnecessary.

Table 3.6. Results of the HMM model comparison for production practice.

	Model comparison						
	$\log(L_i)$	AICc_i	$\Delta_i(\text{AICc})$	$w_i(\text{AICc})$	BIC_i	$\Delta_i(\text{BIC})$	$w_i(\text{BIC})$
1-state	-15224.14	30454.28	3252.15	≈ 0	30473.92	3238.94	≈ 0
2-state	-13596.06	27202.13	0	.997	27234.98	0	.999
3-state	-13600.20	27214.42	12.29	.002	27260.40	25.42	≈ 0
	Power-function parameters						
	I	β_{State1}	β_{State2}	β_{State3}	α		
1-state	0.51	17.34	–	–	-0.22		
2-state	0.00	17.49	11.46	–	-0.13		
3-state	0.00	17.24	10.92	0.38	-0.13		

Figure 3.12 summarizes the probability of stage occupancy across the production practice trials. As with comprehension practice, all participants in production practice began from Stage

1, but they eventually transitioned to Stage 2, which became the majority state for the participants at Trial 103. Table 3.7 summarizes the results of regressing the number of trials to reach Stage 2 on the four indices of cognitive individual differences. As with comprehension practice, the participants' declarative memory capacity was negatively associated with the number of trials to reach Stage 2 in production practice, which meant that those participants with higher declarative memory capacity required fewer practice trials to move to Stage 2. In contrast, ASRT$_{15}$ and SL were positively associated with the number of trials to reach Stage 2, showing that those participants who had higher procedural memory capacity or those who relied more on procedural memory to perform the task tended to take longer time to reach Stage 2. Psychomotor ability, on the other hand, seemed to play a facilitative role such that those with higher psychomotor ability (with lower values on the variable) required fewer trials to move to Stage 2.

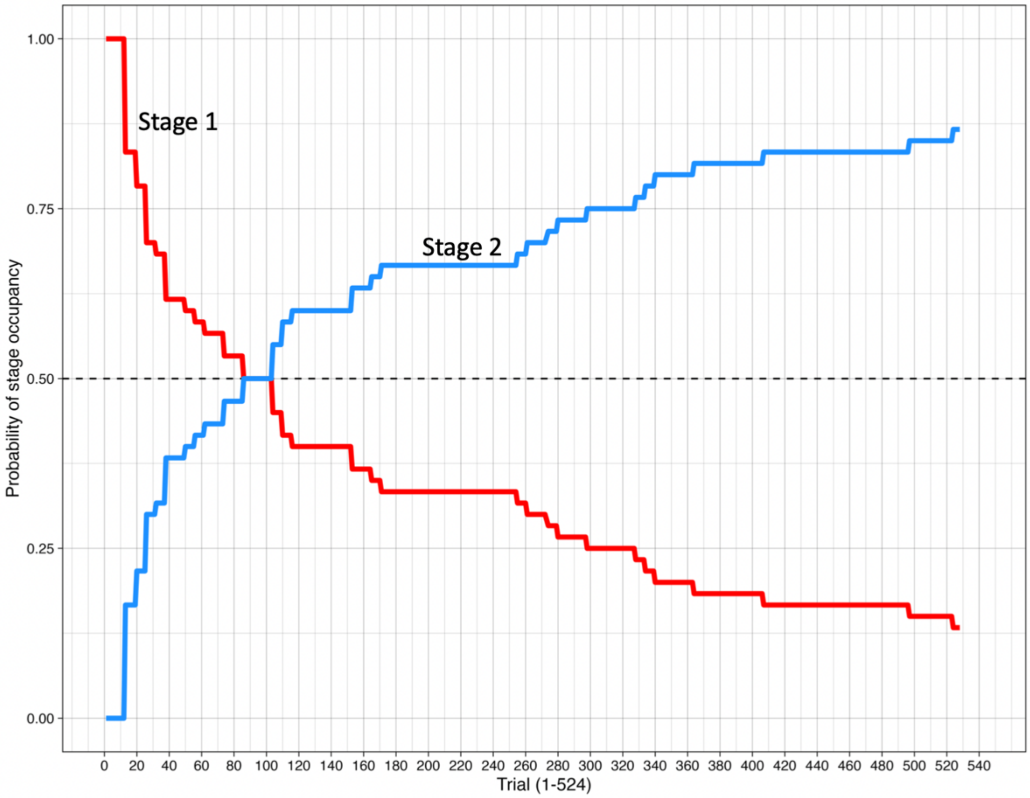


Figure 3.12. Probability of state occupancy in production practice.

Table 3.7. Summary of Poisson regression model for trials to reach Stage 2.

	Stage 2				
	b	*SE*$_b$	95% CrI		$\hat{R}$
Intercept	4.910	0.012	4.886	4.934	1.000
Declarative	-0.429	0.012	-0.452	-0.406	1.000
ASRT$_{15}$	0.170	0.010	0.151	0.190	1.000
SL	0.306	0.010	0.285	0.326	1.000
Psychomotor	0.604	0.010	0.585	0.624	1.000

In summary, the HMM analysis revealed that the acquisition of comprehension skills in Mini-Nihongo can be best encapsulated as a three-stage process, whereas that of production skills took place in two stages. The analysis of the relationship between cognitive individual differences and the number of trials to reach Stage 2 and Stage 3 indicated that for both comprehension and production practice, having higher declarative memory capacity and psychomotor ability was associated with fewer trials to reach Stage 2, and for comprehension practice, those participants with higher procedural memory capacity required fewer trials to reach Stage 3. Assuming that cognitive individual differences imply the learning process facilitated or hampered by the cognitive ability (DeKeyser, 2012), declarative memory facilitated the faster transition to Stage 2 for both comprehension and production practice, and procedural memory facilitated the faster transition to Stage 3 for comprehension practice. An important question is what kinds of cognitive processes are involved in each of these stages of learning?

3.3 The Nature of Skill Acquisition Stages

In this section, I report the results of the regression analysis to address the second research question of the study: the nature of the skill acquisition stages identified by the hidden Markov modeling (HMM) analysis. Based on the findings of the HMM analysis, I assumed that the participants underwent three learning stages to acquire comprehension skills while their learning of production skills was presupposed to have taken place in two stages. In the remainder of this chapter, I thus present the results of the regression analysis based on such conditional assumption (i.e., comprehension skills were acquired in three stages and production skills in two stages). I will first present the results of the analysis on the three dependent variables from comprehension practice, in the order of accuracy, CV, and RT, and then move onto those from production practice.

Because the stage occupancy (i.e., which stage the participants resided in) was dummy-coded in the models, the regression coefficients for two-way interaction terms such as the two-way interaction between Stage 2 (dummy-coded) and declarative memory only shows the average difference between the slope of declarative memory in Stage 1 and that in Stage 2. It does not indicate whether one's declarative memory positively or negatively predicted the dependent variables in Stage 2. This is different from Stage 1, where the main effect directly reveals the role of declarative memory in Stage 1. To interpret the model parameters in the context of the research question in this study, I examined the effects of cognitive individual differences by plotting model-based predictions of the relationship between the dependent variables and the cognitive variables in each stage. I then calculated the posterior probability of whether the relationship is different from zero: $\Pr(b > \text{or} < 0)$. Henceforth, I will abbreviate the posterior probability to Pr in text. This probability encodes how likely the effect of a given variable is present given the current dataset and the model. I used the *conditional_effects* function in the *brms* package to draw model-based expected values of the dependent variables by simulating predicted data points from the corresponding posterior predictive distribution. Note that a *posterior distribution* refers to a distribution of model parameters such as regression coefficients, whereas a *posterior predictive distribution* is a distribution of future (predicted) data based on the model parameters. Due to the numerosity of parameters estimated by the current

GLMMs, I only present estimates of the fixed effects parameters. Appendix B provides full numerical details of the results in a tabular format.

3.3.1 Comprehension

3.3.1.1 Accuracy

Posterior estimates from the binomial model of accuracy in comprehension practice are summarized in Figure 3.13. The dots in the figure display the point estimate of the model parameters, and the error bars represent the 95% credible intervals. Additionally, I have displayed the posterior probability of a given regression coefficient being larger or smaller than 0. The posterior probability corresponds to the color of the dots and error bars; when $\mathrm{Pr}(b > \text{or} < 0) \rightarrow 1$, the color becomes red, and when $\mathrm{Pr}(b > \text{or} < 0) \rightarrow .50$, it approaches to blue. The first observation is that transitioning from Stage 1 to Stage 2 (b = -0.055, 95% CrI [-0.295, 0.173], Pr[b < 0] = .675) and to Stage 3 (b = -0.051, 95% CrI [-0.372, 0.262], Pr[b < 0] = .618) did not relate to any increase in the participant's accuracy of performance, which suggested that it was not accuracy which had to be improved to transition to Stage 2 and Stage 3. However, individual differences in the participants' declarative memory capacity positively predicted accuracy in Stage 1 (Declarative: b = 0.438, 95% CrI [0.168, 0.706], Pr[b > 0] = .999), whereas other cognitive measures such as ASRT_{15}, SL, and psychomotor ability did not seem to affect accuracy in Stage 1 (ASRT_{15}: b = 0.050, 95% CrI [-0.225, 0.326], Pr[b < 0] = .636; SL: b = -0.055, 95% CrI [-0.323, 0.214], Pr[b < 0] = .655; Psychomotor: b = -0.065, 95% CrI [-0.324, 0.197], Pr[b < 0] = .690). One standard deviation increase in declarative memory capacity was associated with 1.4% increase in accuracy ($\mathrm{logit}[3.124 + 0.438] - \mathrm{logit}[3.124]$).

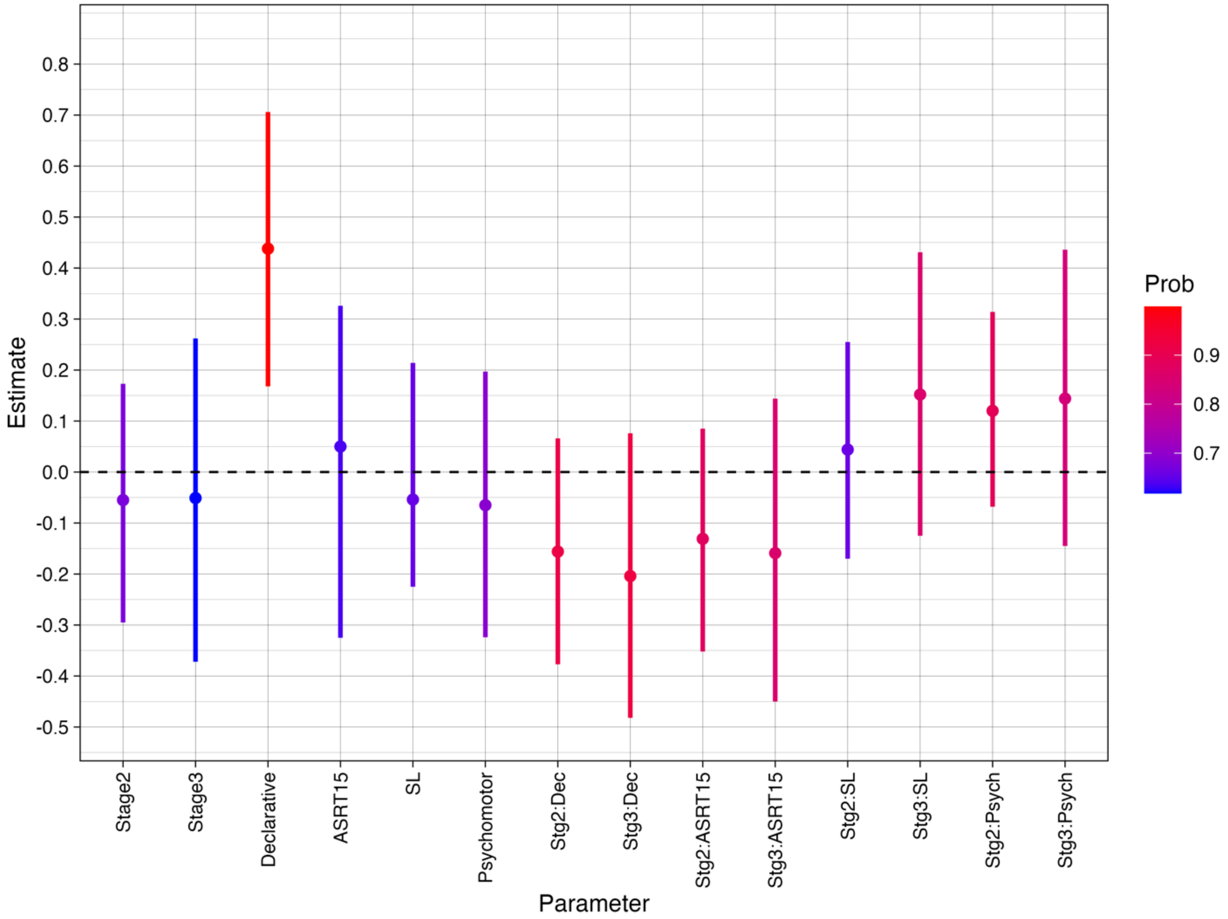


Figure 3.13. Posterior estimates of the fixed effects parameters in binomial GLMM for accuracy in comprehension practice.

In Stage 2 and Stage 3, the effect of declarative memory seemed to weaken in comparison to Stage 1, with the posterior probability larger than .90 (Stage 2:Declarative: *b* = -0.156, 95% CrI [-0.377, 0.066], Pr[*b* < 0] = .920; Stage 3:Declarative: *b* = -0.204, 95% CrI [-0.482, 0.076], Pr[*b* < 0] = .925). However, this result by itself does not show whether one's declarative memory was still facilitative of promoting accuracy in Stage 2 and Stage 3. Figure 3.14 presents the model-based predictions of accuracy in relation to the cognitive individual difference variables. It was clear that the effect of declarative memory was larger in Stage 1, but its influence was still maintained in Stage 2 (Pr[*b* > .0] = .995) and Stage 3 (Pr[*b* > .0] = .974). On the assumption that cognitive individual differences may show the learning process that is enabled by the underlying cognitive ability (DeKeyser, 2012), this shows that declarative memory is facilitative of improving accuracy throughout the process of acquiring comprehension skills, but its effect is more prominent in Stage 1 than in Stage 2 and 3. No other cognitive variables reliably predicted the participants' accuracy in comprehension practice: ASRT_{15} (Stage 1: Pr[$b > 0$] = .636; Stage 2: Pr[$b < 0$] = .765; Stage 3: Pr[$b < 0$] = .803), SL (Stage 1: Pr[$b < 0$] = .655; Stage 2: Pr[$b < 0$] = .540; Stage 3: Pr[$b > 0$] = .785), and psychomotor ability (Stage 1: Pr[$b < 0$] = .690; Stage 2: Pr[$b > 0$] = .687; Stage 3: Pr[$b > 0$] = .715).

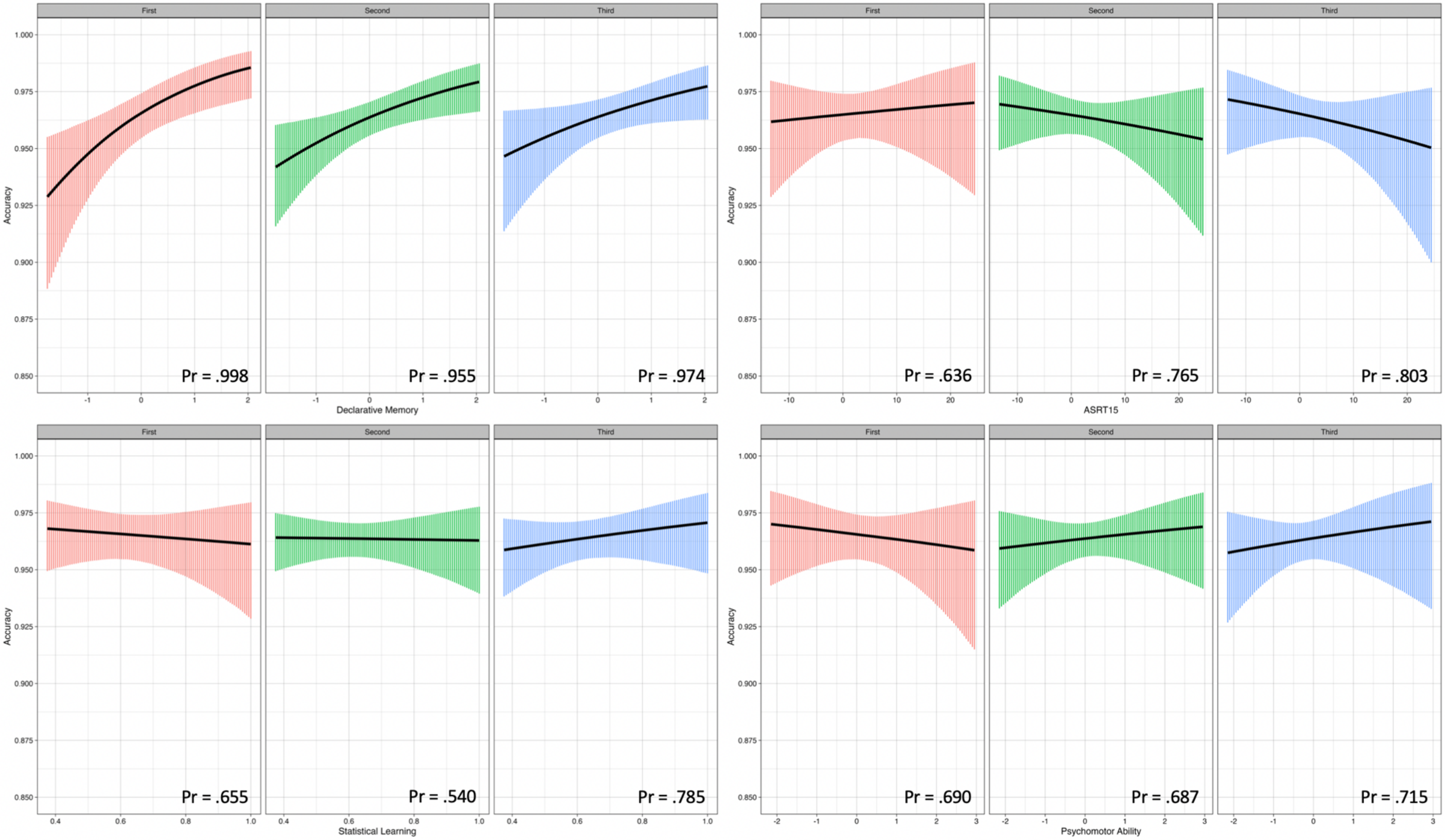


Figure 3.14. Posterior predictions of the relationship between comprehension accuracy and the cognitive variables.

3.3.1.2 CV

Figure 3.15 shows the posterior estimates from the normal GLMM of the CV in comprehension practice. Unlike accuracy, the CV decreased as the participants transitioned from Stage 1 to Stage 2 (b = -0.009, 95% CrI [-0.023, 0.004], Pr[b < 0] = .909) and from Stage 2 to Stage 3 (b = -0.039, 95% CrI [-0.057, -0.020], Pr[b < 0] ≈ 1.000). Hence, the stability of processing (which the CV operationalizes) may be one aspect of skill performance the participants needed to develop to transition to Stage 2 and 3. Interestingly, the decrease in the CV was estimated to be larger in Stage 3 than in Stage 2, which was unexpected because its decrease should be most noticeable in Stage 2, according to the learning mechanism posed by ACT-R. This result may in part be due to the fact that there was extreme variability in how the CV of individual participants changed as a function of practice and that for some participants, there was some initial increase in the CV before it began to decrease (as observed in Hui, 2020). Hence, it is naïve to think that the participants' CV simply decreased (i.e., their processing became more stable) when transitioning to Stage 2 and 3.

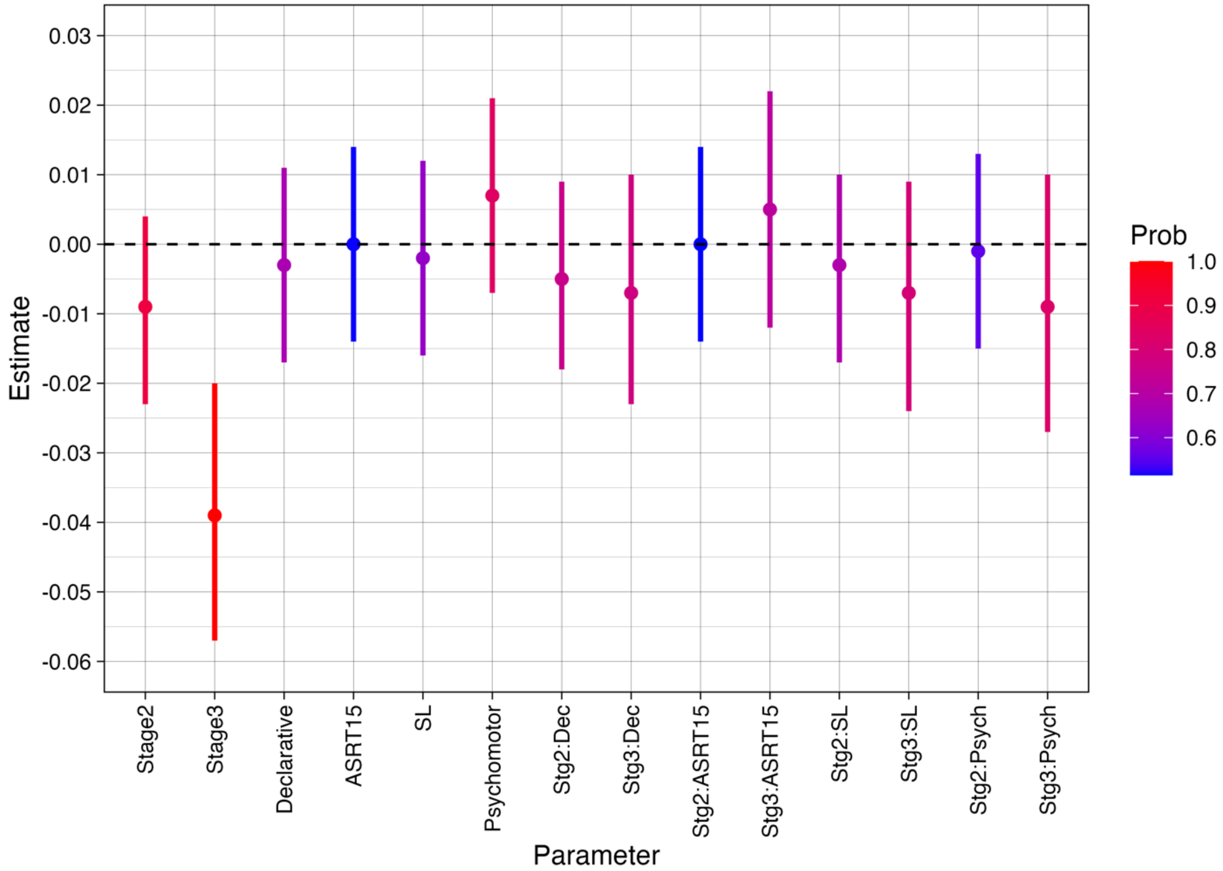


Figure 3.15. Posterior estimates of the fixed effects parameters in normal GLMM for CV in comprehension practice.

Figure 3.16 shows the model-based predictions of the CV in relation to the cognitive individual difference variables. The only reliable predictor was one's declarative memory ability in Stage 2 (Pr[*b* < 0] = .954) and Stage 3 (Pr[*b* < 0] = .940) and SL in Stage 3 (Pr[*b* < 0] = .932), suggesting that those participants with higher declarative memory capacity or statistical learning ability tended to show lower values of the CV. Although no other variables were reliably predictive of the CV, one variable that may predict the change in the CV was psychomotor ability in Stage 1 (Pr[*b* > 0] = .839) and 2 (Pr[*b* < 0] = .882), but again, the standard error around the estimate seemed large, and hence no definitive conclusion can be made.

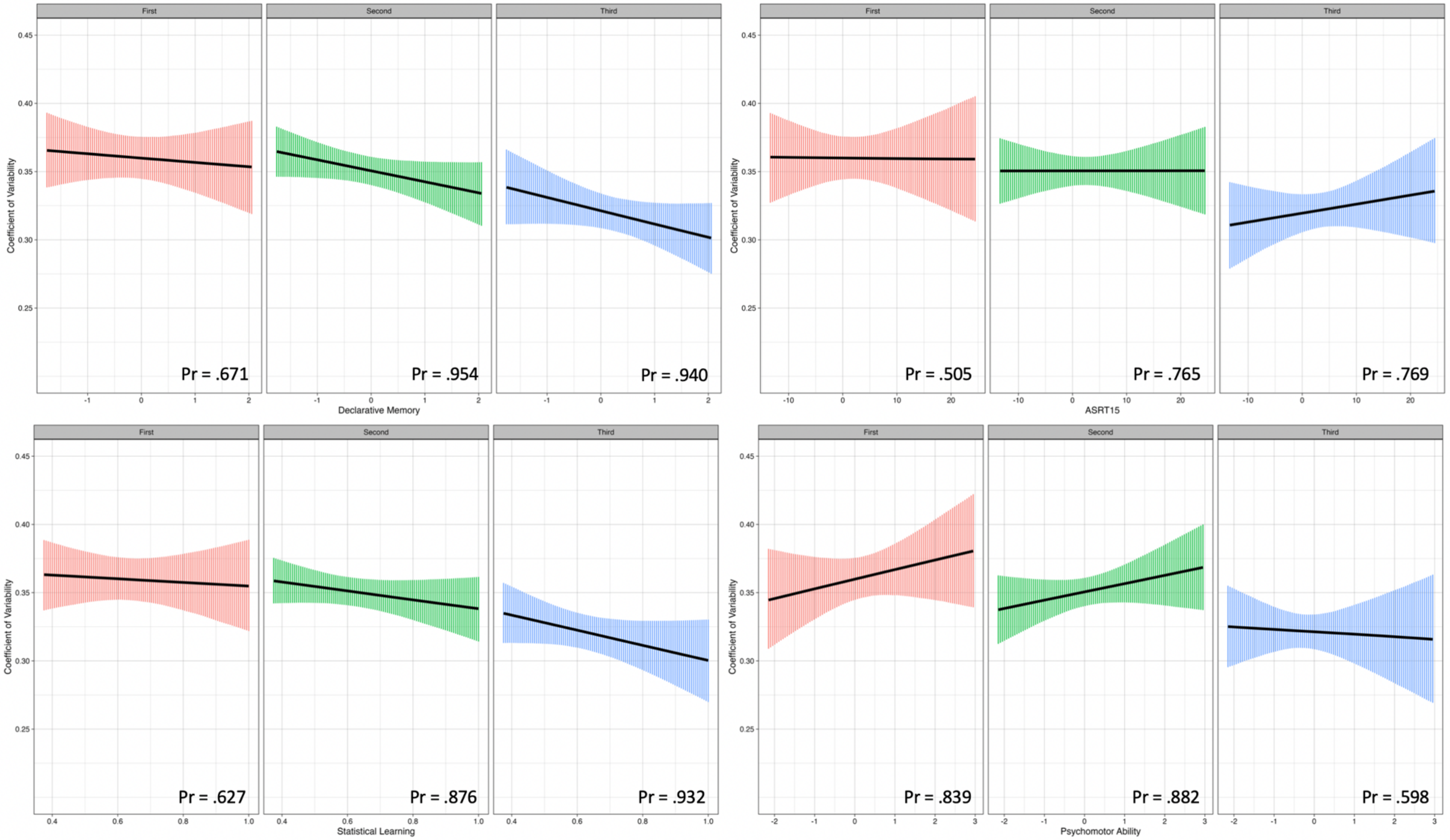


Figure 3.16. Posterior predictions of the relationship between the CV in comprehension and the cognitive variables.

3.3.1.3 RT

Figure 3.17 displays the posterior estimates from the normal GLMM of RT in comprehension practice. Interestingly many variables seemed to play a role in predicting the participant's RT at each stage of skill acquisition. First, transitioning to Stage 2 (b = -0.039, 95% CrI [-0.060, -0.018], Pr[b < 0] = .999) and to Stage 3 (b = -0.114, 95% CrI [-0.143, -0.085], Pr[b < 0] ≈ 1.000) was associated with faster RT. In Stage 1, declarative memory capacity negatively predicted RT (b = -0.072, 95% CrI [-0.109, -0.035], Pr[b < 0] ≈ 1.000), indicating that those participants who had better-functioning declarative memory tended to perform faster than those with lower declarative memory capacity. The same relationship was found for ASRT_{15} (b = -0.048, 95% CrI [-0.084, -0.011], Pr[b < 0] ≈ 1.000), showing that those participants with higher procedural memory capacity tended perform faster in Stage 1. Psychomotor ability also showed a facilitative relationship with RT in Stage 1 (b = 0.121, 95% CrI [0.083, 0.161], Pr[b > 0] ≈ 1.000) as the participants with higher psychomotor ability (i.e., with lower values) were associated with short RT.

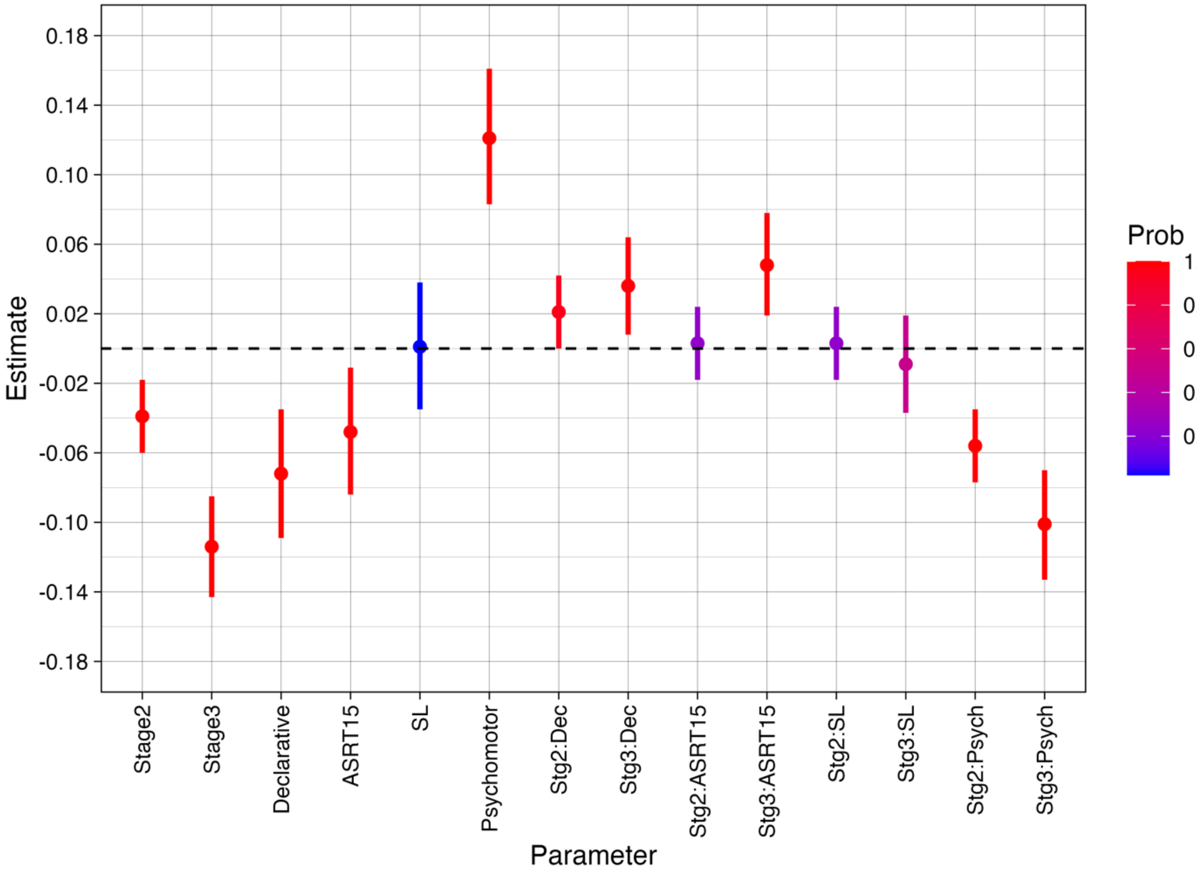


Figure 3.17. Posterior estimates of the fixed effects parameters in normal GLMM for RT in comprehension practice.

Figure 3.18 graphically displays the model-based predictions of RT in relation to the cognitive individual difference variables. Declarative memory remained facilitative of promoting RT in Stage 2 (Pr[*b* < 0] = .999) and Stage 3 (Pr[*b* < 0] = .987), but the two-way interaction of declarative memory with Stage 2 and Stage 3 (see Figure 3.17) showed that its effect was more prominent in Stage 1 than in Stage 2 (*b* = 0.021, 95% CrI [0.000, 0.042], Pr[*b* > 0] = .974) and in Stage 3 (*b* = 0.036, 95% CrI [0.008, 0.064], Pr[*b* > 0] = .995). The pairwise comparison of the regression slopes further showed that the effect was larger in Stage 2 than in Stage 3 (Pr[*b* > 0] = .994). Hence, the participants relied on declarative memory to improve the speed of performance throughout the entire process of skill acquisition in comprehension practice, but its effect gradually weakened from Stage 1 through Stage 2 to Stage 3.

Similarly, ASRT_{15} was facilitative of promoting RT, with its effect only evident in Stage 1 ($\text{Pr}[b < 0] = .994$) and Stage 2 ($\text{Pr}[b < 0] = .997$) but not in Stage 3 ($\text{Pr}[b > 0] = .507$). The interaction between ASRT_{15} and Stage 2 showed that declarative memory was equally influential in Stage 1 and Stage 2 ($b = 0.003$, 95% CrI [-0.018, 0.024], $\text{Pr}[b > 0] \approx 0.617$), showing that the participants relied on procedural memory to improve the speed of performance in comprehension practice, but the effect dissipated in Stage 3. The same relationship was also observed for psychomotor ability (Stage 1: $\text{Pr}[b > 0] \approx 1.000$; Stage 2: $\text{Pr}[b > 0] \approx 1.000$), but the effect was larger in Stage 1 than in Stage 2 ($b = -0.055$, 95% CrI [-0.077, -0.035], $\text{Pr}[b < 0] \approx 1.000$).

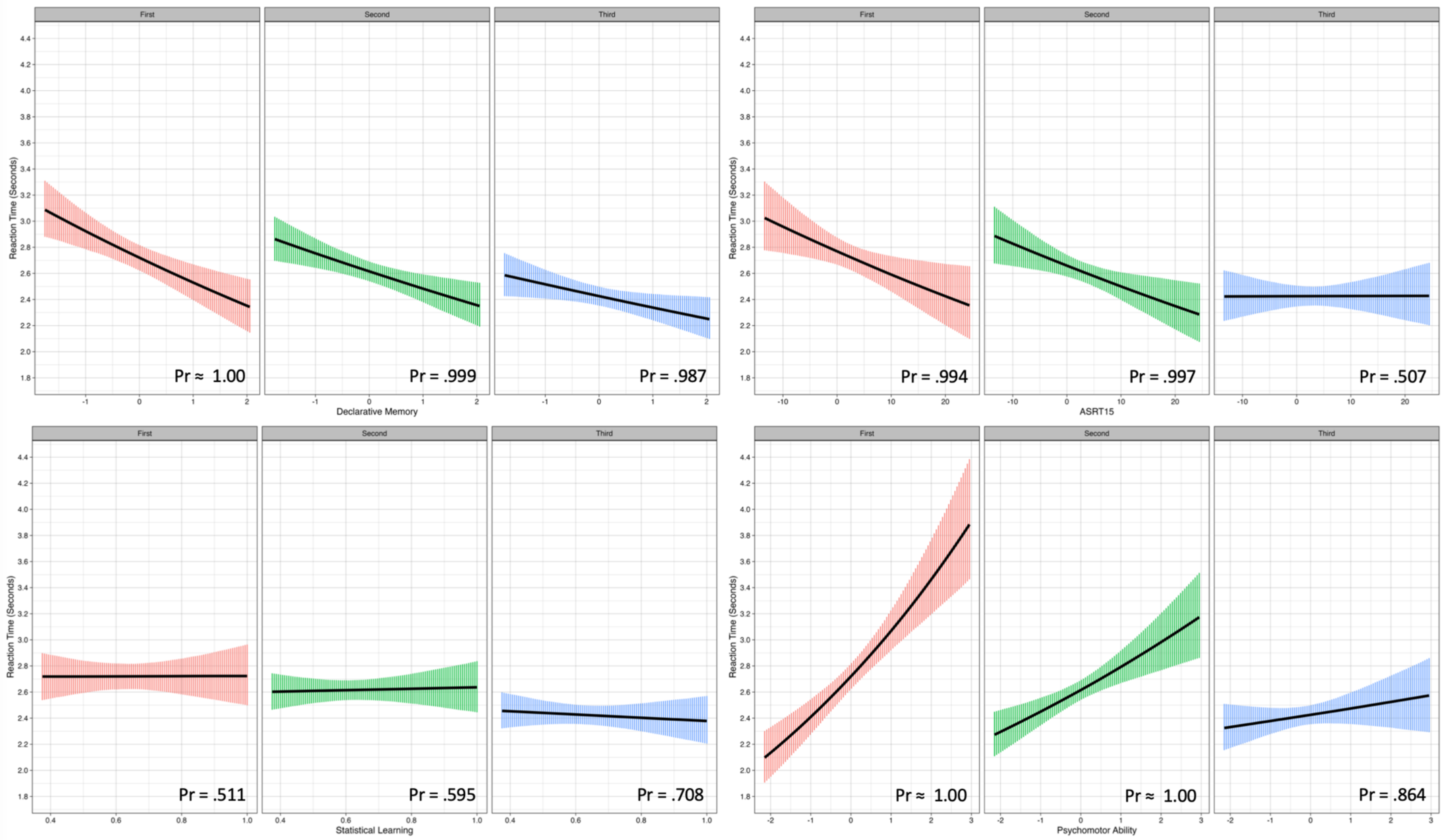


Figure 3.18. Posterior predictions of the relationship between comprehension RT and the cognitive variables.

3.3.2 Production

3.3.2.1 Accuracy

Posterior estimates from the zero-one inflated beta model of accuracy in production practice are summarized in Figure 3.19. Transitioning to Stage 2 (b = 0.034, 95% CrI [-0.007, 0.075], Pr[b > 0] = .946) was associated with higher accuracy rates, showing that for production practice, high accuracy may be characteristic of the transition to Stage 2. In Stage 1, the participant's declarative memory capacity was associated with their accuracy of performance in production practice (b = 0.156, 95% CrI [0.089, 0.219], Pr[b > 0] ≈ 1.000). However, the effect became feeble in Stage 2, as expressed by the two-way interaction between declarative memory and Stage 2 (b = -0.119, 95% CrI [-0.157, -0.081], Pr[b < 0] ≈ 1.000). Figure 3.20 shows the relationship between accuracy in production practice and the cognitive individual difference variables that was predicted by the model. It was clear that declarative memory promoted accuracy in Stage 1 (Pr[b > 0] ≈ 1.000), but its effect became weaker in Stage 2 (Pr[b > 0] = .895) even though it was still in the expected direction.

No other variables reliably predicted the accuracy of performance in production practice: ASRT_{15} (Stage 1: Pr[b > 0] = .518; Stage 2: Pr[b > 0] = .845), SL (Stage 1: Pr[b < 0] = .644; Stage 2: Pr[b < 0] = .635), and psychomotor ability (Stage 1: Pr[b < 0] = .547; Stage 2: Pr[b < 0] = .690). It was noteworthy, however, that procedural memory seemed to predict accuracy in Stage 2, but this result is provisional due the uncertainty over the estimate.

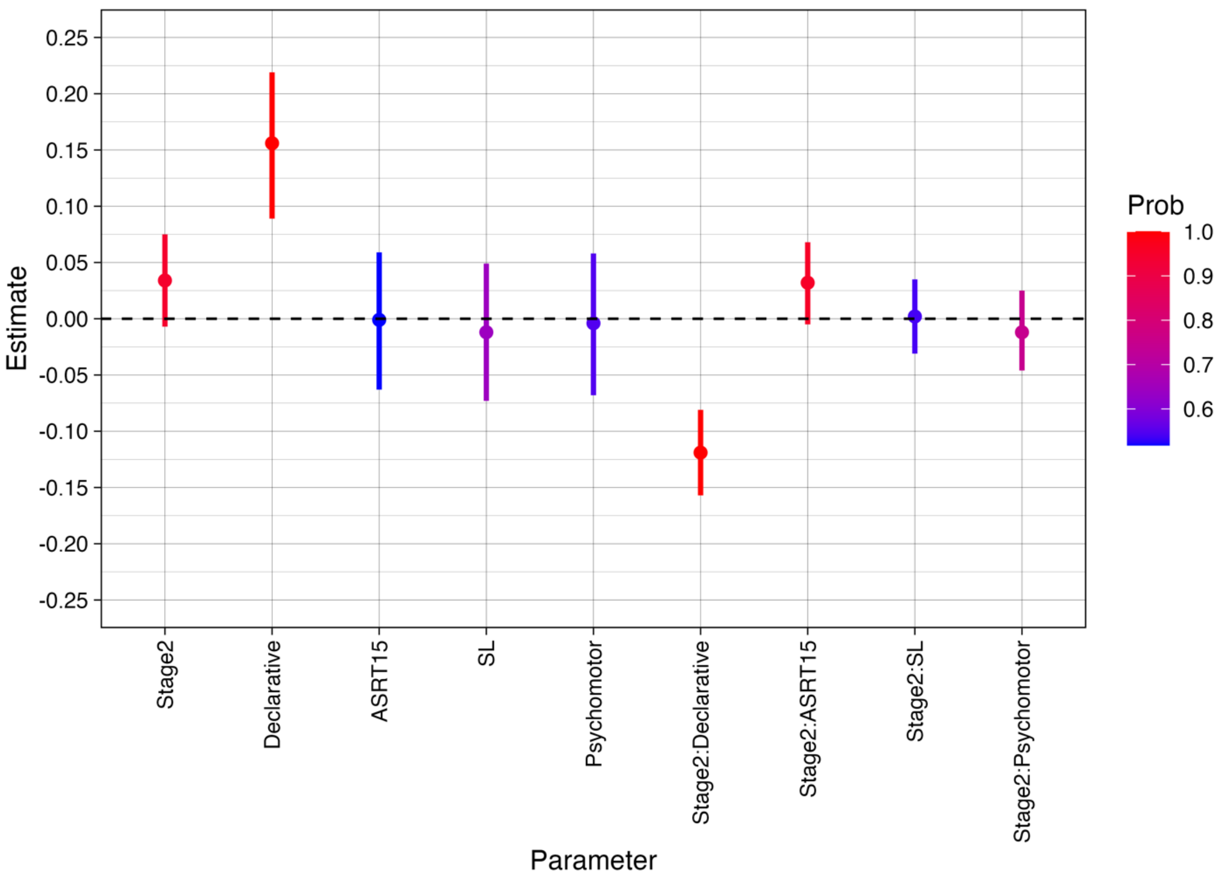


Figure 3.19. Posterior estimates of the fixed effects parameters in zero-one inflated beta GLMM for accuracy in production practice.

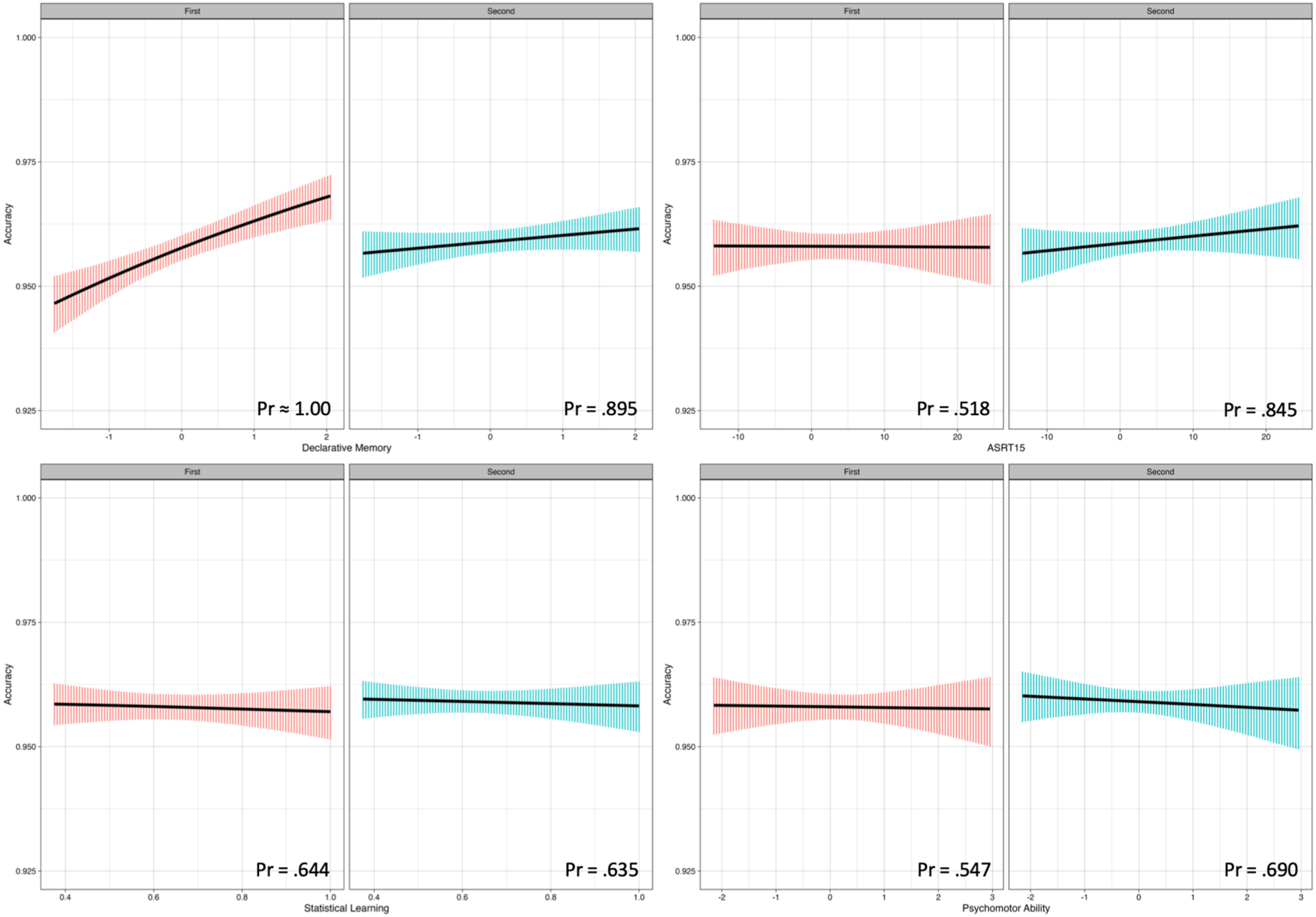


Figure 3.20. Posterior predictions of the relationship between production accuracy and the cognitive variables.

3.3.2.2 CV

Figure 3.21 shows the posterior estimates from the normal GLMM of the CV in production practice. Transitioning to Stage 2 (*b* = -0.009, 95% CrI [-0.022, 0.003], Pr[*b* < 0] = .925) was associated with smaller values of the CV, indicating that processing stability (operationalized by the CV) may be one aspect of skill performance that must be improved to transition to Stage 2. ASRT_{15} and psychomotor ability were similarly related to the CV in Stage 1 such that those participants who had higher procedural memory capacity or psychomotor ability tended to perform more stably in terms of the speed of performance (ASRT_{15}: *b* = -0.010, 95% CrI [-0.023, 0.002], Pr[*b* < 0] = .943; Psychomotor: *b* = 0.014, 95% CrI [0.002, 0.027], Pr[*b* > 0] = .986). Figure 3.22 shows the model-based predictions of the relationship between the CV and the cognitive individual difference variables. In Stage 2, the only predictor that seemed to be associated with the CV was psychomotor ability such that when with higher psychomotor ability, the participants were associated with smaller values of the CV (Pr[*b* > 0] = .927). No other predictors were firmly related to the CV: declarative memory (Stage 1: Pr[*b* < 0] = .757; Stage 2: Pr[*b* < 0] = .638), ASRT_{15} (Stage 2: Pr[*b* < 0] = .735), and SL (Stage 1: Pr[*b* > 0] = .605; Stage 2: Pr[*b* > 0] = .686).

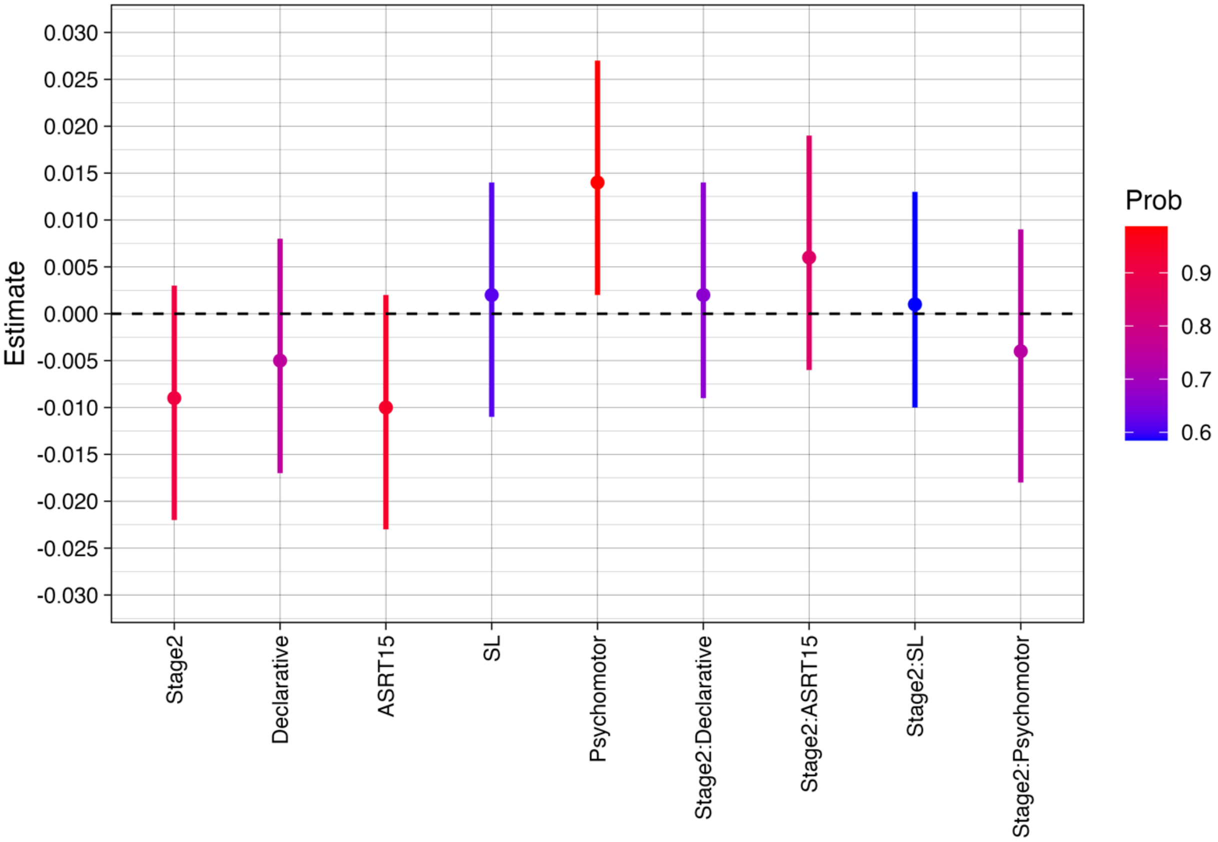


Figure 3.21. Posterior estimates of the fixed effects parameters in normal GLMM for CV in production practice.

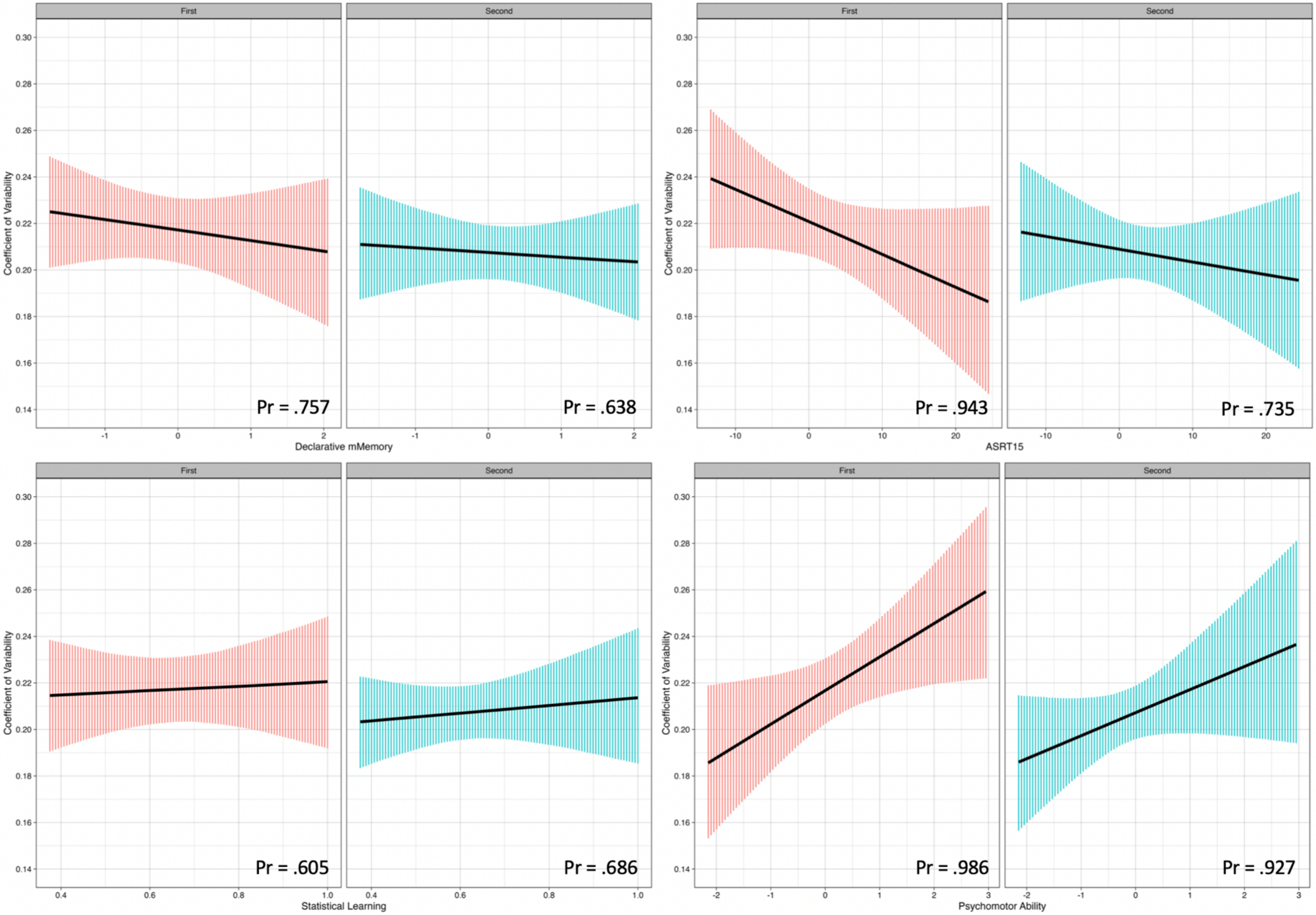


Figure 3.22. Posterior predictions of the relationship between the CV in production and the cognitive variables.

3.3.2.3 RT

Figure 3.23 displays the posterior estimates from the normal GLMM of RT in production practice. All predictor variables in the model were predictive of RT throughout the process of skill acquisition. In Stage 1, declarative memory capacity, ASRT_{15}, and psychomotor ability played a facilitative role in predicting RT such that those participants with higher-functioning declarative memory capacity (b = -0.099, 95% CrI [-0.135, -0.062], Pr[b < 0] ≈ 1.000), procedural memory capacity (ASRT_{15}: b = -0.069, 95% CrI [-0.109, -0.030], Pr[b < 0] = .999), or psychomotor ability (b = 0.154, 95% CrI [0.114, 0.193], Pr[b > 0] ≈ 1.000) tended to perform the skill faster than those participants with lower capacity on the ability. On the other hand, SL was negatively associated with RT, indicating that those with higher procedural memory capacity tended to perform more slowly in Stage 1 (ASRT_{15}: b = 0.069, 95% CrI [0.030, 0.107], Pr[b > 0] = .999; SL: b = 0.057, 95% CrI [0.018, 0.095], Pr[b > 0] = .997).

Figure 3.24 displays the model-based predictions of the relationship between RT and the cognitive individual differences in each learning stage. Declarative memory remained facilitative of promoting RT (Pr[b < 0] = .999), but the two-way interaction between declarative memory and Stage 2 showed that the effect was larger in Stage 1 than in Stage 2 (b = 0.032, 95% CrI [0.019, 0.045], Pr[b > 0] ≈ 1.000). Similarly, psychomotor ability maintained its influence on the participants' RT in Stage 2 (Pr[b > 0] ≈ 1.000), but its effect was more prominent in Stage 1 (b = -0.063, 95% CrI [-0.078, -0.049], Pr[b < 0] ≈ 1.000). This larger role of one's cognitive ability in Stage 1 also applied to ASRT_{15} (b = 0.052, 95% CrI [0.039, 0.064], Pr[b > 0] ≈ 1.000), but the effect weakened to the point that procedural memory was not influential anymore.

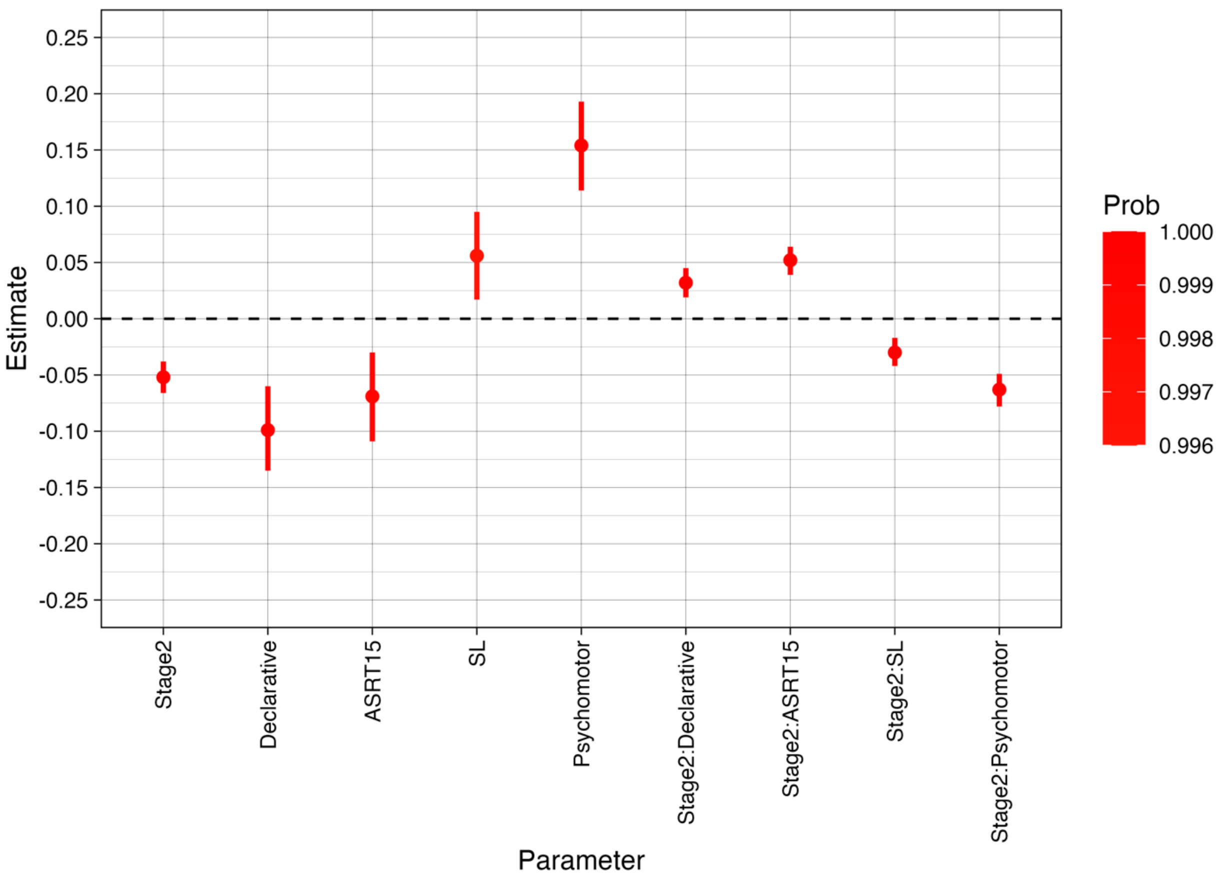


Figure 3.23. Posterior estimates of the fixed effects parameters in normal GLMM for RT in production practice.

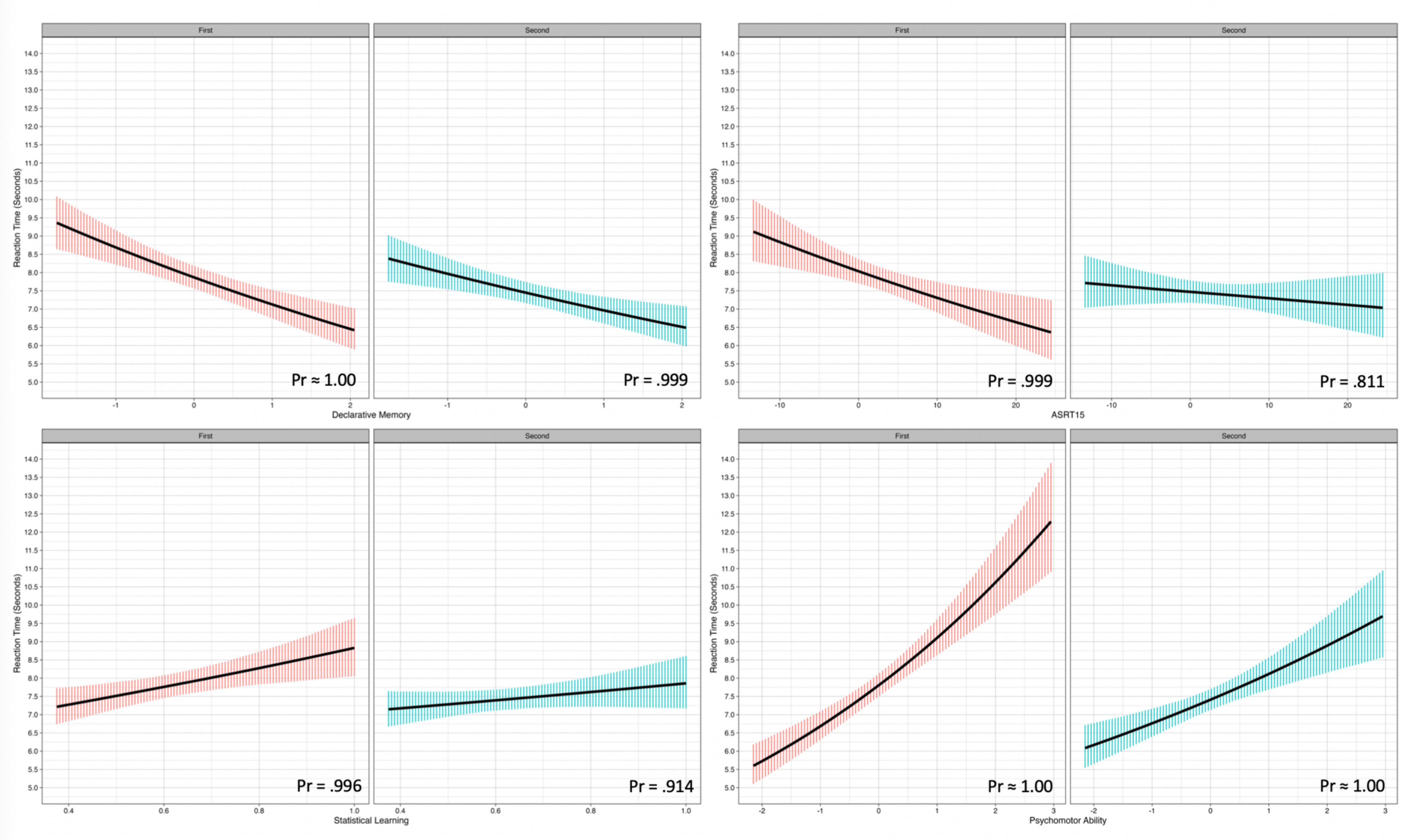


Figure 3.24. Posterior predictions of the relationship between production RT and the cognitive variables.

3.4 Summary of Results

Table 3.8 summarizes the relationship between cognitive individual difference variables and learning in each stage of skill acquisition found in this study. Based on the results from the HMM analysis, the regression analysis assumed three learning stages for comprehension skills, whereas production skills were assumed to have involved two stages. The table displays (a) the posterior probability of a given effect being present (** for Pr. ≥ .95 and * for .90 ≤ Pr. < .95) and (b) whether the effect is larger in one stage than in the other stage (using the inequality sign).

Overall, declarative memory showed the strongest tie with the acquisition of L2 skills in Mini-Nihongo. For comprehension practice, those participants with higher declarative memory capacity tended to perform more accurately and quickly, but the relationship gradually weakened as the participants transitioned from Stage 1 to Stage 2 and Stage 3, especially for RT. Declarative memory also played a facilitative role in promoting processing stability in Stage 2 and 3 of comprehension practice (and SL but to a much less extent). ASRT_{15}, on the other hand, did not bear any relationship with the CV, but it played a facilitative role in promoting speedup (RT decrease) in Stage 1 and 2. In contrast to declarative memory, however, ASRT_{15} remained equally influential in Stage 1 and 2. Psychomotor ability positively predicted the participants' RT in Stage 1 and Stage 2 such that those participants with higher-functioning psychomotor ability showed faster RT in Stage 1 and 2. For production practice, declarative memory similarly predicted accuracy and RT with a more prominent role in Stage 1. One interesting difference for production practice was that ASRT_{15} was predictive of the CV in Stage 1 such that those participants with higher procedural memory capacity tended to show more stability while performing the task in Stage 1. In Chapter 4, I discuss the findings in Table 3.8 in terms of the existing models of skill acquisition.

Table 3.8. Key takeaways from the results of the regression analysis.

	Comprehension								
	Accuracy			CV			RT		
	Stage 1	Stage 2	Stage 3	Stage 1	Stage 2	Stage 3	Stage 1	Stage 2	Stage 3
Declarative	**	**	**		**	*	**	>**	>**
ASRT$_{15}$							**	**	
SL						*			
Psychomotor							**	>**	
	Production								
	Accuracy			CV			RT		
	Stage 1	Stage 2	–	Stage 1	Stage 2	–	Stage 1	Stage 2	–
Declarative	**		–			–	**	>**	–
ASRT$_{15}$			–	*		–	**		–
SL			–			–	**!	*!	–
Psychomotor			–	**	*	–	*	>**	–

Note. ** for Pr. ≥ .95 and * for .90 ≤ Pr. < .95. The inequality signs (>) shows that the relationship was stronger in one stage than the other with the posterior probability larger than .95. The exclamation signs (!) indicates that the relationship was in the opposite direction from the predicted direction; For instance, for declarative memory, ASRT$_{15}$, and psychomotor ability, the effects were facilitative.

CHAPTER 4: DISCUSSION

In this chapter, I discuss the results of the current study in relation to the existing cognitive models of skill acquisition reviewed in Section 1.3. I first discuss the evidence found in the study for the parallelism between (cognitive) skill acquisition and L2 learning to support the widely upheld view in SLA that L2 learning can be considered another type of skill acquisition. Then, I will lay out the results in terms of the number and the nature of skill acquisition stages identified in the study. By discussing these two aspects of the process of L2 skill acquisition, I address the two research questions that guided the study (see Section 2.1).

4.1 Evidence for L2 Skill Acquisition

The power-law of practice and skill specificity are two well-attested empirical phenomena in skill acquisition research (Section 1.1). Consistent with the power-law of practice, the participants in this study decreased their RT as a power function of practice blocks for both comprehension (Figure 3.6) and production practice (Figure 3.7). I found that a regression of the logarithm of RT on the logarithm of practice blocks produced an almost perfect linear line for both comprehension (R^2 = .97) and production skills (R^2 = .96). Additionally, the power function was more consistent with the data than an exponential function (R^2 = .79 and .81 for comprehension and production skills, respectively), which was in contrast to the existing criticism of the power-law of practice (e.g., Haider & Frensch, 2002; Heathcote et al., 2020). These critics have objected that a power function does not apply well to unaggregated data and that an exponential function usually provides a better fit to unaggregated data than a power function. In my study, however, a power function was still a better approximation of the data than an exponential function even in the trial-level analysis (see Figure 3.5), although the fit of the power function did become substantially worse when it was applied to the unaggregated data.

The results of the study thus lend support to the power-law of practice in general but are consistent with the view that applying a power function to aggregated data (averaged across participants and/or items) may mask important patterns in how individual learners actually improve their skill.

There is no doubt that the power-law of practice is based on an idealization of how individual learners develop the knowledge of performing a skill. In L2 learning, R. Ellis (2015) provided an exemplary discussion on the role of idealization in formulating a theoretical model of L2 learning that is postulated to be generalizable across individual learners. In my study, it was clear that individual learners exhibited significant variability in how they improved the speed of performance (see Figure 3.3 and 3.4) and that their developmental paths did not necessarily follow a power function. An important question is “whether such an idealization [here, applied to how learners improve the speed of skill performance] is unwarranted and therefore should be abandoned or whether, in combination with a specification of the constraints and limits (i.e., de-idealization), it can continue to represent actual behavior” (R. Ellis, 2005, p. 204, my addition). Given the robust nature of the power-law of practice across a wide range of skills and task domains (Newell & Rosenbloom, 1981), abandoning it seems like throwing out a baby with the bathwater. The participants decreased their RT as a power function of practice even at the trial level, though not precisely, as shown in a close examination of Figure 3.3 and 3.4. Then, it seems logical to conclude that as long as one identifies how and why individuals may deviate from a power function, which are the “constraints” and “limits” in Ellis’s words, the power-law of practice could provide an informative model of how individual learners actually improve their performance through practice.

In skill acquisition research, some researchers have shown that the accuracy of performance (Anderson, 1995) and the standard deviation of RT (Logan, 1988, 1992, 2002) also follow the power-law of practice. In the current study, the power function applied fairly well to the standard deviation of RT (R^2 = .88 and .79 for comprehension and production skills), but it did not apply to the accuracy data (R^2 = .36 and .32). The poor fit of the power function to accuracy is likely because the participants were already highly accurate from the first block of practice (*M* = .91 and .93). However, the fact the standard deviation decreased following a power function suggests that the variance in RT decreased proportionately to the rate of the RT decrease (because they followed the same form of decrease). In this light, it comes as little surprise that the decrease in coefficient of variability (CV) of RT was minimal (*M* = 0.37 to 0.33 for comprehension; *M* = 0.22 to 0.20 for production) because RT and the standard deviation simply decreased to the same extent.

One unresolved issue in SLA is the question of how the CV changes as a result of practice. Segalowitz and Segalowitz (1993) originally posited that the decrease in the CV (in some meaningful way) can be used as an indication of higher processing stability and by extension, a higher degree of automaticity. Contrary to the previous assumption of a linear decrease in the CV following practice, Solovyeva and DeKeyser (2018) made the first observation that the CV could initially increase before it decreases as an index of automaticity. This was later confirmed in an experiment by Hui (2020), who showed that the change in the CV could follow an inverted U-shape over the course of development. In the current study, this patten of change (characterized by an initial increase followed by a decrease in the CV) only held true for the acquisition of production skills (Figure 3.9), but in comprehension practice, the CV first seemed to decrease and then increase. However, what was more striking was the extreme variability in how the CV changed for individual participants. Hence, the developmental pattern observed in the participant-averaged data (i.e., the blue line in Figure 3.8 and 3.9) may actually be due to an artifact of data averaging (This also seems to be the case for Hui, 2020; see Figure 2 in p. 343, where the group-level change in the CV followed an inverted U-shape, but there was extreme variability in how the CV change for individual learners). In the case of RT, individual curves resembled the group-level pattern (see Figure 3.3 and Figure 3.4), but for the CV, individual patterns looked nothing like their participant-averaged curve. This implies that when the variability in performance (CV) is concerned, one cannot make a generalized claim about how individual learners learn by simply looking at the group-level developmental patterns.

In Figure 3.3 and 3.4, most participants seemed to have reached an asymptote in terms of the speed (or RT) of performance. In skill acquisition research, this is generally taken as evidence of automaticity. However, if Segalowitz and Segalowitz (1993) were correct in that the evidence of automaticity must come from a simultaneous decrease in (a) RT and (b) the CV and (c) a positive correlation between RT and the CV (see Section 1.2.2), the results of my study suggest that the participants did not reach automaticity, as the CV barely decreased for the entire sample in both comprehension (0.37 [.35, .39] to 0.33 [0.31, 0.35]) and production practice (0.22 [0.20, 0.24] to 0.20 [.18, .23]), and the RT and CV positively correlated for production skills (r = .56 [.36, .71]) but not for comprehension skills (r = .18 [-.05, .41]). This is clearly at odds with the evidence of automaticity from the RT data (i.e., the asymptomatic performance by the participants), which begs the question as to which side of the evidence, RT or the CV, should be given more weight. Given that the power-law of practice has been widely replicated across many skill domains (Newell & Rosenbloom, 1981), it seems dubious to conclude that the participants did not achieve automaticity even long after they reached the point at which they could not improve the speed anymore. Hence, I would argue that processing stability is needed to achieve automaticity (in addition to the high speed of performance; see Segalowitz & Segalowitz, 1993), but empirically demonstrating such a phenomenon is highly challenging, at least using the CV. At the group level, evidence for the CV reduction is conflicting at best (but see below for potential moderator variables), and at the level of individual learners, variability looms large, suggesting that even the decrease observed in the group-level analysis may be an artifact of data averaging. To conclude, it is safe to state that the CV still needs further validation work as an index of processing stability (and automaticity, by extension). Additionally, future studies must be conducted in research settings where one can observe how the CV incrementally changes for individual participants so that the researcher can identify the exact function of the change. Below, I have provided further discussions on the use of the CV in L2 skill acquisition research (see Section 4.2.2).

4.2 Identifying the Skill Acquisition Stages in L2 Learning

I utilized the hidden Markov modeling (HMM) analysis to identify the number of learning stages that the participants likely would have gone through while learning and practicing how to comprehend and produce the target language. For comprehension practice, the analysis showed that the three-state model was most consistent with the data, while the two-state model provided the best summary of the data from production practice. As far as the number of stages is concerned, the results suggest that the learning mechanism of ACT-R (Anderson, 2007; Anderson et al., 2004, 2019; Anderson & Fincham, 1994; Anderson & Lebiere, 1998) is most explanatory for the acquisition of comprehension skills. However, the same results also suggest that the acquisition of production skills is more consistent with the learning mechanism posed by the CMPL theory (Bajic & Rickard, 2011; Rickard, 1997, 1999, 2004). Note that further in the discussion, I will exclude the Race model because the one-stage model was found categorically incompatible with the current dataset (see Table 3.4 and 3.6). However, the CMPL theory is an extension of the Race model, so the primary learning mechanism of the Race model (i.e., instance-based learning of previous performance episodes) is represented in the CMPL theory.

Of course, the number of learning stages is distinct from the learning mechanisms that may underlie them. Just because the number of learning stages matched a given model of skill acquisition does not necessarily mean that the result validates the specific learning mechanism in that model. To gain more insight into the learning mechanisms, I conducted a regression analysis to identify which cognitive individual difference variables predicted learning at each stage identified by the HMM analysis. The underlying assumption was that if a cognitive variable correlated with a learning outcome in a given learning stage, the specific cognitive ability was responsible for learning at the specific stage. Although not explicitly stated, this type of interpretation abounds in SLA (see DeKeyser, 2012 for a review), and the previous research on individual differences in L2 skill acquisition (e.g., Li, 2017; Maie, 2021; Pili-Moss et al., 2021; Y. Suzuki, 2018) rested on the same premise to make an inference about the underlying learning mechanism.

4.2.1 Accuracy

For the accuracy of performance, the only reliable predictor was one's declarative memory capacity. Participants with higher declarative memory scores also obtained higher accuracy scores. This finding held true throughout the process of skill acquisition for comprehension practice, but declarative memory had the largest impact in Stage 1. For production practice, the effect was restricted to Stage 1. Returning to the cognitive models of skill acquisition, both the CMPL theory and ACT-R expect that general-purpose learning mechanisms (including declarative memory) are driving the initial stage of skill acquisition, which, in this study, corresponded to Stage 1 for comprehension and production practice. However, the two models offer different accounts of (a) the continuing role of declarative memory in all three stages of comprehension practice and (b) the specific effect of declarative memory in Stage 1 of production practice. Because the CMPL theory posits instance-based learning as the primary mechanism, it naturally expects declarative memory to be dominant throughout the process of skill acquisition. On the other hand, ACT-R expects proceduralization to take place between Stage 1 and Stage 2 (and to be completed following the transition to Stage 3), but ACT-R additionally allows for a complementary mechanism of declarative strengthening, which improves the accuracy and the speed of retrieving declarative knowledge (see Section 1.3). Although skill performance gradually comes to rely predominantly on procedural memory in later stages of learning, ACT-R does not exclude the use of declarative memory in those stages in case the task-specific production rules may lead to performance errors (Anderson, 1982, 1983b). Hence, both the CMPL theory and ACT-R account for the continuous (but gradually decreasing) impact of declarative memory on the participants' skill performance. The models, however, assign different weights to the role of declarative memory in Stage 2 and 3. In contrast, the specific effect of declarative memory in Stage 1 for production practice can be explained only by the learning mechanism of ACT-R (i.e., the transition from declarative to procedural memory) because the CMPL theory predicts declarative memory to be the primary mechanism in all stages of learning.

4.2.2 Coefficient of Variability

Among the three dependent variables analyzed in the current study, the CV was least susceptible to the effect of cognitive individual differences (see Figure 3.14 and 3.17). First, declarative memory reliably predicted the CV in Stage 2 and 3 for comprehension practice.

Those participants with higher declarative memory capacity performed more stably when comprehending Mini-Nihongo. In production practice, $ASRT_{15}$ and psychomotor ability played a facilitative role in promoting processing stability in Stage 1 (and also in Stage 2 for psychomotor ability). While the finding on declarative memory in comprehension practice is consistent with both the CMPL theory (i.e., instance-based learning) and ACT-R (i.e., declarative strengthening), the finding on procedural memory in production practice is only consistent with the latter model (i.e., proceduralization). In part, the superiority of ACT-R over the CMPL theory owes to the flexibility of the cognitive mechanisms espoused by ACT-R, allowing for the complementary role of declarative memory in later stages of skill acquisition. Nevertheless, the major avenue to improving one's skill performance in ACT-R is by developing skill-specific production rules through knowledge compilation; hence, it is unclear why procedural memory did not show any more substantial link with the CV. As discussed below, this may be due to the validity and reliability of the CV as a measure of proceduralization (or automatization) and $ASRT_{15}$/SL as a measure of one's procedural memory capacity.

First, the CV requires further validation work to pin down what exactly its changes may signify. In my study, the CV was defined as an index of processing (or performance) stability and hence as an indication that proceduralization has taken place (as opposed to the original definition by Segalowitz & Segalowitz, 1993; see Section 1.2.2). Conceptually, it makes sense to justify the use of the CV because its mathematical underpinning maps directly onto the concept of processing stability: the CV quantifies the variability in one's speed of performance corrected for the average speed for the person (see Segalowitz & Segalowitz, 1993). However, adopting the CV in empirical studies has proved highly challenging. While some SLA studies have shown that practice can lead to reductions in the CV (Hulstijn et al., 2009 for vocabulary; Lim &

Godfroid, 2015; Segalowitz et al., 1998; Pili-Moss et al., 2020), others have obtained null results, showing that the participants did not achieve automaticity (Hulstijn et al., 2009 for grammar; Hui, 2020; Maie, 2021). Partly, the disparity in results could be explained by methodological differences, including the target of learning (e.g., vocabulary vs. grammar), the operationalization of practice (e.g., between-subjects: participants of different proficiency levels; within-subjects: observing the same participants repeatedly), and the granularity with which the effect of practice was documented (e.g., pretest-posttest vs. the number of practice trials/blocks). Among the existing studies, my study and Pili-Moss et al. (2020) were the most similar in design (i.e., grammar, within-subjects, and 1,440 practice trials), yet the results of the present study do not coincide with those of Pili-Moss et al. If, as DeKeyser (1997, p. 216) discussed, proceduralization can take place in the early phases of L2 practice, it is all the more puzzling why some studies (including the current study) failed to find evidence of proceduralization in the CV. One potential route to tackling the issue may be to synthesize the data from the existing studies using a meta-analytic approach and investigate how conceptual and methodological variables moderate the development of the CV.

Second, the results on the CV could also be attributed to the uncertainty over the validity and reliability of measuring one’s procedural memory ability. The assessment of procedural memory ability has been the subject of heated discussions in recent SLA research (e.g., Buffington et al., 2021; Buffington & Morgan-Short, 2019; Perruchet, 2021). This is a timely and important endeavor given the prominent role of procedural memory as one of the primary cognitive mechanisms in major theoretical approaches in SLA. These approaches include the skill acquisition theory (DeKeyser, 2020; Lyster & Sato, 2013; Y. Suzuki, 2022) and the D/P model (Hamrick et al., 2018; Ullman, 2004, 2016, 2020). A recent study by Buffington et al.

(2021) showed that three procedural memory tasks frequently used in L2 research (i.e., the alternating serial reaction task, Tower of London task, and weather prediction task) failed to show convergent validity. Similarly, Godfroid and Kim (2021) found that their four procedural memory measures, including ASRT and SL that were equivalent to the tasks in my study, did not correlate and converge to the same latent variable. Godfroid and Kim found that the participants' scores from ASRT was the only significant predictor of the latent variable of implicit knowledge. In my study, ASRT_{15} was the only (facilitative) predictor of the CV in Stage 1, which is consistent with Godfroid and Kim, but this association was only present for production practice. One caveat of the ASRT scores in both the current study and the previous research is that the average scores often revolve around 0; that is, the mean difference between the pattern and random trials tends to be close to 0 (e.g., Buffington et al., 2021: $M = 2.99$ ms; Godfroid & Kim, 2021; $M = -1.15$ ms; the current study: $M = -2.54$ ms). Perruchet (2021) pointed out that such a trend in the data indicates that there was no learning effect at the group level, and he questioned the validity of using the ASRT scores (under this condition) to operationalize the participants' procedural memory ability. Hence, it is wise to be cautious when one interprets the results on the CV in the current study.

4.2.3 Reaction Time

In contrast to the CV, the participants' reaction time (RT) was most susceptible to the effect of cognitive individual differences (Figure 3.18 and 3.24). This finding was unexpected given the close relationship between RT and the CV (i.e., RT is part of the equation to compute the CV), and hence I expected some extent of overlap in the results. Instead, the current results indicated that that RT and the CV tapped into different aspects of the process of L2 skill acquisition. Declarative memory was predictive of RT in all stages of learning for both comprehension and production practice, but its impact was most influential in Stage 1. This result converges with that on accuracy in that they both signal the dominant role of declarative memory in Stage 1. ASRT$_{15}$, on the other hand, proved equally important for Stage 1 and 2, at least for comprehension practice. The fact that ASRT$_{15}$ was predictive of RT only in Stage 1 for production practice may be because Stage 1 of production practice carries the characteristics of both Stage 1 and 2 of comprehension practice (see below, Section 4.2.4). As with accuracy of performance, it is evident that procedural memory begins to support the acquisition of L2 skills in Mini-Nihongo. This is clear evidence for the declarative-procedural transition conceptualized in ACT-R. Interestingly, SL also seemed to bear a similar relationship with RT in production practice, but it was in an unexpected direction. Those participants with higher scores on SL actually tended to perform the task more slowly. It is unclear why SL showed the inverse relationship with RT. While SL correlated positively with declarative memory scores (r = .21 [-.03, .43], it correlated negatively with psychomotor ability scores (r = -.25 [-.47, -.01]) and did not show any relationship with ASRT$_{15}$ (r = -.01 [.-.25, .23]). My initial hypothesis was that SL unexpectedly tapped into one's declarative memory rather than procedural memory, but this does not explain why declarative memory consistently played a facilitative role in promoting the speed of performance, whereas, for SL, the effect was in the reverse direction. As discussed in the previous section, this surprising finding may owe to the uncertainty over the measurement of procedural memory capacity.

4.2.4 Synthesis of the Findings

The analysis of the role of cognitive individual differences in L2 skill acquisition carried several implications for the underlying learning mechanism that may be at work in each stage of skill acquisition. The clearest evidence came by for the predominant role of declarative memory in the first stage of learning (Stage 1). Although the results were less clear for the CV, declarative memory potently promoted one's accuracy and speed of performance (RT) in all stages of learning with an exception that it was only predictive of accuracy in Stage 1 for production practice, which only ACT-R could explain.

The second most consistent evidence was on the emerging role of procedural memory in Stage 2. $ASRT_{15}$ predicted the speed of performance in Stage 1 and 2 for comprehension practice, but unlike for accuracy, it remained equally influential in both stages. For production practice, $ASRT_{15}$ promoted the speed and the stability of performance in Stage 1 (only) such that those participants with higher procedural memory capacity were associated with lower values of RT and the CV. Again, these results map nicely onto the learning mechanism in ACT-R, especially onto the mechanism of proceduralization. Hence, the results of the study are very much in line with the three-stage model of skill acquisition. Then, the question must be asked as to why the two-stage model was found most consistent with the data for production practice. Currently, there are three possible explanations. These three possibilities must be interpreted in light of the fact that (a) declarative memory was only predictive of production accuracy in Stage 1, while the same was true for Stage 1, 2, and 3 of comprehension practice, (b) $ASRT_{15}$ promoted RT only in Stage 1, while the same was true for Stage 1 and 2 for comprehension practice, and (c) $ASRT_{15}$ predicted the CV only in Stage 1 of production practice, which was not observed for comprehension practice but is expected (by ACT-R) to take place in the second stage of learning. As far as the nature of the stages was concerned, Stage 1 of production practice thus carried the same characteristics as both Stage 1 and 2 of comprehension practice. Hence, it is logical to speculate that those two stages observed in comprehension practice may have been conflated in a single stage in production practice.

The first possibility is that aggregating adjacent practice trials into bins may have conflated Stage 1 and 2 in a single stage. In the HMM analysis, I grouped every six trials into a single bin (for production practice) so that the probability of each participant residing in a given state becomes scalable. As claimed by DeKeyser (1997), it is the transition from the first to the second stage that can be completed within a relatively brief period of time. A close examination of Figure 3.11 (see p. 121) suggests, however, that this data wrangling was unlikely to have affected the estimation of the number of stages because if the participants on average required 37 trials to transition to Stage 2 in comprehension practice, they must have required more trials for production practice, which was more complex in terms of both cognitive and task demands. Hence, simply aggregating six trials into a bin could not have masked the transition that would have taken place after several dozens of practice trials.

The second possibility, which may be more credible, is that the maze task for production practice was not suited to observe three stages of learning. In the task, the participants produced a sentence in Mini-Nihongo by making a single word-level decision 11 times (see Section 2.2.4 and Figure 2.5). Although I expected the task to tap into the participants' productive knowledge of morphosyntactic structures, cognitive processes involved in the task may have been restricted to those that were specific to lexical processing. In other words, while sentence production in general involves the processing and integration of lexical items and syntactic frames (e.g., Bock, 1987; Garrett, 1980; Levelt, 1989), the maze task in the study may have tapped only into cognitive processing at the lexical level. As a point of comparison, Maie (2021) found that skill acquisition in L2 vocabulary learning (i.e., retrieving the knowledge of form-meaning mappings) is most likely to consist of two phases with the first stage controlled by declarative memory and the second stage by procedural memory. This study offers a good point of comparison for the current study because it utilized a word-picture matching task in which participants chose a word out of two options that best matched the meaning of a picture prompt. Since the participants in the current study similarly chose a word out of two options based on a picture prompt, the maze task in this study may be considered an extension of the word-picture matching task in Maie (2021). While the current study not only dealt with learning of vocabulary items but also the acquisition of morphosyntactic structures (which is different from Maie, 2021), the maze task used to examine the learning necessitated the participants to make a decision at the word level (or at the morpheme level when the decision pertained to grammar). Hence, it is true that the participants constructed a sentence by making 11 word-level decisions, which involved vocabulary and grammar, but cognitive processing involved in those decisions may have been specific to lexical items or morphemes. Hence, one conclusion is that the HMM analysis indeed excavated how the participants acquired production skills, but the maze task adopted in the study only allowed them to develop the skills at the lexical (or morphemic) level of processing.

Assuming that production skills were indeed acquired in two stages (though only at the level of lexical processing), which existing model is most consistent with the results of the regression analysis? The results summarized in Table 3.8 provide good evidence that the declarative-procedural transition not only took place in comprehension practice but also in production practice. This learning process can only be explained by the learning mechanism of ACT-R. If this is true, what does it mean that production practice was missing what it seemed to be the intermediary stage (Stage 2) in comprehension practice? In an attempt to explain the best-fitting two-state HMM for L2 vocabulary skill acquisition, Maie (2021) pointed out that making a decision on lexical items is cognitively simpler and hence less attentionally demanding than doing the same on rule-based grammar rules, which could require (far) fewer practice trials for proceduralization to complete and hence directly send the learners to the last stage of skill acquisition. Although the maze task used in the current study tested the participants' knowledge of both vocabulary and grammar, the decision made to produce a response was always specific to the lexical (or morphemic) level of processing. In this light, it is possible to project the results of Maie (2021) onto those of the current study.

The last possibility of finding two stages for production practice, which is as credible as the second option, is that summing eleven (word-level) responses to calculate the sentence-level RT in production practice may have concealed the incremental changes in the participant's skill performance. In comprehension practice, each trial was based on a single response, each of which reflected the entirety of sentence processing from the beginning to the end. In contrast, a single trial in production practice consisted of eleven responses, and each of these responses required a decision on different aspects of Mini-Nihongo (i.e., four nouns, four verbs, two numerical classifiers, and three case-markers). It is possible that by mixing responses of different kinds to form a trial-level data point, I may have masked the subtlety contained in the data. Nevertheless, the comprehension and production tasks dealt with the same set of sentences, so it is logically sound to expect that the data in production practice (computed by summing eleven responses) should encode the same (or at least comparable) information as those in comprehension practice. Based on the current results, this assumption seems to be false, and it seems that the sum of parts does not necessarily equal the whole of sentence processing.

To conclude this chapter, the results of the present study lend support to the influential three-stage model of L2 skill acquisition (DeKeyser, 2020; Lyster & Sato, 2013; Y. Suzuki, 2022), but the evidence is restricted to comprehension practice. However, this does not mean that the acquisition of production skills takes places in a different number of stages. Rather, the finding for production practice is confounded by the specific design of the task used to gauge the participants' production skills. One interesting corollary of conceptualizing L2 skill acquisition in terms of ACT-R is that the results of the current study can be compared with those of previous studies. For instance, DeKeyser (1997) claimed that in his study, participants were likely to have gone through proceduralization as early as by the first 16 trials for comprehension practice. In the current study, the participants transitioned from Stage 1 to Stage 2 at Trial 38 and from Stage 2 to Stage 3 at Trial 288. This meant that on average, proceduralization required 37 trials to initiate and additional 251 trials to complete (287 trials in total). DeKeyser (1997) only inferred the timing of proceduralization by examining the deviation of RT data from the power-law of practice, which was based on the assumption that performance based on procedural memory (or skill-specific production rules) follows a power function whereas performance based on declarative knowledge largely does not (see Anderson 1982, 1983b). In contrast, I identified the stage transitions using a HMM analysis and followed up with the individual difference analysis to identify the nature of the learning stages. Hence, the current study may present more objective and reliable estimates of the timing of the stage transitions.

CHAPTER 5: CONCLUSION AND LIMITATIONS

This study had several conceptual and methodological limitations that need to be considered when interpreting the results. First, this study was experimental in nature, and hence the findings should not be blindly extrapolated to L2 learning in classroom contexts. In particular, the automaticity of comprehension and production skills shown in the study is specific to the experimental tasks adopted, and hence the observed developmental paths may not generalize to other linguistic tasks. Additionally, the findings on the declarative-procedural transition are likely specific to the adult and educated sample of participants recruited in this study, who were explicitly taught the vocabulary and grammar of the language before engaging in comprehension and production practice. When one deals with a younger population of L2 leaners and/or a context in which learners are exposed to the language through communicative use, the learning mechanism proposed in ACT-R may not necessarily be the best to account for the learning process (as observed by DeKeyser, 1997).

Second, it is unclear to what extent the results on the CV and procedural memory (ASRT_{15} and SL) may be reliable and hence speak to the nature of L2 skill acquisition identified by the regression analysis. As discussed in Section 4.2.2, the construct of the CV and the assessment of procedural memory ability require further research to investigate the validity and reliability of the measurement. The findings presented here must be interpreted with a caveat that future research may find different results when one chooses to adopt a different operationalization of processing stability or a different assessment of procedural memory capacity.

Third, this study only examined cognitive individual difference variables pertaining to three dimensions of cognitive ability that were projected by Ackerman (1988, 1992). Future research of L2 skill acquisition can sample other dimensions of cognitive abilities to extend the findings of the current study. One candidate may be to examine the role of short-term or working memory capacity. The mechanism of knowledge compilation in ACT-R consists of two sub-processes, composition and proceduralization. In the current study, I only focused on the process of proceduralization, but composition is also an integral process to knowledge compilation. When a skill involves a complex procedure and hence must be carried out by multiple production rules, the process of composition collapses the productions into a single large production before proceduralization restructures the macro-production to the form that is specific to the skill (see Section 1.3.1). In order for composition to take place, learners must practice applying the productions while holding the relevant declarative information in working memory (Anderson, 1982). The prediction is that the more declarative information one can maintain in working memory, the larger number of production rules the process of composition can unify. Hence, when the target skill is complex, those with higher working memory capacity should experience the faster (or more efficient) rate of proceduralization. In the regression analysis, this effect may surface as a three-way interaction of practice trials, working memory, and procedural memory.

Fourth, the estimation method used for the HMM analysis was limited in that when applied to a high number of practice trials such as the one in the current study (i.e., 524 trials), the algorithm encountered an issue of scalability. This is the reason why I had to aggregate every four or six practice trials into a bin. In future, a more flexible scaling method (e.g., the Naïve Bayes algorithm) needs to be implemented so that the trial-level data can be directly analyzed.

Fifth, the experimental tasks used in the study only allowed me to investigate the acquisition of vocabulary and grammar rules of Mini-Nihongo as an integrated whole. One interesting avenue for future research is to investigate how L2 knowledge of vocabulary and grammar develops independently and interactively through practice. A recent study by Monaghan, Ruiz, and Rebuschat (2021) is a good example of devising an experiment such that they can separate the simultaneous acquisition of vocabulary and grammar skills. In addition, there are only few studies in L2 skill acquisition research that investigated the acquisition of skills other than vocabulary and morphosyntax (cf. Li, 2017; Li & DeKeyser, 2017). When it comes to longitudinal research that documents the effect of practice across a large number of trials (e.g., Cornillie et al., 2017; DeKeyser, 1997; Ferman et al., 2009; Pili-Moss et al., 2020), there is no study that has investigated the acquisition of L2 skills other than morphosyntax. Hence, longitudinal research of L2 skill acquisition that focus on linguistic targets other than morphosyntax is an avenue for future research.

Lastly, while ACT-R was most consistent with the findings of the current study, there were many details in the results that defied an explanation. In part, this is due to the nature of psychological research where one cannot expect all results to be in line with the researcher's predictions. However, some of the unexpected results in this study are likely attributed to the fact that the level at which I analyzed the theoretical models of skill acquisition was not appropriate to capture the development of automaticity specific to the context of the current study (i.e., the target structures selected as the learning material, the tasks used to elicit skill performance, etc.). This is most notably evident in the finding that the two-stage model was found most consistent with production practice because the experimental task used to elicit production data was only able to tap into cognitive processing at the lexical level.

Most accounts of L2 skill acquisition to date (DeKeyser, 2020; Lyster & Sato, 2013; Suzuki, 2022) describe the process of L2 skill acquisition by importing the mechanism of learning from the existing cognitive models (e.g., declarative/procedural memory, proceduralization, and restructuring). As discussed in Section 1.5, L2 learning is often more complex in scope than typical skill acquisition studied in cognitive psychology (e.g., arithmetic tasks, associative memory tasks) because successful performance in L2 requires the coordination of multiple levels of linguistic knowledge (e.g., phonology, vocabulary, morphosyntax, and pragmatics). A cautionary tale from the results on production practice is that depending on linguistic targets (e.g., vocabulary vs. morphosyntax), learners may experience different numbers of learning stages and different types of learning mechanisms. While the existing cognitive models of skill acquisition (including the three-stage model) have been invaluable as a guiding framework of what skill acquisition make look like in L2 learning, a skill acquisition theory specific to the process of L2 learning may be necessary to further advance our understanding of the process of L2 skill acquisition. In particular, detailed theorization is necessary as to what kinds of cognitive process may be involved in a specific linguistic task learners require to perform for L2 learning. In cognitive psychology, researchers often analyze the process of skill acquisition not only at the level of learning mechanisms (e.g., declarative-transition-procedural) but also at the level of exact cognitive processes entailed by the experimental task at hand. As a case in point, Tenison, Fincham, and Anderson (2016) investigated how transitioning through stages of learning across trials leads to discrete changes in cognitive processing within a trial. In the context of the Pyramid problem (e.g., 8$3 → 8 + 7 + 6 = 21), the researchers demonstrated that as learners move through the three phases of learning (i.e., the cognitive, associative, and autonomous stage), the time they spent to carry out each step of cognitive processing (i.e., encoding-solving-responding) effectively changed. By applying the HMM analysis to fMRI data, Tenison et al. showed that in the cognitive stage, the learners took the longest time solving each problem, but as they moved to the associative stage, this solving time reduced to almost zero. In the autonomous stage, the learners even reduced the time to encode a problem because the skill of solving the Pyramid problem became almost like a reflex at that point.

Since the cognitive revolution in language science (Chomsky, 1959), linguistics has witnessed a countless number of theoretical models of how a language (including L2) can be processed, comprehended, and produced. L2 skill acquisition research thus has a fertile ground to base its own theorical models, and developing a model as detailed as the one exemplified by Tenison et al. (2016) is certainly not impossible. In addition, recent developments in ACT-R (Anderson, 2007; Anderson et al., 2004, 2019) not only offer a model of skill acquisition at the mechanism level but also at the level of specific cognitive processes that may be executed when performing a cognitive task. Hence, the newer versions of ACT-R can also be a starting point. Such an endeavor is surely difficult and likely more arduous than doing so in cognitive psychology, but only through such an effort can we do justice to the complexity of L2 learning.

To conclude, this dissertation study was the first in SLA to test and validate the three-stage model of L2 skill acquisition derived from cognitive psychological research. By combining the HMM analysis with the regression analysis, I identified the number and the nature of skill acquisition stages in L2 learning. Overall, the findings of the study lend support to the three-stage model of L2 skill acquisition, but its proposed mechanisms may have to be revised to suit to the specific cognitive processes involved in L2 learning. Furthermore, when applying the skill acquisition theory (or variants thereof) to L2 learning, one may have to analyze the theory not only at the mechanism level (e.g., declarative and procedural memory) but also at the level of cognitive processes that are specific to language tasks at hand. L2 learning thus may require a skill acquisition theory specific to its learning process.

REFERENCES

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information-processing perspectives. *Psychological Bulletin*, *102*(1), 3–27. https://doi.org/10.1037/0033-2909.102.1.3

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General*, *117*(3), 288–318. https://doi.org/10.1037/0096-3445.117.3.288

Ackerman, P. L. (1990). A correlational analysis of skill specificity: Learning, abilities, and individual differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(5), 883–901. https://doi.org/10.1037/0278-7393.16.5.883

Ackerman, P. L. (1992). Predicting individual differences in complex skill acquisition: Dynamics of ability determinants. *Journal of Applied Psychology*, *77*(5), 598–614. https://doi.org/10.1037/0021-9010.77.5.598

Ackerman, P. L., & Cianciolo, A. T. (2000). Cognitive, perceptual-speed, and psychomotor determinants of individual differences during skill acquisition. *Journal of Experimental Psychology: Applied*, *6*(4), 259–290. https://doi.org/10.1037//1076-898x.6.4.259

Ackerman, P. L., Kanfer, R., & Goff, M. (1995). Cognitive and noncognitive determinants and consequences of complex skill acquisition. *Journal of Experimental Psychology: Applied*, *1*(4), 270–304. https://doi.org/10.1037/1076-898X.1.4.270

Adams, J. A. (1957). The relationship between certain measures of ability and the acquisition of a psychomotor criterion response. *The Journal of General Psychology*, *56*(1), 121–134. https://doi.org/10.1080/00221309.1957.9918366

Adams, J. A. (1961). The second facet of forgetting: A review of warm-up decrement. *Psychological Bulletin*, *58*(4), 257–273. https://doi.org/10.1037/h0044798

Adams, J. A. (1987). Historical review and appraisal of research on the learning, retention, and transfer of human motor skills. *Psychological Bulletin*, *101*(1), 41–74. https://doi.org/10.1037/0033-2909.101.1.41

Adi-Japha, E., Karni, A., Parnes, A., Loewenschuss, I., & Vakil, E. (2008). A shift in task routines during the learning of a motor skill: group averaged data may mask critical phases in individuals' acquisition of skilled performance. *Journal of Experimental Psychology: Learning memory and Cognition*, *34*(6), 1544–1551. https://doi.org/10.1037/a0013217

Akamatsu, N. (2008). The effects of training on automatization of word recognition in English as a foreign language. *Applied Linguistics*, *29*(2), 1–19. https://doi.org/10.1017/S0142716408080089

Allen, L. Q. (2000). Form-meaning connections and the French causative: An experiment in processing instruction. *Studies in Second Language Acquisition*, *22*(1), 69–84. https://doi.org/10.1017/S0272263100001030

American Council on the Teaching of Foreign Languages. (2012). *ACTFL proficiency guidelines*. Retrieved from https://www.actfl.org/sites/default/files/guidelines/ACTFLProficiencyGuidelines2012.pdf

Anderson, J. R. (1981). Interference: The relationship between response latency and response accuracy. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 326–343. https://doi.org/10.1037/0278-7393.7.5.326

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. (1983a). Acquisition of proof skills in geometry. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (pp. 191–219). New York, NY: Springer.

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

Anderson, J. R. (1986). Knowledge compilation: The general learning mechanism. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning : An artificial intelligence approach* (Vol. 2, pp. 289–310). Los Altos, CA: Morgan Kaufman.

Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem situations. *Psychological Review*, *94*(2), 192–210. https://doi.org/10.1037/0033-295X.94.2.192

Anderson, J. R. (1992). Automaticity and the ACT theory. *American Journal of Psychology*, *105*(2), 165–180. https://doi.org/10.2307/1423026

Anderson, J. R. (1995). *Learning and memory: An integrated approach.* New York, NY: John Wiley & Sons.

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

Anderson, J. R. (2012). Tracking problem solving by multivariate pattern analysis and Hidden Markov Model algorithms. *Neuropsychologia*, *50*(4), 487–498. https://doi.org/10.1016/j.neuropsychologia.2011.07.025

Anderson, J. R., Borst, J. P., Fincham, J. M., Ghuman, A. S., Tenison, C., Zhang, Q. (2018). The common time course of memory processes revealed. *Psychological Science*, *29*(9), 1463–1474. https://doi.org/10.1177/0956797618774526

Anderson, J. R., Betts, S., Bothell, D., Hope, R., & Lebiere, C. (2019). Learning rapid and precise skills. *Psychological Review*, *126*(5), 727–760. http://dx.doi.org/10.1037/rev0000152

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060. http://dx.doi.org/10.1037/0033-295X.111.4.1036

Anderson, J. R., & Fincham, J. M. (1994). Acquisition of procedural skills from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1322–1340. https://doi.org/10.1037//0278-7393.20.6.1322

Anderson, J. R., & Fincham, J. M. (2014). Discovering the sequential structure of thought. *Cognitive Science*, *38*(2), 322–352. https://doi.org/10.1111/cogs.12068

Anderson, J. R., Kline, P. J., & Beasley, C. M. (1980). Complex learning processes. In R. E. Snow, P. E. Federico, & W. E. Montague (Eds.), *Aptitude, learning, and instruction* (Vol. 2, pp. 199–235). New York, NY: Routledge.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates.

Anderson, J. R., Zhang, Q., Borst, J., Walsh, M. W. (2016). The discovery of processing stages: Extension of Sternberg's method. *Psychological Review*, *123*(5), 481–509. https://doi.org/10.1037/rev0000030

Andringa, S., de Glopper, K., & Hacquebord, H. (2011). Effect of explicit and implicit instruction on free written response task performance. *Language Learning*, *61*(3), 868–903. https://doi.org/10.1111/j.1467-9922.2010.00623.x

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321–324. https://doi.org/10.1111%2F1467-9280.00063

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bock, K. (1987). Exploring levels of processing in sentence production. In G. Kempen (Ed.), *Natural language generation* (pp. 351–363). Dordrecht, The Nerthelands: Martinus Nijhoff Publishers.

Bokander, L., & Bylund, E. (2020). Probing the interval validity of the LLAMA language aptitude tests. *Language Learning*, *70*(1), 11–47. https://doi.org/10.1111/lang.12368

Borst, J. P., Ghuman, A. S., & Anderson, J. R. (2016). Tracking cognitive processing stages with MEG: A spatio-temporal model of associative recognition in the brain. *Neuroimage*, *141*(1), 416–430. https://doi.org/10.1016/j.neuroimage.2016.08.002

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

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*(s3), 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.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York, NY: Springer-Verlag.

Buxton, C. E., & Humphreys, L. G. (1935). The effect of practice upon intercorrelations of motor skills. *Science*, *81*, 441–442. https://doi.org/10.1126/science.81.2105.441

Cadierno, T. (1995). Formal instruction from a processing perspective: An investigation into the Spanish past tense. *The Modern Language Journal*, *79*(2), 179–193. https://doi.org/10.1111/j.1540-4781.1995.tb05430.x

Card, S. K., English, W. K., & Burr, B. J. (1978). Evaluation of mouse, rate-controlled isometric joystick, step keys, and text keys for text selection on a CRT. *Ergonomics*, *21*(8), 601–613. https://doi.org/10.1080/00140137808931762

Card, S. K., Moran, T. P., & Newell, A. (1980). The keystroke-level model for user performance time with interactive systems. *Communications of the ACM*, *23*(7), 396–410.

Chan, H., Verspoor, M., & Vahtrick, L. (2015). Dynamic development in speaking versus writing in identical twins. *Language Learning*, *65*(2), 298–325. https://doi.org/10.1111/lang.12107

Chomsky, N. (1959). A review of B. F. Skinner's *Verbal behavior*. *Language*, *35*(1), 26–58. https://doi.org/10.2307/411334

Council of Europe. (2020). *Common European Framework for Languages: Learning, Teaching, and Assessment*. Retrieved from https://rm.coe.int/1680459f97.

Compton, B. J., & Logan, G. D. (1991). The transition from algorithm to retrieval in memory-based theories of automaticity. *Memory & Cognition*, *19*(2), 151–158. https://doi.org/10.3758/bf03197111

Crossman, E. R. F. W. (1959). A theory of the acquisition of speed-skill. *Ergonomics*, *2*(2), 153–166. https://doi.org/10.1080/00140135908930419

De Jong, N. (2005). Can second language grammar be learned through listening?: An experimental study. *Studies in Second Language Acquisition*, *27*(2), 205–234. https://doi.org/10.1017/S0272263105050114

de Jong, N., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, *61*(2), 533–568. https://doi.org/10.1111/j.1467-9922.2010.00620.x

DeKeyser, R. M. (1994). How implicit can adult second language learning be? *AILA Review*, *11*, 83–96.

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. (1998). Beyond focus on form: Cognitive perspectives on learning and practicing second language grammar. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 42–63). New York, NY: Cambridge University Press.

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. (2007a). Conclusion: The future of practice. In R. M. DeKeyser (Ed.), *Perspectives from applied linguistics and cognitive psychology* (pp. 287−304). New York, NY: Cambridge University Press.

DeKeyser, R. M. (2007b). Introduction: Situating the concept of practice. In R. M. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 1−18). New York, NY: Cambridge University Press.

DeKeyser, R. M. (2012). Interactions between individual differences, treatments, and structures in SLA. *Language Learning*, *62*(s2), 189–200. https://doi.org/10.1111/j.1467-9922.2012.00712.x

DeKeyser, R. M. (2017). Knowledge and skill in ISLA. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 15–32). New York, NY: Routledge.

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.

DeKeyser, R. M., Salaberry, R., Robinson, P., & Harrington, M. (2002). What gets processed in processing instruction? A commentary on Bill VanPatten's "Processing instruction: An update". *Language Learning*, *52*(4), 805–823. https://doi.org/10.1111/1467-9922.00204

DeKeyser, R. M., & Sokalski, K. J. (1996). The differential role of comprehension and production practice. *Language Learning*, *46*(4), 613–642. https://doi.org/10.1111/j.1467-1770.1996.tb01354.x

Delaney, P. F., Reder, L. M., Staszewski, J. J., & Ritter, F. E. (1998). The strategy-specific nature of improvement: The power law applies by strategy within task. *Psychological Science*, *9*(1), 1–7. https://doi.org/10.1111%2F1467-9280.00001

Dienes, Z., Altmann, G., Kwan, L., & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(5), 1322–1338. https://psycnet.apa.org/doi/10.1037/0278-7393.21.5.1322

Dienes, Z. & Scott, R. (2005). Measuring unconscious knowledge: Distinguishing structural knowledge and judgment knowledge. *Psychological Research*, *69*(5-6), 338–351. https://doi.org/10.1007/s00426-004-0208-3

Eichennaum, H. (2017). Memory: Organization and control. *Annual Review of Psychology*, *68*, 19–45. https://doi.org/10.1146/annurev-psych-010416-044131

Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, *61*(2), 367–413. https://doi.org/10.1111/j.1467-9922.2010.00613.x

Ellis, N. C. (2002). Frequency effects in language processing. A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188. https://doi.org/10.1017/S0272263102002024

Ellis, R. (2015). Researching acquisition sequences: Idealization and de-idealization in SLA. *Language Learning*, *65*(1), 181–209. https://doi.org/10.1111/lang.12089

Faretta-Stutenberg, M., & Morgan-Short, K. (2018). The interplay of individual differences and context of learning in behavioral and neurocognitive second language development. *Second Language Research*, *34*(1), 67–101. https://doi.org/10.1177%2F0267658316684903

Ferman, S., Olshtain, E., Schechtman, E., & Karni, A. (2009). The acquisition of a linguistic skill by adults: Procedural and declarative memory interact in the learning of an artificial morphological rule. *Journal of Neurolinguistics*, *22*(4), 382–412. https://doi.org/10.1016/j.jneuroling.2008.12.002

Fitts, P. M. (1964). Perceptual-motor skill learning. In A. W. Melton (Ed.), *Categories of human learning* (pp. 243–285). Cambridge, MA: Academic Press.

Fitts, P. M., & Posner, M. I. (1967). *Human performance.* Belmont, CA: Brooks/Cole.

Fleishman, E. A. (1972). On the relation between abilities, learning, and human performance. *American Psychologist*, *27*(11), 1017–1032. https://doi.org/10.1037/h0033881

Forster, K. I., Guerrera, C., & Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavioral Research Methods*, *41*(1), 163–171. https://doi.org/10.3758/brm.41.1.163

Freedman, S. E., & Forster, K. I. (1985). The psychological status of overgenerated sentences. *Cognition, 19*(2), 101–131. https://doi.org/10.1016/0010-0277(85)90015-0

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Sciences*, *101*(36), 13124–13131. https://doi.org/10.1073/pnas.0404965101

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (Comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-BA117A

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

Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, *19*(10), 1–13. https://doi.org/10.3390/e19100555

Garrett, M. (1980). Levels of processing in sentence production. In B. Butterworth (Ed.), *Language production vol. 1: Speech and talk* (pp. 177–220). London, UK: Academic Press.

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

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431–436. https://doi.org/10.1111%2F1467-9280.00476

Gómez, R., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy*, *7*(2), 183–206. https://doi.org/10.1207/s15327078in0702_4

Haider, H., & Frensch, P. (2002). Why aggregated learning follows the power law of practice when individual learning does not: Comment on Rickard (1977, 1999), Delaney et al. (1998), and Palmeri (1999). *Journal of Experimental Psychology: Learning, Memory and Cognition*, *28*(2), 392–406.

Hama, M., & Leow, R. P. (2010). Learning without awareness revisited: Extending Williams (2005). *Studies in Second Language Acquisition*, *32*(3), 465–491. https://doi.org/10.1017/S0272263110000045

Hamrick, P. (2014). A role for chunk formation in statistical learning of second language syntax. *Language Learning*, *64*(2), 247–278. https://doi.org/10.1111/lang.12049

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

Healy, A., & Bourne, L. (1995). *Learning and memory of knowledge and skills: Durability and specificity*. Thousand Oaks, CA: Sage.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repeated: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*(2), 185–207.

Hui, B. (2020). Processing variability in intentional and incidental word learning: An extension of Solovyeva and DeKeyser (2018). *Studies in Second Language Acquisition*, *42*(2), 327–357. https://doi.org/10.1017/S0272263119000603

Hulstijn, J. H., Van Gelderen, A., & Schoonen, R. (2009). Automatization in second language acquisition: What does the coefficient of variation tell us? *Applied Psycholinguistics*, *30*(4), 555–582. https://doi.org/10.1017/S0142716409990014

Jiang, N. (2012). *Conducting reaction time research in second language studies.* New York, NY: Routledge.

Jurafsky, D., & Martin, J. H. (2021). *Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition* (3rd ed.). Retrieved from https://web.stanford.edu/~jurafsky/slp3/ed3book_jan122022.pdf.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. https://doi.org/10.2307/2291091

Keating, G. D., & Farley, A. P. (2008). Processing instruction, meaning-based output instruction, and meaning-based drills: Impacts on classroom L2 acquisition of Spanish object pronouns. *Hispania*, *91*(3), 639–650.

Kim, J. W., Ritter, F. E., & Koubek, R. J. (2013). An integrated theory for improved skill acquisition and retention in the three stages of learning. *Theoretical Issues in Ergonomics Science*, *14*(1), 22–37. tps://doi.org/10.1080/1464536X.2011.573008

Kolers, P. A. (1975). Memorial consequences of automatized encoding. *Journal of Experimental Psychology: Human Learning and Memory*, *1*(6), 689–701. https://doi.org/10.1037/0278-7393.1.6.689

Kruschke, J. (2015). *Doing Bayesian data analysis: A tutorial introduction with R*, *JAGS, and Stan* (2nd ed.). Cambridge, MA: Academic Press.

Lassaline, M. E., & Logan, G. D. (1993). Memory-based automaticity in the discrimination of visual numerosity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(3), 561–581. https://doi.org/10.1037/0278-7393.19.3.561

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

Lewandowki, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Li, M. (2017). *Temporal distribution of practice and individual differences in the acquisition and retention of L2 Mandarin tonal word production* (Publication No. 10641809) [Doctoral dissertation, University of Maryland–College Park]. ProQuest Dissertations and Theses Global.

Li, M., & DeKeyser, R. M. (2017). Perception practice, production practice, and music ability in L2 Mandarin tone-word learning. *Studies in Second Language Acquisition*, *39*(4), 593–620. https://doi.org/10.1017/S0272263116000358

Lim, H., & Godfroid, A. (2015). Automatization in second language sentence processing: A partial, conceptual replication of Hulstijn, Van Gelderen, and Schoonen's 2009 study. *Applied Psycholinguistics*, *36*(5), 1247–1282. https://doi.org/10.1017/S0142716414000137

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. (1992). Shapes of reaction-time distributions and shapes of learning curve: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(5), 883–914. https://doi.org/10.1037//0278-7393.18.5.883

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

Lyster, R., & Sato, M. (2013). Skill acquisition theory and the role of practice in L2 development. In M. P. García Mayo, J. Gutierrez-Mangado, & M. Martínez Adrián (Eds.), *Contemporary approaches to second language acquisition* (pp. 71–92). Amsterdam, The Netherlands: John Benjamins.

MacKay, D. G. (1982). The problems of flexibility, fluency, and speed–accuracy trade-off in skilled behavior. *Psychological Review*, *89*(5), 483–506. https://doi.org/10.1037/0033-295X.89.5.483

Maie, R. (2021). *Testing skill acquisition stages in language learning: A case of L2 vocabulary learning and practice*. Unpublished qualifying paper. Michigan State University, East Lansing, MI.

Maie, R., & DeKeyser, R. M. (2020). Conflicting evidence of explicit and implicit knowledge from objective and subjective measures. *Studies in Second Language Acquisition*, *42*(2), 359–382. https://doi.org/10.1017/S0272263119000615

Maie, R., & Godfroid, A. (2022). Controlled and automatic processing in the acceptability judgment task: An eye-tracking study. *Language Learning*, *72*(1), 158–197. https://doi.org/10.1111/lang.12474

Marteniuk, R. G. (1974). Individual differences in motor performance and learning. In J. H. Wilmore (Ed.), *Exercise and sport sciences reviews* (Vol. 2, pp. 103-130). New York, NY: Academic Press.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). Boca Raton, FL: CRC Press.

McLaughlin, B., Rossman, T., & McLeod, B. (1983). Second language learning: An information-processing perspective. *Language Learning*, *33*(2), 135–158. https://doi.org/10.1111/j.1467-1770.1983.tb00532.x

McLeod, B., & McLaughlin, B. (1986). Restructuring or automaticity? Reading in a second language. *Language Learning*, *36*(2), 109–123. https://doi.org/10.1111/j.1467-1770.1986.tb00374.x

McManus, K., & Marsden, E. (2019). Signatures of automaticity during practice: Explicit instruction about L1 processing routines can improve L2 grammatical processing. *Applied Psycholinguistics*, *40*(1), 205–234. https://doi.org/10.1017/S0142716418000553

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

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on out capacity for processing information. *Psychological Review*, *63*(2), 81–97. https://doi.org/10.1037/h0043158

Monaghan, P., Ruiz, S., & Rebuschat, P. (2021). The role of feedback and instruction on the cross-situational learning of vocabulary and morphosyntax: Mixed effects models reveal local and global effects on acquisition. *Second Language Research*, *37*(2), 261–289.

Monaghan, P., Schoetensack, C., & Rebuschat, P. (2019). A single paradigm for implicit and statistical learning. *Topics in Cognitive Science*, *11*(3), 536–554. https://doi.org/10.1111/tops.12439

Moors, A. (2016). Automaticity: Componential, causal, and mechanistic explanations. *Annual Review of Psychology*, *67*, 263–287. https://doi.org/10.1146/annurev-psych-122414-033550

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, *132*(2), 297–326. https://doi.org/10.1037/0033-2909.132.2.297

Morgan-Short, K., & Bowden, H. W. (2006). Processing instruction and meaningful output-based instruction: Effects on second language development. *Studies in Second Language Acquisition*, *28*(1), 31–65. https://doi.org/10.1017/S0272263106060025

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

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

Mueller, J. L., Hahne, A., Fujii, Y., & Friederici, A. D. (2005). Native and nonnative speakers's processing of a miniature version of Japanese as revealed by ERPs. *Journal of Cognitive Neuroscience*, *17*(8), 1229–1244. https://doi.org/10.1162/0898929055002463

Mueller, J. L., Hirotani, M., & Friederici, A. D. (2007). ERP evidence for difference strategies in the processing of case markers in native speakers and non-native speakers. *BMC Neuroscience*, *8*, 1–16. https://doi.org/10.1186/1471-2202-8-18

Murakami, A., & Ellis, N. C. (2022). Effects of availability, contingency, and formulaicity on the accuracy of English grammatical morphemes in second language writing. *Language Learning*, 1–42. https://doi.org/10.1111/lang.12500

Myung, I. J., Kim, C., & Pitt, M. A. (2000). Toward an explanation of the power law artifact: Insights from response surface analysis. *Memory & Cognition*, *28*(5), 832–840. https://doi.org/10.3758/bf03198418

Neisser, U., Novick, R., & Lazar, R. (1963). Searching for ten targets simultaneously. *Perceptual and Motor Skills*, *17*(3), 955–961. https://doi.org/10.2466/pms.1963.17.3.955

Neves, D. M., & Anderson, J. R. (1981). Knowledge compilation: Mechanisms for the automatization of cognitive skills. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*(2), 127–162. https://doi.org/10.1016/s0010-0285(03)00128-2

Ospina, R., & Ferrari, S. L. P. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, *56*(6), 1609–1623. https://doi.org/10.1016/j.csda.2011.10.005

Palmer, E. M., Horowitz, T., Torralba, A., & Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 58–71. https://doi.org/10.1037%2Fa0020747

Perruchet, P. (2021). Why is the componential construct of implicit language aptitude so difficult to capture? A commentary on the special issue. *Studies in Second Language Acquisition*, *43*(s3), 677–691. https://doi.org/10.1017/S027226312100019X

Phillips, A., Segalowitz, N., O'Brien, I., & Yamasaki, N. (2004). Semantic priming in a first and second language: Evidence from reaction time variability and event-related brain potentials. *Journal of Neurolinguistics*, *17*(2–3), 237–262.

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

Poldrack, R. A., & Packard, M. G. (2003). Competition among multiple memory systems: Converging evidence from animal and human brain studies. *Neuropsychologia*, *41*(3), 245–251. https://doi.org/10.1016/s0028-3932(02)00157-4

Posner, M. I., & Snyder, C. R. (1975). Facilitation and inhibition in the processing of signals. In P. Rabbitt & S. Dornic (Eds.), *Attention and performance* (Vol. 5, pp. 669–682). Mahwah, NJ: Lawrence Erlbaum Associates.

R Core Team (2022). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286. https://doi.org/10.1109/5.18626

Revelle, W. (2022). *psych: Procedures for psychological, psychometric, and personality research*. Northwestern University, Evanston, Illinois. R package version 2.1.9, https://CRAN.R-project.org/package=psych.

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

Rickard, T. C. (1999). A CMPL alternative account of practice effects in numerosity judgment tasks. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *25*(2), 532–542. https://doi.org/10.1037/0278-7393.25.2.532

Rickard, T. C. (2004). Strategy execution in cognitive skill learning: An item-level test of candidate models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 65–82. http://dx.doi.org/10.1037/0278-7393.30.1.65

Robinson, P. (1997). Generalizability and automaticity 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., & 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

Rodgers, D. M. (2011). The automatization of verbal morphology in instructed second language acquisition. *International Review of Applied Linguistics in Language Teaching*, *49*(4), 295–319. https://doi.org/10.1515/iral.2011.016

Rogers, V., Meara, P., Barnett-Legh, Curry, C., & Davie, E. (2017). Examining the LLAMA aptitude tests. *Journal of the European Second Language Association*, *1*(1), 49–60. http://doi.org/10.22599/jesla.24

Rogoff, B., & Lave, J. (1984). *Everyday cognition: Its development in social context*. Cambridge, MA: Harvard University Press.

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

Rumelhart, D. E., & Norman, D. A. (1978). Accretion, tuning, and restructuring: Three modes of learning. In J. W. Cotton & R. Klatzky (Eds.), *Semantic factors in cognition* (pp. 37–53). Mahwah, NJ: Lawrence Erlbaum Associates.

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language*, *47*(1), 172–196. https://doi.org/10.1006/jmla.2001.2839

Saito, K., Cui, H., Suzukida, Y., Dardon, D. E., Suzuki, Y., Jeong, H., Révész, A., Sugiura, M., & Tierney, A. (2022). Does domain-general auditory processing uniquely explain the outcomes of second language speech acquisition, even once cognitive and demographic variables are accounted for? *Bilingualism: Language and Cognition*, 1–13. https://doi.org/10.1017/S1366728922000153

Saito, K., Macmillan, K., Mai, T., Suzukida, Y., Sun, H., Magne, V., Ilkan, M., & Murakami, A. (2020). Developing, analyzing and sharing multivariate datasets: Individual differences in L2 learning revisited. *Annual Review of Applied Linguistics*, *40*, 9–25. https://doi.org/10.1017/S0267190520000045

Schneider, W., & Chein, J. M. (2003). Controlled & automatic processing: Behavior, theory, and biological mechanisms. *Cognitive Science*, *27*(3), 525–559. https://doi.org/10.1207/s15516709cog2703_8

Schneider, W., & Detweiler, M. (1988). The role of practice in dual-task performance: Toward workload modeling in a connectionist/control architecture. *Human Factors*, *30*(5), 539–566. https://doi.org/10.1177%2F001872088803000502

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1–66. https://doi.org/10.1037/0033-295X.84.1.1

Segalowitz, N. S. (2003). Automaticity and second languages. In C. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 382–408). Malden, MA: Wiley-Blackwell.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York, NY: Routledge.

Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, *26*(2), 173–199.

Segalowitz, N., Watson, V., & Segalowitz, S. (1995). Vocabulary skill: Single-case assessment of automaticity of word recognition in a timed lexical decision task. *Second Language Research*, *11*(2), 121–136. https://doi.org/10.1177%2F026765839501100204

Segalowitz, N. S., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics*, *14*(3), 369–385. https://doi.org/10.1017/S0142716400010845

Segalowitz, S. J., Segalowitz, N. S., & Wood, A. G. (1998). Assessing the development of automaticity in second language word recognition. *Applied Psycholinguistics*, *19*(1), 53–67. https://doi.org/10.1017/S0142716400010572

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127–190. https://doi.org/10.1037/0033-295X.84.2.127

Shintani, N., Li, S., & Ellis, R. (2013). Comprehension versus production-based grammar instruction: A meta-analysis of comparative studies. *Language Learning*, *63*(2), 296–329. https://doi.org/10.1111/lang.12001

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill*. Cambridge, MA: Harvard University Press.

Smith, E. E. (1968). Choice reaction time: An analysis of the major theoretical positions. *Psychological Bulletin*, *69*(2), 77–110. https://doi.org/10.1037/h0020189

Snoddy, G. S. (1926). Learning and stability: a psychophysiological analysis of a case of motor learning with clinical applications. *Journal of Applied Psychology*, *10*(1), 1–36. https://doi.org/10.1037/h0075814

Solovyeva, K., & DeKeyser, R. (2018). Response time variability signatures of novel word learning. *Studies in Second Language Acquisition*, *40*(1), 225–239. https://doi.org/10.1017/S0272263117000043

Squire, L. R., & Wixted, J. T. (2011). The cognitive neuroscience of human memory since H.M. *Annual Review of Neuroscience*, *34*, 259288. https://doi.org/10.1146/annurev-neuro-061010-113720

Stan Development Team (2022). Stan: A C++ library for programming and sampling [Computer software]. http://mc-stan.org

Suzuki, Y. (2018). The role of procedural learning ability in automatization of L2 morphology under different learning schedules: An exploratory study. *Studies in Second Language Acquisition*, *40*(4), 923−937. https://doi.org/10.1017/S0272263117000249

Suzuki, Y. (2021). Probing the construct validity of LLAMA_D as a measure of implicit learning aptitude. *Studies in Second Language Acquisition*, *43*(s3), 663–676. https://doi.org/10.1017/S0272263120000704

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

Suzuki, Y., & Sunada, M. (2018). Automatization in second language sentence processing: Relationship between elicited imitation and maze tasks. *Bilingualism: Language and Cognition*, *21*(1), 32–46. https://doi.org/10.1017/S1366728916000857

Suzuki, Y., & Sunada, M. (2020). Dynamic interplay between practice type and practice schedule in a second language: The potential and limits of skill transfer and practice schedule. *Studies in Second Language Acquisition*, *42*(1), 169–197. https://doi.org/10.1017/S0272263119000470

Suzuki, S., & Kormos, J. (2022). The multidimensionality of second language oral fluency: Interfacing cognitive fluency and utterance fluency. *Studies in Second Language Acquisition*, 1–27. https://doi.org/10.1017/S0272263121000899

Taatgen, N. A. & Anderson, J. R. (2002). Why do children learn to say "broke"? A model of learning the past tense without feedback. *Cognition*, *86*(2), 123–155. https://doi.org/10.1016/s0010-0277(02)00176-2

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

Trueman, R. C., Brooks, S. P., & Dunnett, S. B. (2012). Choice reaction time and learning. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 534–537). Boston, MA: Springer: https://doi.org/10.1007/978-1-4419-1428-6_594

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

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, *3*(1), 1–42. https://doi.org/10.1207/s15473341lld0301_1

Thorndike, E. L. (1906). *The principles of teaching based on psychology*. New York, NY: AG Seiler.

Toth, P. D. (2006). Processing instruction and a role for output in second language acquisition. *Language Learning*, *56*(2), 319–385. https://doi.org/10.1111/j.0023-8333.2006.00349.x

Woodworth, R. S., & Thorndike, E. L. (1901). The influence of improvement in one mental function upon the efficiency of other functions (I). *Psychological Review*, *8*(3), 247–261. https://doi.org/10.1037/h0074898

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. (2016). The declarative/procedural model: A neurobiological model of languagelearning, knowledge, and use. In G. Hickok & S. Small (Eds.), Neurobiology of language (pp. 953–968). Cambridge, MA: Academic Press.

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.

VanLehn, K. (1996). Cognitive skill acquisition. *Annual Review of Psychology*, *47*, 513–539. https://doi.org/10.1146/annurev.psych.47.1.513

VanPatten, B. (2020). Input processing in adult second language acquisition. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition. An introduction* (3rd ed., pp. 105–127). New York, NY: Routledge.

VanPatten, B., & Cadierno, T. (1993a). Explicit instruction and input processing. *Studies in Second Language Acquisition*, *15*(2), 225–243. https://doi.org/10.1017/S0272263100011979

VanPatten, B., & Cadierno, T. (1993b). Input processing and second language acquisition: A role for instruction. *The Modern Language Journal*, *77*(1), 45–57. https://doi.org/10.1111/j.1540-4781.1993.tb01944.x

Walker, N., Monaghan, P., Schoetensack, C., & Rebuschat, P. (2020). Distinctions in the acquisition of vocabulary and grammar: An individual differences approach. *Language Learning*, *70*(s2), 221–254. https://doi.org/10.1111/lang.12395

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

APPENDIX A: DESCRIPTIVE STATISTICS OF THE ANALYZED VARIABLES

Table S1. The accuracy rate per practice block (comprehension practice) (*n* = 65).

	1	2	3	4	5	6	7	8	9	10	11
Mean	.91	.96	.95	.94	.95	.95	.95	.97	.94	.94	.93
SD	.12	.06	.06	.07	.06	.07	.09	.05	.06	.07	.09
Min	.50	.73	.81	.62	.75	62	.38	.81	.75	.67	.59
Max	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	12	13	14	15	16	17	18	19	20	21	22
Mean	.95	.95	.95	.95	.94	.95	.96	.96	.96	.97	.97
SD	.07	.07	.06	.07	.08	.08	.06	.06	.06	.05	.04
Min	.69	.75	.75	.59	.62	.62	.80	.69	.75	.75	.81
Max	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	23	24	25	26	27	28	29	30	31	32	33
Mean	.97	.97	.96	.94	.95	.96	.94	.95	.96	.95	.96
SD	.05	.05	.05	.07	.07	.06	.09	.06	.08	.06	.08
Min	.81	.75	.62	.67	.62	.75	.56	.75	.56	.81	.50
Max	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table S2. The accuracy rate per practice block (production practice) (*n* = 60).

	1	2	3	4	5	6	7	8	9	10	11
Mean	.93	.95	.96	.95	.95	.95	.95	.95	.95	.96	.96
SD	.10	.07	.06	.08	.08	.07	.08	.06	.06	.06	.05
Min	.55	.67	.71	61	.67	.62	.59	.64	.64	.69	.67
Max	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

	12	13	14	15	16	17	18	19	20	21	22
Mean	.96	.96	.95	.95	.95	.96	.97	.96	.96	.96	.96
SD	.05	.05	.06	.05	.05	.05	.03	.04	03	.04	.04
Min	.69	.65	.58	.67	.65	.62	.83	74	.85	.73	.75
Max	1.00	1.00	1.00	1.00	1.00	1.00	.99	1.00	1.00	.99	.99

	23	24	25	26	27	28	29	30	31	32	33
Mean	.96	.95	.95	.96	.96	.96	.96	.95	.95	.95	.95
SD	.04	.05	.04	.03	.04	.05	.04	.05	.05	.05	.04
Min	.75	.66	69	.82	.80	.71	.71	.69	.65	.65	.77
Max	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table S3. RT per practice block (Comprehension practice) (*n* = 65).

	1	2	3	4	5	6	7	8	9	10	11
Mean	5.99	4.82	4.30	4.12	3.95	3.68	3.43	3.50	3.34	2.98	2.91
SD	1.34	1.24	0.94	0.87	0.85	0.76	0.69	0.72	0.69	0.49	0.56
Min	3.28	2.77	2.51	2.58	2.59	2.48	2.18	2.13	2.16	2.06	2.01
Max	8.81	7.71	6.15	6.46	6.10	5.61	5.06	5.57	4.93	4.26	4.40
	12	13	14	15	16	17	18	19	20	21	22
Mean	2.85	2.90	2.79	2.77	2.73	2.78	2.86	2.80	2.72	2.77	2.71
SD	0.53	0.60	0.62	0.51	0.52	0.52	0.61	0.71	0.60	0.65	0.60
Min	1.88	1.85	2.00	1.96	1.72	1.81	1.82	1.66	1.84	1.74	1.75
Max	4.27	4.33	4.01	3.94	4.07	4.23	4.14	4.70	4.20	4.16	4.26
	23	24	25	26	27	28	29	30	31	32	33
Mean	2.65	2.66	2.58	2.35	2.41	2.33	2.31	2.41	2.36	2.35	2.33
SD	0.56	0.62	0.53	0.48	0.47	0.43	0.48	0.49	0.50	0.43	0.48
Min	1.69	1.59	1.76	1.57	1.56	1.70	1.59	1.45	1.57	1.59	1.40
Max	4.10	4.39	4.18	3.86	3.67	3.76	3.74	3.51	3.61	3.68	3.97

Note. The numbers are in seconds.

Table S4. RT per practice block (Production practice) (*n* = 60).

	1	2	3	4	5	6	7	8	9	10	11
Mean	16.58	12.41	11.48	10.84	10.47	10.29	9.77	9.77	9.43	8.87	9.23
SD	4.89	3.69	2.81	2.72	2.66	2.81	2.67	2.43	2.41	2.51	2.79
Min	9.55	7.66	7.38	7.65	6.47	6.44	6.11	6.24	6.52	5.80	5.69
Max	30.12	23.23	18.42	18.8	16.54	16.10	16.80	15.93	15.96	14.43	16.48
	12	13	14	15	16	17	18	19	20	21	22
Mean	9.06	8.96	8.51	8.57	8.51	8.20	7.48	7.81	7.77	7.84	7.86
SD	2.58	2.29	2.13	2.69	2.41	2.09	1.91	1.87	1.81	2.09	2.01
Min	5.51	5.71	5.63	5.54	5.07	5.46	5.06	5.35	5.30	5.30	5.36
Max	14.43	13.59	13.75	15.06	15.72	14.15	13.00	13.12	12.43	13.88	13.43
	23	24	25	26	27	28	29	30	31	32	33
Mean	7.69	7.64	7.65	6.84	7.14	6.90	6.89	7.01	7.28	7.01	7.05
SD	2.19	1.93	2.12	1.65	1.83	1.69	1.53	1.65	1.92	1.99	2.26
Min	4.68	4.65	4.68	4.88	4.73	4.60	4.62	4.66	4.65	4.52	4.25
Max	14.82	11.60	14.18	11.11	11.73	11.44	10.23	11.53	12.44	12.48	14.63

Note. The numbers are in seconds.

Table S5. CV per practice block (Comprehension practice) (*n* = 65).

	1	2	3	4	5	6	7	8	9	10	11
Mean	0.37	0.36	0.34	0.36	0.35	0.36	0.34	0.36	0.35	0.33	0.32
SD	0.08	0.09	0.08	0.07	0.08	0.08	0.08	0.09	0.09	0.07	0.09
Min	0.22	0.19	0.18	0.22	0.18	0.22	0.11	0.16	0.16	0.16	0.14
Max	0.54	0.65	0.47	0.51	0.57	0.52	0.54	0.58	0.59	0.47	0.52
	12	13	14	15	16	17	18	19	20	21	22
Mean	0.32	0.33	0.33	0.33	0.33	0.34	0.35	0.35	0.35	0.35	0.35
SD	0.07	0.09	0.08	0.07	0.08	0.08	0.10	0.09	0.10	0.09	0.10
Min	0.14	0.11	0.13	0.15	0.14	0.14	0.17	0.14	0.14	0.17	0.12
Max	0.59	0.57	0.50	0.53	0.54	0.53	0.68	0.58	0.58	0.56	0.67
	23	24	25	26	27	28	29	30	31	32	33
Mean	0.35	0.35	0.34	0.32	0.33	0.32	0.32	0.34	0.33	0.34	0.33
SD	0.09	0.10	0.09	0.11	0.10	0.09	0.10	0.10	0.09	0.09	0.09
Min	0.17	0.16	0.16	0.11	0.18	0.16	0.12	0.14	0.15	0.17	0.08
Max	0.62	0.62	0.65	0.84	0.55	0.57	0.53	0.60	0.59	0.65	0.54

Note. The numbers are in seconds.

Table S6. CV per practice block (Production practice) (*n* = 60).

	1	2	3	4	5	6	7	8	9	10	11
Mean	0.22	0.2	0.21	0.23	0.21	0.23	0.22	0.23	0.22	0.18	0.22
SD	0.06	0.06	0.08	0.08	0.07	0.07	0.08	0.09	0.07	0.06	0.10
Min	0.09	0.09	0.09	0.08	0.08	0.09	0.08	0.09	0.07	0.07	0.06
Max	0.35	0.35	0.46	0.43	0.47	0.42	0.44	0.51	0.39	0.30	0.58
	12	13	14	15	16	17	18	19	20	21	22
Mean	0.22	0.23	0.21	0.22	0.21	0.21	017	0.20	0.20	0.21	0.22
SD	0.09	0.09	0.08	0.09	0.09	0.07	0.06	0.09	0.08	0.09	0.09
Min	0.09	0.07	0.09	0.09	0.07	0.09	0.08	0.07	0.07	0.06	0.09
Max	0.49	0.51	0.44	0.48	0.48	0.41	0.33	0.58	0.43	0.44	0.48
	23	24	25	26	27	28	29	30	31	32	33
Mean	0.20	0.21	0.22	0.20	0.19	0.19	0.19	0.22	0.22	0.21	0.20
SD	0.08	0.09	0.10	0.08	0.07	0.07	0.07	0.08	0.09	0.09	0.09
Min	0.08	0.06	0.09	0.07	0.07	0.07	0.08	0.08	0.08	0.02	0.09
Max	0.40	0.49	0.57	0.41	0.36	0.36	0.33	0.54	0.51	0.45	0.49

Note. The numbers are in seconds.

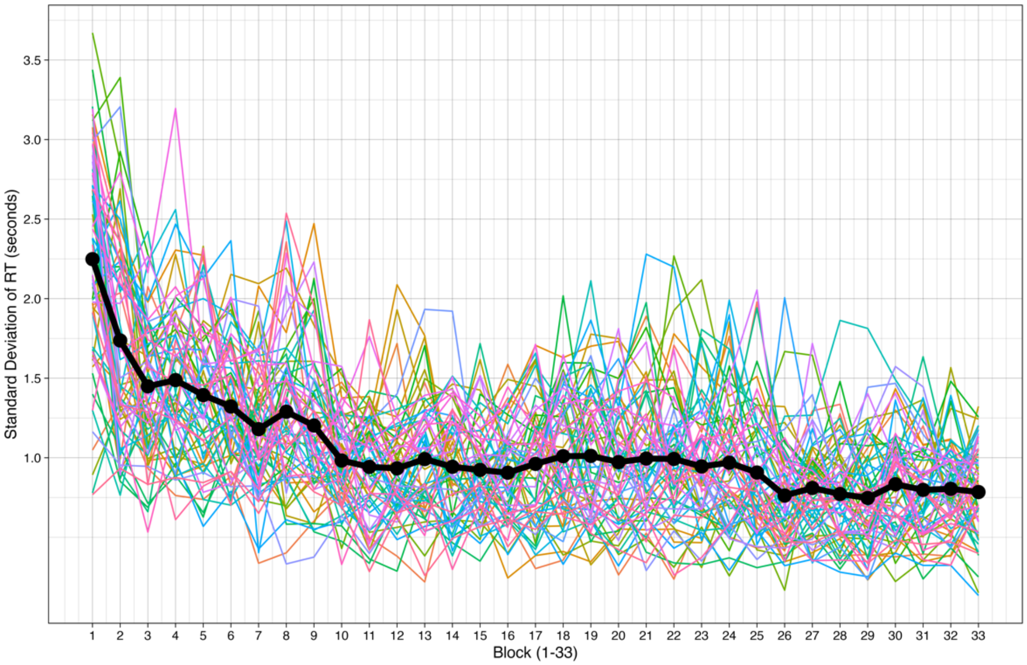


Figure S1. Aggregated SD data at the block level for comprehension practice.
Note. The black dots and line show the means over the participants and colored lines display the individual means.

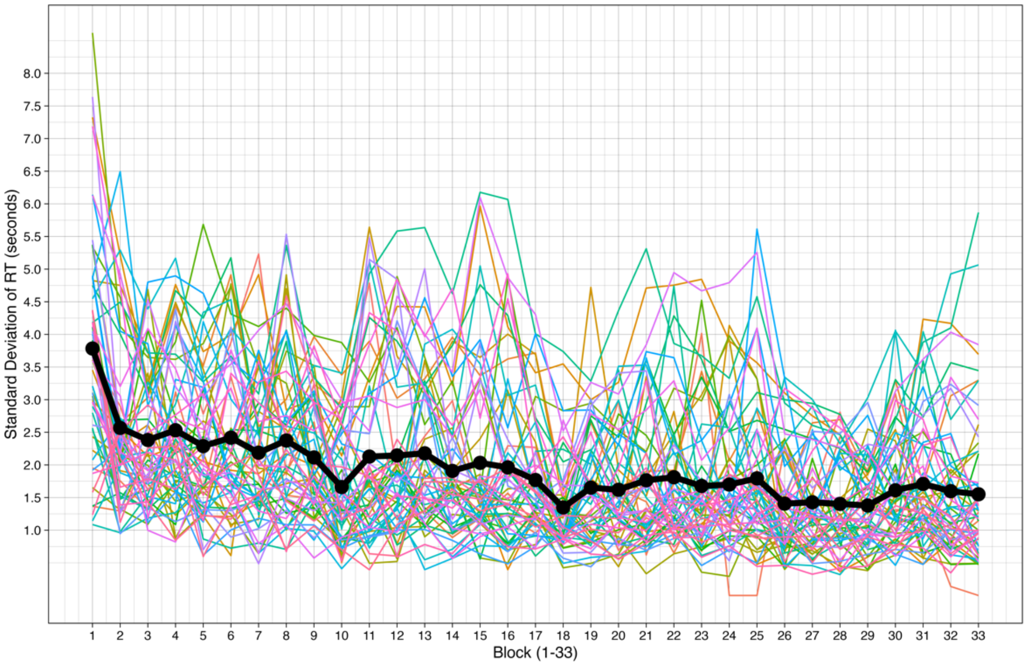


Figure S2. Aggregated SD data at the block level for production practice.
Note. The black dots and line show the means over the participants and colored lines display the individual means.

APPENDIX B: RESULTS OF THE REGRESSION ANALYSIS

Table S7. Binomial model (accuracy, comprehension).

	Fixed Effects				
	Estimate	*SE_b*	95% CrI	Prob.	$\hat{R}$
Intercept	3.124	0.144	0.144, 2.851	≈1.000	1.000
Trial	0.001	0.000	0.000, 0.001	.988	1.000
Stage 2	-0.055	0.120	-0.295, 0.173	.675	1.000
Stage 3	-0.051	0.161	-0.372, 0.262	.618	1.000
Declarative	0.438	0.137	0.168, 0.706	.999	1.000
ASRT_{15}	0.050	0.141	-0.225, 0.326	.636	1.000
SL	-0.055	0.136	-0.323, 0.214	.655	1.000
Psychomotor	-0.065	0.133	-0.324, 0.197	.690	1.000
Stage 2: Declarative	-0.156	0.112	-0.377, 0.066	.920	1.000
Stage 3: Declarative	-0.204	0.142	-0.482, 0.076	.925	1.000
Stage 2: ASRT_{15}	-0.131	0.111	-0.352, 0.085	.890	1.000
Stage 3: ASRT_{15}	-0.159	0.150	-0.450, 0.144	.857	1.001
Stage 2: SL	0.044	0.108	-0.170, 0.255	.659	1.000
Stage 3: SL	0.152	0.143	-0.125, 0.431	.858	1.000
Stage 2: Psychomotor	0.120	0.098	-0.068, 0.314	.891	1.000
Stage 3: Psychomotor	0.144	0.149	-0.145, 0.436	.835	1.000
	Random Effects				
	Estimate	SE_{SD}	95% CrI	Prob.	$\hat{R}$
$\sigma_{\alpha_{\mathrm{subject}}}$	0.782	0.099	0.606, 0.991	≈1.000	1.000
$\sigma_{\beta_{\mathrm{subject}}}$	0.002	0.000	0.001, 0.002	≈1.000	1.000
cor($\sigma_{\alpha_{\mathrm{subject}}}$, $\sigma_{\beta_{\mathrm{subject}}}$)	-.410	.160	-.676 -.060	.986	1.000

Table S8. Normal GLMM (CV, comprehension).

	Fixed Effects				
	Estimate	*SE_b*	95% CrI	Prob.	$\hat{R}$
Intercept	0.360	0.007	0.347, 0.373	≈1.000	1.000
Block	0.000	0.000	-0.001, 0.001	.519	1.000
Stage 2	-0.009	0.007	-0.023, 0.004	.909	1.000
Stage 3	-0.039	0.009	-0.057, -0.020	≈1.000	1.000
Declarative	-0.003	0.007	-0.017, 0.011	.671	1.000
ASRT_{15}	-0.000	0.007	-0.014, 0.014	.516	1.000
SL	-0.002	0.007	-0.016, 0.012	.627	1.001
Psychomotor	0.007	0.007	-0.007, 0.021	.839	1.000
Stage 2: Declarative	-0.005	0.007	-0.018, 0.009	.754	1.000
Stage 3: Declarative	-0.007	0.009	-0.023, 0.010	.777	1.000
Stage 2: ASRT_{15}	0.000	0.007	-0.014, 0.014	.515	1.000
Stage 3: ASRT_{15}	0.005	0.009	-0.012, 0.022	.715	1.000
Stage 2: SL	-0.003	0.007	-0.017, 0.010	.683	1.001
Stage 3: SL	-0.007	0.009	-0.024, 0.009	.796	1.001
Stage 2: Psychomotor	-0.001	0.007	-0.015, 0.013	.555	1.000
Stage 3: Psychomotor	-0.009	0.009	-0.027, 0.010	.827	1.001
	Random Effects				
	Estimate	SE_{SD}	95% CrI	Prob.	$\hat{R}$
$\sigma_{\alpha_{\text{subject}}}$	0.028	0.005	0.018, 0.039	≈1.000	1.001
$\sigma_{\beta_{\text{subject}}}$	0.002	0.000	0.001, 0.002	≈1.000	1.001
cor($\sigma_{\alpha_{\text{subject}}}, \sigma_{\beta_{\text{subject}}}$)	-.363	.194	-.667, .087	.943	1.002

Table S9. Normal GLMM (RT, comprehension).

	Fixed Effects				
	Estimate	SE_b	95% CrI	Prob.	$\hat{R}$
Intercept	2.123	0.036	2.052, 2.195	≈1.000	1.001
Trial (log)	-0.201	0.007	-0.215, -0.188	≈1.000	1.001
Stage 2	-0.039	0.011	-0.060, -0.018	.999	1.000
Stage 3	-0.114	0.015	-0.143, -0.085	≈1.000	1.001
Declarative	-0.072	0.019	-0.109, -0.035	≈1.000	1.002
ASRT_{15}	-0.048	0.019	-0.084, -0.011	.994	1.001
SL	0.001	0.019	-0.035, 0.038	.504	1.001
Psychomotor	0.121	0.020	0.083, 0.161	≈1.000	1.003
Stage 2: Declarative	0.021	0.011	0.000, 0.042	.974	1.000
Stage 3: Declarative	0.036	0.014	0.008, 0.064	.995	1.000
Stage 2: ASRT_{15}	0.003	0.011	-0.018, 0.024	.616	1.000
Stage 3: ASRT_{15}	0.048	0.015	0.019, 0.078	.999	1.000
Stage 2: SL	0.003	0.011	-0.018, 0.024	.617	1.000
Stage 3: SL	-0.009	0.014	-0.037, 0.019	.734	1.001
Stage 2: Psychomotor	-0.055	0.011	-0.077, -0.035	≈1.000	1.002
Stage 3: Psychomotor	-0.101	0.016	-0.133, -0.070	≈1.000	1.001
	Random Effects				
	Estimate	SE_{SD}	95% CrI	Prob.	$\hat{R}$
$\sigma_{\alpha_{\text{subject}}}$	0.281	0.028	0.233, 0.341	≈1.000	1.002
$\sigma_{\beta_{\text{subject}}}$	0.048	0.005	0.039, 0.058	≈1.000	1.002
cor($\sigma_{\alpha_{\text{subject}}}$, $\sigma_{\beta_{\text{subject}}}$)	-.913	.023	-.950, -.861	≈1.000	1.002

Table S10. Zero-one inflated beta (production, accuracy).

	Fixed Effects				
	Estimate	*SE_b*	95% CrI	Prob.	$\hat{R}$
Intercept	1.927	0.033	1.862, 1.994	≈1.000	1.003
Stage 2	0.034	0.021	-0.007, 0.075	.946	1.000
Declarative	0.156	0.033	0.089, 0.219	≈1.000	1.001
ASRT_{15}	-0.001	0.031	-0.063, 0.059	.518	1.005
SL	-0.012	0.032	-0.073, 0.049	.644	1.001
Psychomotor	-0.004	0.032	-0.068, 0.058	.547	1.002
Stage 2: Declarative	-0.119	0.019	-0.157, -0.081	≈1.000	1.001
Stage 2: ASRT_{15}	0.032	0.019	-0.005, 0.068	.955	1.002
Stage 2: SL	0.002	0.017	-0.031, 0.035	.538	1.000
Stage 2: Psychomotor	-0.012	0.018	-0.046, 0.025	.747	1.001
	Random Effects				
	Estimate	SE_{SD}	95% CrI	Prob.	$\hat{R}$
$\sigma_{\alpha_{\text{subject}}}$	0.204	0.021	0.168, 0.249	≈1.000	1.002

Table S11. Normal GLMM (CV, production).

	Fixed Effects				
	Estimate	SE_b	95% CrI	Prob.	$\hat{R}$
Intercept	0.220	0.007	0.207, 0.234	≈1.000	1.000
Block	-0.000	0.000	-0.001, 0.000	.735	1.000
Stage 2	-0.009	0.007	-0.022, 0.003	.925	1.000
Declarative	-0.005	0.006	-0.017, 0.008	.757	1.000
ASRT$_{15}$	-0.010	0.006	-0.023, 0.002	.943	1.000
SL	0.002	0.006	-0.011, 0.014	.605	1.001
Psychomotor	0.014	0.006	0.002, 0.027	.986	1.001
Stage 2: Declarative	0.002	0.006	-0.009, 0.014	.653	1.000
Stage 2: ASRT$_{15}$	0.006	0.006	-0.006, 0.019	.842	1.000
Stage 2: SL	0.001	0.006	-0.010, 0.013	.587	1.000
Stage 2: Psychomotor	-0.004	0.007	-0.018, 0.009	.743	1.000
	Random Effects				
	Estimate	SE_{SD}	95% CrI	Prob.	$\hat{R}$
$\sigma_{\alpha_{\text{subject}}}$	0.039	0.005	0.029, 0.050	≈1.000	1.000
$\sigma_{\beta_{\text{subject}}}$	0.002	0.000	0.001, 0.002	≈1.000	1.001
cor($\sigma_{\alpha_{\text{subject}}}, \sigma_{\beta_{\text{subject}}}$)	-.392	.153	-.650, -.052	.987	1.001

Table S12. Normal GLMM (RT, production).

	Fixed Effects				
	Estimate	*SE*$_b$	95% CrI	Prob.	$\hat{R}$
Intercept	3.078	0.037	3.006, 3.157	≈1.000	1.002
Trial (log)	-0.183	0.007	-0.198, -0.169	≈1.000	1.004
Stage 2	-0.052	0.007	-0.066, -0.038	≈1.000	1.000
Declarative	-0.099	0.019	-0.135, -0.062	≈1.000	1.001
ASRT$_{15}$	-0.069	0.020	-0.109, 0.-0.030	.999	1.001
SL	0.056	0.020	0.017, 0.095	.996	1.001
Psychomotor	0.154	0.020	0.114, 0.193	≈1.000	1.001
Stage 2: Declarative	0.032	0.006	0.019, 0.045	≈1.000	1.000
Stage 2: ASRT$_{15}$	0.052	0.006	0.039, 0.064	≈1.000	1.000
Stage 2: SL	-0.030	0.006	-0.042, -0.017	≈1.000	1.000
Stage 2: Psychomotor	-0.063	0.007	-0.078, -0.049	≈1.000	1.000
	Random Effects				
	Estimate	*SE*$_{SD}$	95% CrI	Prob.	$\hat{R}$
$\sigma_{\alpha_{\text{subject}}}$	0.293	0.028	0.243, 0.355	≈1.000	1.002
$\sigma_{\beta_{\text{subject}}}$	0.054	0.005	0.044, 0.064	≈1.000	1.003
cor($\sigma_{\alpha_{\text{subject}}}$, $\sigma_{\beta_{\text{subject}}}$)	-.877	.033	-.929, -.801	≈1.000	1.002