Studies in Second Language Acquisition, **43** (2021), 606–634
doi:10.1017/S0272263121000085

Research Article

THE CONTRIBUTIONS OF IMPLICIT-STATISTICAL LEARNING APTITUDE TO IMPLICIT SECOND-LANGUAGE KNOWLEDGE

Aline Godfroid*

Michigan State University

Kathy MinHye Kim

Boston University

Abstract

This study addresses the role of domain-general mechanisms in second-language learning and knowledge using an individual differences approach. We examine the predictive validity of implicit-statistical learning aptitude for implicit second-language knowledge. Participants ($n = 131$) completed a battery of four aptitude measures and nine grammar tests. Structural equation modeling revealed that only the alternating serial reaction time task (a measure of implicit-statistical learning aptitude) significantly predicted learners' performance on timed, accuracy-based language tests, but not their performance on reaction-time measures. These results inform ongoing debates about the nature of implicit knowledge in SLA: they lend support to the validity of timed, accuracy-based language tests as measures of implicit knowledge. Auditory and visual statistical learning were correlated with medium strength, while the remaining implicit-statistical learning aptitude measures were not correlated, highlighting the multicomponential nature of implicit-statistical learning aptitude and the corresponding need for a multitest approach to assess its different facets.

We gratefully acknowledge Bronson Hui, Daniel Isbell, Wenjing (Wendy) Li, Jungmin Lim, and Jeffrey Maloney for their help with materials development, piloting, data coding, and analysis. We thank Chad Bousley, Leoni Klem, Wenyue (Melody) Ma, Marisol Masso, and Sophia Zhang for assisting with data collection. Noam Siegelman provided us with a computerized version of the auditory and visual statistical learning tasks and Kara Morgan-Short provided us with a computerized version of the Tower of London task. Hope Akaeze and Steven Pierce from the Center for Statistical Consulting at Michigan State University and Michael Clark from Consulting for Statistics, Computing and Analytics Research at University of Michigan provided expert guidance on data analysis. Brittany Finch, Joanne Koh, and Wenyue (Melody) Ma provided editorial assistance. This research was supported by a Language Learning Early Career Research Grant, a research fellowship from the Center for Language Teaching Advancement, and a summer fellowship from the College of Arts and Letters at Michigan State University to Aline Godfroid.

* Correspondence concerning this article should be addressed to Aline Godfroid, Second Language Studies Program, Michigan State University, B-253 Wells Hall, 619 Red Cedar Road, East Lansing, Michigan 48824, United States. E-mail: godfroid@msu.edu

INTRODUCTION

Understanding the relationship between implicit (unconscious) learning and knowledge is fundamental to second language acquisition (SLA) theory and pedagogy. In recent years, researchers have turned to measures of *language aptitude* (an individual's ability to learn language) to better understand the nature of the different types of linguistic knowledge. Results have shown that explicit aptitude predicts the knowledge that results from explicit instruction (Li, 2015, 2016; Skehan, 2015); however, evidence for the effects of implicit-statistical learning aptitude on implicit knowledge has been limited in the field of SLA (compare Granena, 2013; Suzuki & DeKeyser, 2017). In this project, we address two questions related to implicit-statistical learning aptitude and second language (L2) knowledge: (1) whether implicit-statistical learning aptitude is a componential mechanism (convergent validity) and (2) the extent to which different types of implicit-statistical learning tasks predict implicit knowledge (predictive validity). We expand the number of implicit-statistical learning aptitude measures beyond serial reaction time to obtain a more comprehensive assessment of learners' implicit-statistical aptitude. Alongside, we will administer a battery of linguistic knowledge tests designed to measure explicit and implicit L2 knowledge. By doing so, we are able to examine how implicit-statistical learning aptitude predicts the development of implicit L2 knowledge.

IMPLICIT-STATISTICAL LEARNING APTITUDE

Implicit-statistical learning denotes one's ability to pick up regularities in the environment (Frost et al., 2019).[1] Learners with greater implicit-statistical learning aptitude, for instance, can segment word boundaries (statistical learning) and detect regularities in artificial languages (implicit language learning) better than those with lower implicit-statistical learning ability (for a comprehensive review of the unified framework of implicit-statistical learning, see Christiansen, 2019; Conway & Christiansen, 2006; Perruchet & Pacton, 2006; Rebuschat & Monaghan, 2019). This process of implicit-statistical learning is presumed to take place incidentally, without instructions to learn or the conscious intention on the part of the learner to do so.

Traditionally, implicit-statistical learning ability has been conceptualized as a unified construct where learning from different modes, such as vision, audition, and sense of touch, is interrelated and there is a common implicit-statistical learning mechanism governing the extraction of patterns across different modes of input. Recently, however, a growing body of research has evidenced that implicit-statistical learning may operate differently in different modalities and stimuli, yet still be subserved by domain-general computational principles (for reviews, see Arciuli, 2017; Frost et al., 2015; Siegelman et al., 2017a). In this view, implicit-statistical learning is modality and stimulus constrained (as the encoding of the information in different modalities relies on different parts of our body and different cortices) but this modality specific information is subject to domain-general processing principles that invoke shared brain regions. Implicit-statistical learning is thus modality specific at the level of encoding while also obeying domain-general computational principles at a more abstract level. If implicit-statistical

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

learning is a componential ability, it follows that a more comprehensive approach to measurement is needed that brings together different tasks tapping into different components of implicit-statistical learning. Our first aim, accordingly, is to test the convergent validity of implicit-statistical learning measures by assessing the interrelationships between different measures of implicit-statistical learning. Doing so will inform measurement and help illuminate the theoretical construct of implicit-statistical learning.

In SLA, researchers have relied on different measures to capture implicit learning, statistical learning, and the related construct of procedural memory (see Appendix S1 in online Supplementary Materials).[2] For instance, implicit learning aptitude has been measured with the LLAMA D test of phonemic coding ability (Granena, 2013, 2019; Yi, 2018), the serial reaction time task (Granena, 2013, 2019; Hamrick, 2015; Linck et al., 2013; Suzuki & DeKeyser, 2015, 2017; Tagarelli et al., 2016; Yi, 2018), and the alternating serial reaction time (ASRT) task (Faretta-Stutenberg & Morgan-Short, 2018; Tagarelli et al., 2016). The ASRT task doubles as a measure of procedural memory (Buffington & Morgan-Short, 2018; Buffington et al., 2021; Faretta-Stutenberg & Morgan-Short, 2018; Hamrick, 2015). Other measures of procedural memory are the Tower of London (TOL) (Antoniou et al., 2016; Ettlinger et al., 2014; Morgan-Short et al., 2014; Pili-Moss et al., 2019; Suzuki, 2017) and the Weather Prediction Task (Faretta-Stutenberg & Morgan-Short, 2018; Morgan-Short et al., 2014; Pili-Moss et al., 2019). Lastly, statistical learning has only been measured in the auditory modality in L2 research to date, with different tests of verbal auditory statistical learning (Brooks & Kempe, 2013; McDonough & Trofimovich, 2016; Misyak & Christiansen, 2012).

These different measures can provide insight into the nature of the learning processes that individuals draw on in different language learning tasks. Specifically, when performance on the linguistic task and the aptitude measure share variance, a common cognitive process (i.e., implicit-statistical learning or procedural memory) can be assumed to guide performance on both tasks. To illustrate, Yi (2018) found that native English speakers' performance on a serial reaction time task predicted (i.e., shared variance with) their phrasal acceptability judgment speed. A similar association for L2 speakers between their explicit aptitude and phrasal acceptability judgment accuracy led the author to conclude that L1 speakers process collocations implicitly and L2 speakers process them more explicitly.

Although the use of implicit-statistical learning aptitude measures in L2 research is rising, there is a need to justify the use of these measures from a theoretical and a psychometric perspective more strongly. The possibility that implicit-statistical learning may not be a unitary construct highlights the need to motivate the choice of specific aptitude measure(s) and examine their construct validity, with due consideration of the measures' input modality (Frost et al., 2015). The questions of convergent validity (correlation with related measures) and divergent validity (dissociation from unrelated measures) have implications for measurement as well as SLA theory. Indeed, if implicit-statistical learning aptitude is to fulfill its promise as a cognitive variable that can explain the learning mechanisms that operate in different L2/foreign language contexts, for different target structures, and for learners of different L2 proficiency levels, valid and reliable measurement will be paramount.

In recent years, some researchers have begun to examine the construct validity of implicit-statistical learning aptitude measures by exploring their relationship to implicit memory (Granena, 2019), procedural memory (Buffington et al., 2021; Buffington & Morgan-Short, 2018), and working memory and explicit learning aptitude (Yi, 2018). For measures of implicit learning aptitude, Granena (2019) found that the serial reaction time task loaded onto a different factor than the LLAMA D in an exploratory factor analysis (EFA), suggesting the two measures did not converge. Similarly, Yi (2018) reported that the serial reaction time task and LLAMA D were uncorrelated and the reliability of LLAMA D was low. In a study combining measures of implicit learning aptitude and procedural memory, Buffington et al. (2021) also observed a lack of convergent validity between the ASRT, the Weather Prediction Task, and the TOL. These results do not support a unitary view of implicit-statistical learning aptitude or procedural memory. Furthermore, this research is yet to include measures of statistical learning as another approach to the same phenomenon (Christiansen, 2019; Conway & Christiansen, 2006; Perruchet & Pacton, 2006; Reber, 2015; Rebuschat & Monaghan, 2019). More research is needed to advance our understanding of these important issues.

With this study, we aim to advance this research agenda. We consider multiple dimensions of implicit-statistical learning aptitude, their reliabilities, and interrelationships (convergent validity). Of the various measures used as implicit-statistical learning aptitude in SLA and cognitive psychology, we included measures that represent different modes of input streams: visual statistical learning (VSL) for visual input, auditory statistical learning (ASL) for aural input, and ASRT for motor and visual input. In addition, we included the TOL task in recognition of its wide use in SLA research as a measure of procedural memory along with the ASRT task.

IMPLICIT, AUTOMATIZED EXPLICIT, AND EXPLICIT KNOWLEDGE

It is widely believed that language users possess at least two types of linguistic knowledge: explicit and implicit. Explicit knowledge is conscious and verbalizable knowledge of forms and regularities in the language that can be acquired through instruction. Implicit knowledge is tacit and unconscious linguistic knowledge that is gained mainly through exposure to rich input, and therefore cannot be easily taught. A third type of knowledge, automatized explicit knowledge, denotes explicit knowledge that language users are able to use rapidly, in time-pressured contexts, as a result of their extensive practice with the language. While the use of (nonautomatized) explicit knowledge tends to be slow and effortful, both implicit and automatized explicit knowledge can be deployed rapidly, with little or no conscious effort, during spontaneous communication (DeKeyser, 2003; Ellis, 2005). Consequently, it has been argued that implicit and automatized explicit knowledge are "functionally equivalent" (DeKeyser, 2003), in that it may be impossible to discern between the two in practice.

In a landmark study, Ellis (2005) proposed a set of criteria to guide the design of tests that could provide relatively separate measures of explicit and implicit knowledge. Using principal component analysis, Ellis showed that time-pressured grammar tests that invite a focus on meaning (content creation) or form (linguistic accuracy) loaded onto one component (i.e., an oral production [OP] task, elicited imitation [EI], and a timed grammaticality judgment test [GJT]), which Ellis termed implicit knowledge. Untimed

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

grammar tests that focus learners' attention on form (i.e., ungrammatical items on an untimed GJT and a metalinguistic knowledge test [MKT]) loaded onto a different component, which Ellis labeled explicit knowledge (see Ellis & Loewen, 2007, for a replication of these findings with confirmatory factor analysis). Subsequent studies using factor analysis on similar batteries of language tests also uncovered at least two dimensions of linguistic knowledge, termed explicit and implicit, which was largely consistent with Ellis's initial results (e.g., Bowles, 2011; Kim & Nam, 2017; Spada et al., 2015; Zhang, 2015; but see Gutiérrez, 2013).

The advent of reaction-time measures, however, invited new scrutiny of the construct validity of traditional measures of implicit knowledge such as the EI task and the timed written GJT (compare Ellis, 2005; Suzuki & DeKeyser, 2015, 2017; Vafaee et al., 2017). The theoretical debate surrounding this issue was the distinction between implicit and automatized explicit knowledge, described previously, and whether, aside from differences in neural representation, the two types of knowledge can be differentiated behaviorally, in L2 learners' language use. Departing from Ellis (2005), researchers have hypothesized that timed, accuracy-based tests (e.g., EI) may be more suited to tap into learners' automatized explicit knowledge because timed tests do not preclude learners from accessing their explicit knowledge, but merely make it more difficult for learners to do so (DeKeyser, 2003; Suzuki & DeKeyser, 2015). Reaction-time tests such as self-paced reading (SPR), however, require participants to process language in real time, as it unfolds, and could therefore hypothetically be more appropriate to capture learners' implicit knowledge (Godfroid, 2020; Suzuki & DeKeyser, 2015; Vafaee et al., 2017). In the implicit-statistical learning literature, Christiansen (2019) similarly argued for the use of processing-based measures (e.g., reaction time tasks) over reflection-based tests (e.g., judgment tasks) to measure the effects of implicit-statistical learning. He did not, however, attribute differences in construct validity to them (i.e., both types of tests are assumed to measure largely implicit knowledge, but at different levels of sensitivity or completeness).

Using confirmatory factor analysis, Suzuki (2017) and Vafaee et al. (2017) confirmed that timed, accuracy-based tests and reaction-time tests represent different latent variables, which they interpreted as automatized explicit knowledge and implicit knowledge, respectively. The researchers did not include measures of (nonautomatized) explicit knowledge, however, which leaves the results open to alternative explanations. Specifically, for automatized explicit knowledge to be a practically meaningful construct, it needs to be distinguishable from implicit knowledge and (nonautomatized) explicit knowledge simultaneously, within the same statistical analysis. Doing so requires a more comprehensive approach to measurement, with tests of linguistic knowledge being sampled from across the whole explicit/automatized explicit/implicit knowledge spectrum. Hence, current evidence for the construct validity of reaction-time tasks as measures of implicit knowledge is still preliminary.

More generally, all the previous validation studies have included only a subset of commonly used explicit/implicit knowledge tests in SLA, which limits the generalizability of findings. Differences in test batteries may explain the conflicting findings for tests such as the timed written GJT (see Godfroid et al., 2015). This is because the results of confirmatory factor analysis are based on variance-covariance patterns for the tests included in the analysis and hence different test combinations may give rise to different

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

statistical solutions. To obtain a more comprehensive picture, Godfroid et al. (2018) synthesized 12 years of test validation research since Ellis (2005) by including all previously used measures in one study—the word monitoring test (WMT), SPR, EI, OP, timed/untimed GJTs in the aural and written modes, and the MKT. The results suggested that both a three-factor model (EI and timed written GJT as "automatized explicit knowledge"; Suzuki & DeKeyser, 2015, 2017) and a two-factor model (EI and timed written GJT as "implicit knowledge"; Ellis, 2005) provided a good fit for the data and that the two models did not differ significantly. These results support the viability of a three-way distinction between explicit, automatized explicit, and implicit knowledge. As with all factor analytic research, however, the nature of the latent constructs was left to the researchers' interpretation. Other sources of validity evidence, such as different patterns of aptitude-knowledge associations, examined here, could support the proposed interpretation and bolster the case for the distinction between implicit and automatized explicit knowledge.

CONTRIBUTIONS OF IMPLICIT-STATISTICAL LEARNING APTITUDE TO IMPLICIT KNOWLEDGE

Three studies to date have examined aptitude-knowledge associations in advanced L2 speakers, with a focus on measurement validity. We will review each study in detail because of their relevance to the current research. Granena (2013) compared Spanish L1 and Chinese-Spanish bilinguals' performance on measures of explicit and implicit knowledge, using both agreement and nonagreement structures in Spanish. The participants had acquired Spanish either from birth, early in life, or postpuberty. Granena wanted to know whether the participants' starting age impacted the cognitive processes they drew on for language learning. She found that early and late bilinguals' performance on agreement structures correlated with their implicit-statistical learning aptitude, as measured by a serial reaction time task (early learners) or LLAMA D (late learners). These results suggested that bilinguals who do not acquire the language from birth may still draw on implicit-statistical learning mechanisms, albeit to a lesser extent than native speakers do; hence, the bilinguals' greater sensitivity to individual differences in implicit-statistical learning aptitude compared to native speakers.

Suzuki and DeKeyser (2015) compared the construct validity of EI and the WMT as measures of implicit knowledge. L1 Chinese-L2 Japanese participants performed an EI test with a built-in monitoring task. They were asked to listen to and repeat sentences, as is commonly done in an EI test, but in addition, they were asked to monitor the spoken sentences for a given target word (i.e., built-in word monitoring). The researchers found that performance on the two test components correlated with different criterion variables; specifically, EI correlated with performance on a MKT (a measure of explicit knowledge), whereas the WMT correlated with performance on the serial reaction time task (a measure of implicit-statistical learning aptitude), albeit only in a subgroup of participants who had lived in Japan for at least 2.5 years. Based on these results, the authors concluded that the WMT is a measure of implicit linguistic knowledge, whereas the EI test (traditionally considered a measure of implicit knowledge as well) is best considered a measure of automatized explicit knowledge.

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

In a follow-up study, Suzuki and DeKeyser (2017) examined the relationships among implicit knowledge, automatized explicit knowledge, implicit-statistical learning aptitude, explicit learning aptitude, and short-term memory. Different from Granena (2013) and Suzuki and DeKeyser (2015), the researchers found no significant association between serial reaction time (a measure of implicit-statistical learning aptitude) and either implicit or automatized explicit knowledge. Rather, they found that advanced Japanese L2 students' performance on LLAMA F (a measure of explicit learning aptitude) predicted their automatized explicit knowledge. The authors also tested the explanatory value of adding a *knowledge interface* (i.e., a directional path) between automatized explicit and implicit knowledge in the structural equation model (SEM). This path was indeed significant, meaning automatized explicit knowledge predicted implicit knowledge, but the interface model as a whole was not significantly different from a noninterface model that did not include such a path. The researchers interpreted their results as evidence that automatized explicit knowledge *directly* impacts the acquisition of implicit knowledge (through the interface), and that explicit learning aptitude *indirectly* facilitated the development of implicit knowledge. Thus, in their study no direct predictors of implicit knowledge were found.

Taken together, Granena (2013) and Suzuki and DeKeyser (2015) found a positive correlation between implicit knowledge test scores (i.e., sensitivity on a WMT) and implicit-statistical learning aptitude, in line with the view that the WMT, a reaction-time measure, may index implicit knowledge. In Suzuki and DeKeyser's (2017) SEM, however, the same implicit-statistical learning aptitude test had no association with the implicit knowledge construct, which was composed of three reaction-time measures, including a WMT (incidentally, none of the three reaction-time measures loaded onto the implicit knowledge factor significantly, which may have signaled a problem with these measures or with the assumption that they were measuring implicit knowledge). Critically, the three studies have only used a very limited set of implicit-statistical learning aptitude measures (serial reaction time and, in Granena's study, LLAMA D) that examine implicit-statistical motor learning and phonetic coding ability, respectively. Given that the implicit-statistical learning construct is modality specific (i.e., implicit-statistical learning can occur in visual, aural, and motor modes), the limited range of implicit-statistical learning aptitude tests in these studies limits the generalizability of the results to the tests with which they were obtained. Another issue concerns the low reliability of aptitude and knowledge measures obtained from reaction time data (Draheim et al., 2019; Rouder & Haaf, 2019), which may obscure any aptitude–knowledge relationships. In recognition of these gaps, we included a battery of four implicit-statistical learning aptitude tests (VSL, ASL, ASRT, and TOL) in order to examine the predictive validity of implicit-statistical learning aptitude for implicit, automatized explicit, and explicit L2 knowledge.

RESEARCH QUESTIONS

In this study, we triangulate performance on a battery of nine linguistic knowledge tests with data from four measures of implicit-statistical learning aptitude with an aim to validate a new and extended set of measures of implicit, automatized explicit, and explicit knowledge. The following research questions guided the study:

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

1. Convergent validity of implicit-statistical learning aptitude:
   To what extent do different measures of implicit-statistical learning aptitude interrelate?
2. Predictive validity of implicit-statistical learning aptitude:
   To what extent do measures of implicit-statistical learning aptitude predict three distinct dimensions of linguistic knowledge, referred to as explicit knowledge, automatized explicit knowledge, and implicit knowledge?

METHOD

PARTICIPANTS

Participants were 131 nonnative English speakers (Female = 69, Male = 51, Not reported = 11) who were pursuing academic degrees at a large Midwestern university in the United States. The final sample was obtained after excluding 26 participants who completed only one out of the four aptitude tests. Nearly half of the participants were native speakers of Chinese (n = 66). The remaining participants' L1s included Korean, Spanish, Arabic, Russian, Urdu, Malay, Turkish, and French, among others. The participants' average length of residence in an English-speaking country was 41 months (SD = 27.21, range 2–200 months). The participants were highly proficient English speakers with an average TOEFL score of 96.00 (SD = 8.80). Their mean age was 24 years (SD = 4.64) and their average age of arrival in the United States was 20 years old (SD = 5.68). They received $50 as compensation for their time.

TARGET STRUCTURES

The target structures included six grammatical features: (1) third-person singular *-s*, (2) mass/count nouns, (3) comparatives, (4) embedded questions, (5) *be* passive, and (6) verb complement. We selected these three syntactic (4–6) and three morphological (1–3) structures to measure a range of English grammar knowledge (see Table 1 for examples). These structures emerge in different stages of L2 acquisition (e.g., Ellis, 2009) and thus were deemed appropriate to represent English morphosyntax.

TABLE 1. Six grammatical structures and examples of ungrammatical sentences for each.

Morphological	Syntactical
Third person *-s*	Embedded questions
* *The old woman* ***enjoy*** *reading many different famous novels.*	* *He wanted to know why* ***had he*** *studied for the exam.*
Mass/count nouns	*Be* passive
* *The boy had* ***rices*** *in his dinner bowl.*	* *The flowers were* ***pick*** *last winter for the festival.*
Comparatives (double marking)	Verb Complement (*ask*, *have*, *need*, *want*)
* *It is* ***more harder*** *to learn Japanese than to learn Spanish.*	* *Jim is told his parents* ***want buying*** *a new house.*

Note: Critical region in boldface and underlined.

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

TABLE 2. Summary of measures.

Test constructs: two knowledge factors	Test constructs: three knowledge factors	Test	No. of items (total)	No. of items (critical)	Grammaticality	Dependent variable
Implicit knowledge	Implicit knowledge	Word monitoring	126	96	48 grammatical, 48 ungrammatical	Grammatical sensitivity (ms): $RT_{ungrammatical} - RT_{grammatical}$
		Self-paced reading	126	96	48 grammatical, 48 ungrammatical	Grammatical sensitivity (ms): $RT_{ungrammatical} - RT_{grammatical}$
	Automatized explicit knowledge	Timed aural GJT	40	24	12 grammatical, 12 ungrammatical	Accuracy: proportion
		Timed written GJT	40	24	12 grammatical, 12 ungrammatical	Accuracy: proportion
		Elicited imitation	32	24	12 grammatical, 12 ungrammatical	Accuracy: correct usage in obligatory contexts
		Oral production	250-word story	250-word story	Grammatical sentences only	Accuracy: correct usage in obligatory contexts
Explicit knowledge	Explicit knowledge	Untimed written GJT	40	24	12 grammatical, 12 ungrammatical	Accuracy: proportion
		Untimed aural GJT	40	24	12 grammatical, 12 ungrammatical	Accuracy: proportion
		Metalinguistic knowledge test	12	12	12 ungrammatical	Accuracy: error correction and error explanation
Implicit-statistical learning aptitude		Alternating serial reaction time	10 blocks of 40 trials	10 blocks of 40 trials	20 pattern, 20 random	RT difference (correct responses only): (RT of block 1 − RT of block 10 pattern trials) − (RT of block 1 − RT of block 10 random trials)
		Auditory statistical learning	42	42	NA	Accuracy: proportion
		Visual statistical learning	42	42	NA	Accuracy: proportion
		Tower of London	28	28	four 3-move, eight 4-move, eight 5-move, eight 6-move trials	Overall solution time: (RT on the first trial − RT on the final trial)/ RT on the first trial

INSTRUMENTS

We administered nine linguistic tests of L2 grammar knowledge: WMT, SPR, OP, Timed Aural Grammaticality Judgment Test (TAGJT), Timed Written Grammaticality Judgment Test (TWGJT), EI, Untimed Aural Grammaticality Judgment Test (UAGJT), Untimed Written Grammaticality Judgment Test (UWGJT), and MKT. Based on previous literature, it was hypothesized that these tests represented either the (1) *implicit*, *automatized explicit*, and *explicit* knowledge constructs (i.e., an extension of the Suzuki and DeKeyser [2017] model) or the (2) *implicit* and *explicit* knowledge constructs (i.e., the Ellis [2005] model). We also administered four implicit-statistical learning aptitude tests: VSL, ASL, ASRT, and TOL. Table 2 summarizes the characteristics of the nine linguistic and four aptitude tests.

LINGUISTIC TESTS

Word Monitoring Task

The WMT is a dual processing task that combines listening comprehension and word-monitoring task demands. Participants first saw a content word (e.g., *reading*), designated as the target for word monitoring. They were instructed to press a button immediately as they heard the word in a spoken sentence (e.g., *The old woman* ***enjoys*** *reading many different famous novels).* Importantly, the monitor word was always preceded by one of the six linguistic structures in either a grammatical (e.g., *enjoys*) or ungrammatical (e.g., *enjoy*) form. Exhibiting grammatical sensitivity—that is, slower reaction times on content words when the prior word is ungrammatical than when it is grammatical—indicated knowledge of the grammatical target structure.

Self-Paced Reading

In the SPR task, participants read a sentence word-by-word in a self-paced fashion. They progressed to the next word in a sentence by pressing a button. As with the WMT, participants read grammatical and ungrammatical sentences. Evidence for linguistic knowledge was based on grammatical sensitivity—that is, slower reaction times to the ungrammatical version than the matched, grammatical version of the same sentence. In particular, we analyzed reaction times for the spillover region (i.e., the word or words immediately following the critical region) for each sentence and created a difference score for the ungrammatical and grammatical sentences.

Oral Production

The OP task was a speaking test where participants had to retell a picture-cued short story that contained multiple tokens of the six target structures. After reading the story two times without a time limit, participants had to retell the story in as much detail as possible in two and a half minutes. The percentage of correct usage of each target structure in all obligatory occasions of use (i.e., obligatory contexts) was used as a dependent variable. Obligatory contexts were defined relative to the participants' own production. Two coders independently coded OP. The reliability of interrater coding (Pearson *r*) was .96.

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

Elicited Imitation

Similar to the OP task, the EI was a speaking test where participants were asked to listen to a sentence, judge the semantic plausibility of the sentence, and repeat the sentence in correct English. No explicit instructions directed participants to correct the erroneous part of the sentence. Following Erlam's (2006) scoring system, correct usage in obligatory context was used for analysis.

Grammaticality Judgment Tests

In the GJTs, participants either read or listened to a sentence in a timed or an untimed test condition. The participants were instructed to determine the grammaticality of the sentence. The time limit for each sentence in the timed written and the timed aural GJT was set based on the length of the audio stimuli in the aural GJT. We computed the median audio length of sentences with the same number of words and added 50%. This resulted in a time limit of 4.12 seconds for a seven-word sentence and up to 5.7 seconds for a 14-word sentence for the timed GJTs. Two sets of sentences were created and counterbalanced for grammaticality and each set was rotated between the four tests, resulting in eight sets of sentences in total. In each of the four GJTs (timed written, untimed written, timed aural, untimed aural), one point was given per accurate judgment.

Metalinguistic Knowledge Test

The MKT required participants to read 12 sentences that contained a grammatical error. Their task was to (1) identify the error, (2) correct the error, and (3) explain in as much detail as possible why it was ungrammatical. We only scored the error correction and explanation parts of the test; as such, a total of two points were given per question. The maximum score was 24 and the total score was converted to a percentage. See Appendix S2 in online Supplementary Materials for the scoring rubric.

IMPLICIT-STATISTICAL LEARNING APTITUDE TESTS

ASRT

The ASRT (Howard & Howard, 1997) was used to measure implicit-statistical learning aptitude. In the ASRT, participants viewed four empty circles in the middle of a computer screen that would fill as a black circle one at a time. The sequence of the filled circles followed a pattern that alternated with random (nonpatterned) trials, creating a second-order relationship (e.g., 2r4r3r1r, where *r* denotes a random position). Participants were instructed to press the corresponding key on a keyboard that mirrored the position of the filled circle as quickly and accurately as possible. To capture learning, we calculated the change in reaction time to pattern trials from block 1 to block 10 and subtracted the change in reaction time to random trials from block 1 to block 10. Positive values indicate a greater improvement in sequence learning over the course of the task.

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

Auditory Statistical Learning

The ASL (Siegelman et al., 2018, experiment 1b) served as another implicit-statistical learning task. In the ASL test, participants heard 16 nonverbal, familiar sounds that were randomly organized into eight triplets (sequences of three sounds). Four triplets had a transitional probability of one (i.e., they were fixed) and four triplets had a transitional probability of .33 (i.e., every sound was followed by one of three other sounds, with equal likelihood). Each triplet was repeated 24 times during a continuous familiarization stream. Participants were asked to listen to the input very carefully as they would be tested on it after the training. The test consisted of 42 trials: 34 four-alternative forced-choice questions measuring recognition of triplets and eight pattern completion trials measuring recall. Performance on the test yielded an accuracy percentage score.

Visual Statistical Learning

The VSL (Siegelman et al., 2017b) was used to measure learners' ability to learn visual patterns implicitly. As the visual counterpart of the ASL, the VSL presented participants with 16 complex visual shapes that were difficult to describe verbally and were randomly organized into eight triplets (sequences of three shapes). The triplets had a transitional probability of one. Each triplet was repeated 24 times during the familiarization phase. In the testing phase, participants completed 42 trials: 34 four-alternative forced-choice items measuring recognition of triplets and eight pattern completion trials measuring recall. Performance on the test yielded an accuracy percentage score.

Tower of London

The TOL (Kaller et al., 2011) was administered to measure learners' implicit-statistical learning ability during nonroutine planning tasks. Participants were presented with two spatial configurations that consisted of three pegs with colored balls on them. These configurations were labeled as "Start" or "Goal." The participants' task was to move the colored balls on the pegs in the "Start" configuration to match the "Goal" configuration in the given number of moves. There was a block of four 3-move trials, eight 4-move trials, eight 5-move trials, and eight 6-move trials (Morgan-Short et al., 2014). We will present the results for overall solution time in what follows, which is the sum of initial thinking time and movement execution time. All three measures yielded similar results. To capture learning, we calculated a proportional change score for each block of trials (i.e., 3-move, 4-move, 5-move, and 6-move separately) for each participant using the following computation: (RT on the first trial − RT on the final trial)/RT on the first trial. Positive values indicate a greater improvement in planning ability from the beginning to the end of each block in the experiment.

PROCEDURE

Participants met with a trained research assistant for three separate sessions. As seen in Table 3, the first session included the WMT, SPR, timed aural GJT, and untimed aural GJT; the second session started with OP followed by EI, the timed written GJT, untimed

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

TABLE 3. The sequencing of linguistic and aptitude tests.

Session 1	Session 2	Session 3
Word monitoring	Oral production	Visual statistical learning
Self-paced reading	Elicited imitation	Auditory statistical learning
Timed aural GJT	Timed written GJT	Alternating serial reaction time
Untimed aural GJT	Untimed written GJT	Tower of London
	Metalinguistic knowledge test	

written GJT, and MKT; in the last session, participants completed all aptitude tests starting with VSL, and ended with the MLAT 5 (which is not discussed in this article). Sessions 1 and 2 started with the more implicit knowledge measures to minimize the possibility of participants becoming aware of the target features in the implicit tasks.

DATA ANALYSIS

Descriptive Statistics and Correlations

Overall, 6% of the data were missing and they were missing completely at random (Little's MCAR test: $\chi 2 = 1642.159$, $df = 1744$, $p = .960$). To explore the associations among measures of implicit-statistical learning aptitude and between implicit-statistical learning aptitude and linguistic knowledge, respectively, we calculated descriptive statistics and Spearman correlations (abbreviated as *rs* in what follows) for all measures of L2 morphosyntactic knowledge and cognitive aptitude. All such analyses were carried out in *R* version 1.2.1335 (R Core Team, 2018).

Factor Analysis

To address research question 1, "to what extent do different measures of implicit-statistical learning aptitude interrelate (convergent validity)?," we conducted an EFA to explore the association between the four implicit-statistical learning aptitude measures. The EFA was performed with an oblique rotation (oblimin) that permits factors to correlate with each other. The model was computed using weighted least squares to account for the violation of multivariate normality assumption for the four tests (Mardia's skewness coefficient was 36.93 with a *p*-value of 0.012; Mardia's kurtosis coefficient was 2.28 with a *p*-value of 0.023). Finally, we used a factor loading cutoff criterion of .40 to interpret the factor loadings.

To address research question 2, "to what extent do measures of implicit-statistical learning aptitude predict three distinct dimensions of linguistic knowledge (i.e., explicit knowledge, automatized explicit knowledge, and implicit knowledge (predictive validity)?," we built confirmatory factor analysis (CFA) and SEM models using the *lavaan* package in R. To examine the psychometric dimensions underlying the nine linguistic tests, we constructed two CFA models, a two-factor and a three-factor model. These models were specified based on theory and previous empirical findings from CFA studies by Ellis (2005) and Suzuki and DeKeyser (2017). To evaluate the CFA models, we

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

used a model test statistic (chi-square test), standardized residuals (<|1.96|) and three model fit indices (Hu & Bentler, 1999): the comparative fit index (CFI => .96), the root mean square error of approximation (root mean square error of association [RMSEA] =< .06), and the standardized root mean square residual (standardized root mean square [SRMR] =< .09). We then built a SEM. In combination with a measurement model (CFA), SEM estimates the directional effects of independent variables (measures of implicit-statistical learning aptitude) on the latent dependent variables (the knowledge type constructs). Full-information maximum likelihood estimation was used to evaluate different models and Robust Maximum Likelihood was adopted as an estimation method for both the CFA and SEM analyses to account for the violation of multivariate normality assumption.

RESULTS

DESCRIPTIVE STATISTICS

Table 4 shows the descriptive statistics for all linguistic and aptitude measures. Participants showed a wide range of abilities in their performance on the linguistic knowledge measures. Reliabilities of the individual differences measures ranged from satisfactory to high and were generally on a par with those reported in previous studies: ASRT intraclass correlation coefficient (ICC) = .96 (this study) and ASRT ICC = .99 (Buffington & Morgan-Short, 2018), VSL α = .75 (this study) and VSL α = .88 (Siegelman et al., 2017b), ASL α = .68 (this study) and ASL α = .73 (Siegelman et al., 2018, Experiment 1b), and TOL ICC = .78 (this study) and TOL split-half reliability = .59 (Buffington & Morgan-Short, 2018).

RESEARCH QUESTION 1: CONVERGENT VALIDITY OF IMPLICIT-STATISTICAL LEARNING APTITUDE: TO WHAT EXTENT DO DIFFERENT MEASURES OF IMPLICIT-STATISTICAL LEARNING APTITUDE INTERRELATE?

Correlational Analysis Among Aptitude Measures

To examine the unidimensionality of implicit-statistical learning aptitude and the interrelationships between different aptitude measures, we ran a correlation matrix between the four implicit-statistical learning aptitude measures. Table 5 presents the Spearman correlation matrix of ASRT, VSL, ASL, and TOL. We note a medium correlation between the VSL and ASL tasks (rs = .492, p < .001). At the same time, correlations of the ASRT and TOL with other tasks are low ($-.146 \leq rs \leq .054$). These results suggest that ASL and VSL may tap into a common underlying ability, statistical learning, whereas performance on other measures of implicit-statistical learning aptitude was essentially unrelated. In sum, the correlation analysis provides initial evidence for the lack of convergent validity of measures of implicit-statistical learning aptitude.

Exploratory Factor Analysis

As the second and final step in answering research question 1, we conducted an EFA with the same four measures. The Kaiser–Meyer–Olkin (KMO) measure suggested that, at the

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

TABLE 4. Descriptive statistics of all measures of L2 morphosyntactic knowledge and all cognitive measures.

	n	Mean	SD	Min	Max	Skewness	Kurtosis	*k*	Reliability
SPR (*z*, Δ msec)	110	0.06	0.30	−0.60	0.92	0.13	−0.21	96	α = .66
WMT (*z*, Δ msec)	101	0.06	0.25	−0.67	0.67	−0.22	0.07	96	α = .62
OP (%)	115	0.81	0.15	0.38	1.00	−0.68	−0.15	250-word story	α = .96[c]
EI (total correct)[f]	101	1.51	0.31	0.70	2.10	−0.11	−0.70	24	α = .73[d]
TAGJT (%)	124	0.56	0.15	0.21	1.00	0.29	0.42	24	α = .67[e]
TWGJT (%)	113	0.60	0.14	0.13	0.92	−0.30	0.29	24	α = .51[e]
UAGJT (%)	124	0.45	0.22	0.00	1.00	0.41	−0.21	24	α = .59[e]
UWGJT (%)	113	0.62	0.24	0.00	1.00	−0.47	−0.61	24	α = .65[e]
MKT (%)	120	0.64	0.17	0.21	1.00	−0.42	−0.28	12	α = .78
ASRT (Δ msec)	109	−1.15	27.78	−115.13	105.64	−0.61	4.08	10 blocks of 40 trials	ICC = .96[b]
VSL (total correct)	112	24.02	6.30	9.00	39.00	0.40	−0.64	42	α = .75
ASL (total correct)	108	20.75	5.36	9.00	33.00	0.24	−0.59	42	α = .68[a]
TOL (Δ msec)	121	0.02	0.30	−0.86	0.62	−0.36	−0.25	28 sets	ICC = .78

Abbreviations: ASL, Auditory Statistical Learning; ASRT, Alternating Serial Reaction Time; EI, elicited imitation; MKT, metalinguistic knowledge test; OP, oral production; SPR, self-paced reading; TAGJT, Timed Aural Grammaticality Judgment Test; TWGJT, Timed Written Grammaticality Judgment Test; TOL, Tower of London; UAGJT, Untimed Aural Grammaticality Judgment Test; UWGJT, Untimed Written Grammaticality Judgment Test; VSL, Visual Statistical Learning; WMT, word monitoring test.
Notes: *z*, standardized score; Δ, difference score; msec, milliseconds.
[a]Four items with negative item-total correlation were excluded from reliability analysis and from the final dataset.
[b]The reliability was computed separately for Random and Pattern trials and both were above .957.
[c]Pearson *r* intercoder correlation.
[d]The reliability of EI is an average score of two versions.
[e]The reliability scores of the four GJTs are an average score of the structure-level reliability of eight versions.
[f]Total correct on the EI was rescaled by a factor of .10, yielding a total score out of 2.4.

TABLE 5. Intercorrelation between four cognitive aptitude measures.

	ASRT	VSL	ASL	TOL
ASRT	1			
VSL	0.054	1		
ASL	0.038	0.492***	1	
TOL	−0.146	0.040	−0.070	1

Abbreviations: ASL, Auditory Statistical Learning; ASRT, Alternating Serial Reaction Time; TOL, Tower of London; VSL, Visual Statistical Learning.
Note: ***$p < .001$.

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

TABLE 6. Summary of EFA.

	Factor 1	Factor 2	Factor 3
ASRT	**1.00**	0.00	0.00
VSL	0.00	0.05	**0.71**
ASL	0.00	−0.08	**0.63**
TOL	0.00	**0.98**	0.00
Eigenvalues	1.46	1.08	0.92
% of variance	0.25	0.24	0.22
Cumulative variance	0.25	0.49	0.72

Abbreviations: ASL, Auditory Statistical Learning; ASRT, Alternating Serial Reaction Time; TOL, Tower of London; VSL, Visual Statistical Learning.
Note: Bold values indicate loadings above 0.40.

group-level, the sampling for the analysis was close to the minimum KMO of .50 (KMO = .49). At an individual test level, most tests were near the .50 cutoff point (ASRT = .52; VSL = .49; ASL = .49) with TOL reaching a bit short (.43). Despite the low KMO, we decided to keep all measures in the analysis because they were theoretically motivated. Bartlett's test of sphericity, $\chi^2(6) = 31.367$, $p < .001$, indicated that the correlations between tests were sufficiently large for an EFA. Using an eigenvalue cutoff of 1.0, there were three factors that explained a cumulative variance of 72% (the third factor accounted for a substantial increase in the explained variance, that is, 22%, and was thus included even though the eigenvalue was slightly short of 1.0). Table 6 details the factor loadings post rotation using a factor criterion of .40. As can be seen in Table 6, factor 1 represents motor sequence learning (ASRT), factor 2 represents procedural memory (TOL), and the last factor represents statistical learning, with VSL and ASL loading together.

RESEARCH QUESTION 2: PREDICTIVE VALIDITY OF IMPLICIT-STATISTICAL LEARNING APTITUDE: TO WHAT EXTENT DO MEASURES OF IMPLICIT-STATISTICAL LEARNING APTITUDE PREDICT EXPLICIT, AUTOMATIZED EXPLICIT, AND IMPLICIT KNOWLEDGE?

Confirmatory Factor Analysis

To address the second research question, we first constructed measurement models as a part of SEM to examine the number of dimensions in the nine linguistic tests. As seen in Table 2, we specified two CFA models based on SLA theory: a two-factor model distinguishing implicit versus explicit knowledge (Ellis, 2005) and a three-factor model distinguishing implicit versus automatized explicit versus explicit knowledge (an extension of Suzuki & DeKeyser, 2017). The models differed critically with regard to whether the reaction-time tasks (WMT, SPR) and the timed, accuracy-based measures (OP, EI, TAGJT, TWGJT) loaded onto the same factor, "implicit knowledge," in the two-factor solution, or different factors, "implicit knowledge" and "automatized explicit knowledge," in the three-factor solution (see Table 2).

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

TABLE 7. Summary of fit indices for the measurement models (n = 131).

	χ^2	df	CFI	SRMR	RMSEA [lower, upper]	BIC
Criterion	Nonsignificant		>.96	<.09	Lower bound: <.06	
Two-factor	0.56	27	1.00	0.07	0.00 [0.00, 0.06]	1659.36
Three-factor	0.67	26	1.00	0.07	0.00 [0.00, 0.06]	1661.19

Abbreviations: CFI, comparative fit index; RMSEA, root mean square error of approximation; SRMR, standardized root mean square.

The summary of the fit indices for the measurement models in Table 7 suggests that both models fit the data well, meeting general guidelines by Hu and Bentler (1999). At the same time, the two-factor model demonstrates a better fit than the three-factor model with a Bayesian information criterion value smaller than the three-factor model (ΔBIC ranging between 2 and 6 denotes a positive difference in favor of the model with the lower BIC; see Kass & Raftery, 1995).

Correlational Analysis of Aptitude Measures and Knowledge Measures

Before running the SEM, we first explored the correlations between implicit-statistical learning aptitude and linguistic knowledge measures. Figure 1 contains Spearman correlation coefficients (above the diagonal) of the 13 variables, scatterplots for variable pairs (below the diagonal), and density plots for each variable (on the diagonal). The results suggest that ASRT correlated significantly and positively with the WMT ($rs = .335$, $p = .002$) and the TWGJT ($rs = .229$, $p = .024$). In contrast, VSL ($rs = -.341$, $p = .001$) and, to a lesser extent, ASL ($rs = -.184$, $p = .095$) correlated negatively with the WMT. TOL did not correlate significantly with any of the linguistic knowledge measures ($-.128 \leq rs \leq .069$).

Structural Equation Model

As the final step in answering research question 2, we fitted the structural model to the measurement model to examine aptitude-knowledge relationships. In light of the EFA findings where VSL and ASL clustered into a single factor, we built a latent predictor variable called *statistical learning* (SL), which combined the ASL and VSL. Consequently, we retained three measures of implicit-statistical learning aptitude (SL, TOL, and ASRT) and treated these as predictor variables of different knowledge constructs to examine the aptitude-knowledge relationships. In the measurement model, we allowed for the different knowledge constructs (i.e., explicit, automatized explicit, and implicit knowledge) to correlate because they represent different subcomponents of language proficiency and thus we assumed that they would be related. Figures 2 and 3 show the results of the analyses.[3]

Table 8 details model fit indices for the two-factor and three-factor SEM models. Two out of the four global fit indices, namely the chi-square test and CFI, fell short of the cutoff points proposed by Hu and Bentler (1999); the SRMR was slightly above the .09

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

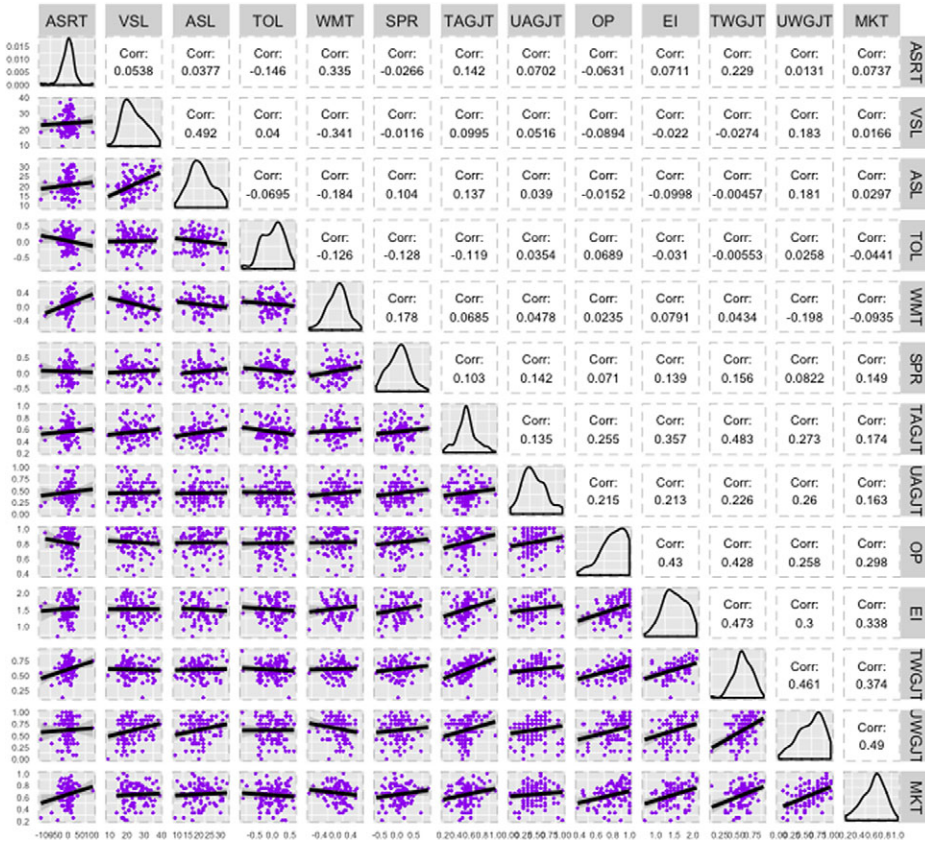

FIGURE 1. Relationships among implicit-statistical aptitude measures and linguistic tests.

threshold. To diagnose any sources of model misspecification, we inspected the modification indices and standardized residuals. In the two-factor SEM model, two modification indices were larger than 3.84, signaling localized areas of potential ill fits. Both modification indices concerned the WMT, which had a low factor loading onto implicit knowledge. The modifications were not implemented, however, as they lacked theoretical justifications (i.e., one recommended WMT as an explicit measure [MI = 4.89] and another suggested WMT as a SL measure [MI = 4.65]). No standardized residual for any of the indicators was greater than |1.96| (largest = 1.73). In the three-factor model, 12 modification indices were larger than 3.84. Based on this information, we modified the model by adding a method effect (error covariance) between EI and OP to account for the fact that EI and OP are both production tasks. Other modification indices lacked a theoretical or methodological underpinning and, hence, were not pursued further. As detailed in Table 8, adding the error covariance changed the global fit of the modified three-factor model mostly positively (i.e., chi-square p value: 0.02 → 0.03; CFI: .843 → .863; lower bound RMSEA: 0.028 → 0.019) but also negatively (i.e., SRMR: 0.094 → 0.095). No standardized residual for any of the variables was greater than |1.96|;

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

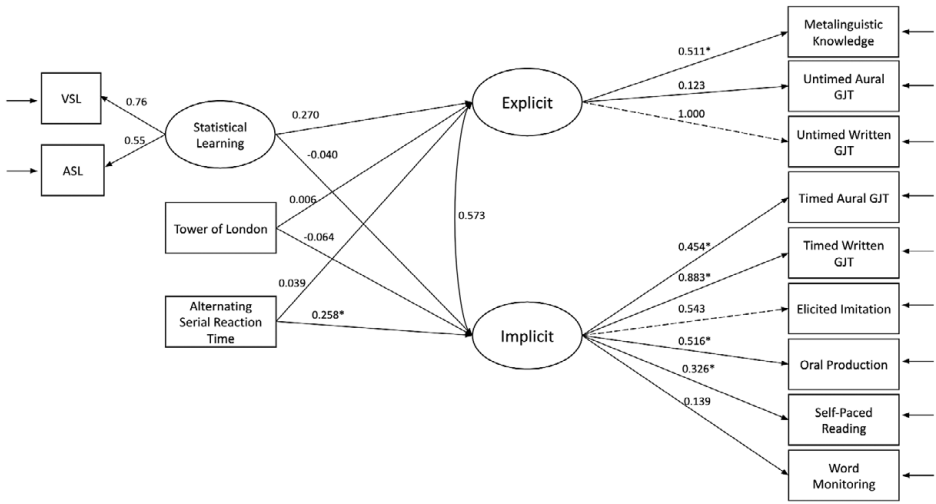

FIGURE 2. Two-factor SEM model.
*$p < .05$

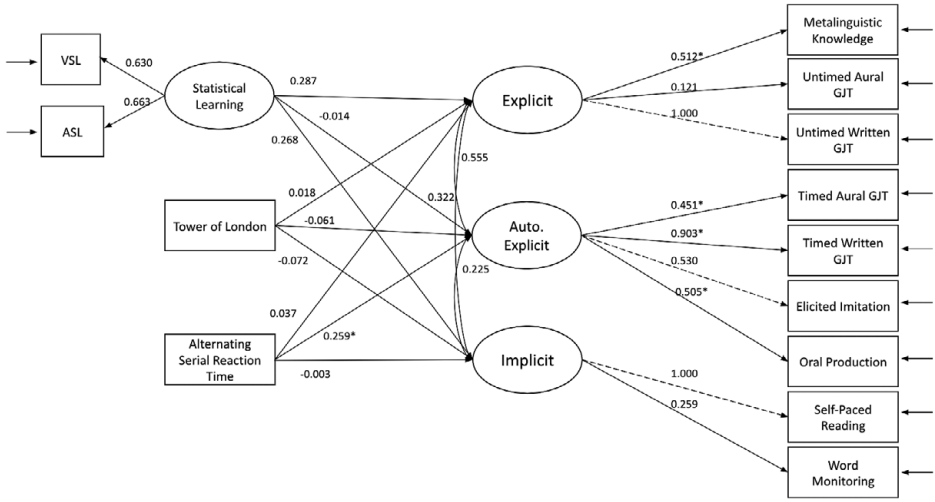

FIGURE 3. Three-factor SEM model.
*$p < .05$

however, the standardized residual for the WMT-ASRT covariance was slightly above the threshold (Std. residual = 1.97), indicating an area of local strain.

Taken together, the two-factor model exhibited a better local fit than the three-factor model, which suggested that it represented our data best. Global fit indices were somewhat low, possibly due to sample size limitations, but importantly, the underlying measurement models (CFA) demonstrated a good fit (see Table 7). As such, we proceeded to interpret the parameter estimates of the two-factor model.

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

TABLE 8. Summary of fit indices for the SEM models (n = 131).

	χ^2	df	CFI	SRMR	RMSEA [lower, upper]
Criterion	Nonsignificant		>.96	<.09	Lower bound: <.06
Two-factor	0.01	60	.811	.099	0.069 [0.036, 0.097]
Three-factor	0.02	56	.843	.094	0.065 [0.028, 0.095]
Modified three-factor	0.03	55	.863	.095	0.061 [0.020, 0.092]

Abbreviations: CFI, comparative fit index; RMSEA, root mean square error of approximation; SRMR, standardized root mean square.

TABLE 9. Two-factor SEM model parameter estimates.

Path	Estimate	SE	p	Standardized est.
ASRT →				
Implicit	0.002	0.001	0.007	0.258
Explicit	0.000	0.001	0.656	0.039
SL →				
Implicit	−0.015	0.061	0.805	−0.040
Explicit	0.147	0.129	0.255	0.270
TOL →				
Implicit	−0.036	0.070	0.609	−0.064
Explicit	0.005	0.094	0.957	0.006

Abbreviations: ASRT, alternating serial reaction time; SL, statistical learning; TOL, Tower of London.

Table 9 and Figure 2 detail parameter estimates for the two-factor SEM model. As seen in Table 9, the regression path from ASRT to implicit knowledge was significant, r = .258, p = .007. None of the other aptitude measures were significantly predicting ability in implicit or explicit knowledge in the model.

DISCUSSION

SUMMARY OF RESULTS

We aimed to contribute to the theorization of implicit-statistical learning aptitude as an individual differences variable that may be of special importance for attaining an advanced L2 proficiency (Linck et al., 2013). To measure implicit-statistical learning aptitude more comprehensively, we included two new measures—ASL and VSL (Siegelman et al., 2017b, 2018)—to the better-known measures of ASRT and TOL. Overall, only ASL and VSL showed a medium-strong correlation (r = .49) and loaded onto the same factor, whereas the remaining measures were not correlated (RQ1). This underlines that implicit-statistical learning aptitude is a multidimensional, multifaceted construct and that input modality is an important facet of the construct. A multitest approach, measuring aptitude in different input streams and task conditions, is best suited to ensure its predictive validity for language learning.

Given the theoretical importance of implicit-statistical learning aptitude, we also examined its predictive validity for implicit language knowledge, using a battery of nine

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

L2 grammar tests. The final SEM consisted of three aptitude measures regressed on a two-factor measurement model—explicit and implicit knowledge. We found that only ASRT predicted implicit knowledge, which was a latent variable composed of timed, accuracy-based measures and reaction-time tasks. These results inform ongoing debates about the nature of implicit knowledge in SLA (Ellis, 2005; Suzuki & DeKeyser, 2017) and do not lend support to the view that reaction time measures are inherently superior for measuring L2 speakers' implicit knowledge (Suzuki & DeKeyser, 2015; Vafaee et al., 2017).

MULTIDIMENSIONAL NATURE OF IMPLICIT-STATISTICAL LEARNING APTITUDE (RQ1)

Research on implicit-statistical learning aptitude can be traced back to different research traditions within cognitive and developmental psychology (Christiansen, 2019). The domain-general mechanisms that enable implicit-statistical learning have been linked to a range of different linguistic behaviors—from speech segmentation and vocabulary acquisition, to syntactic processing and literacy development (see Armstrong et al., 2017; Monaghan & Rebuschat, 2019, for recent theoretical discussions). Given the explanatory power of implicit-statistical learning aptitude in language research, we first examined the convergent validity of different measures used to assess learners' aptitude.

The results of our EFA did not support the unidimensionality of the different implicit-statistical learning aptitude measures (see Table 6). At a descriptive level, bivariate correlations between the different aptitude measures were close to 0, with the exception of ASL and VSL, which showed a .49 correlation. Correspondingly, in the EFA, the three-factor solution indicated that the battery of aptitude tests does not represent a unitary construct of implicit-statistical learning aptitude. Three factors were extracted: Factor 1 [ASRT] = .25; Factor 2 [TOL] = .24; Factor 3 [ASL and VSL] = .22, which together accounted for 72% of the total variance.

The medium strength correlation between the measures of statistical learning replicated Siegelman et al. (2018, experiment 2), who reported a .55 correlation between the ASL and VSL. The ASL and VSL are similar in terms of the nature of the embedded statistical regularity, length of training, and the way statistical learning is assessed (Siegelman et al., 2017a, 2018). Given that the tests are similar other than with regard to their input modality, these measures jointly offer a relatively pure test of the role of input modality in statistical learning. The results of the EFA showed that a common underlying ability, statistical learning, accounted for approximately 22% of the variance in participants' ASL and VSL performance, while differences in input modality accounted for some portion of the remaining 78% of variance. Input modality is therefore likely to be an important source of individual differences in statistical learning (Frost et al., 2015). These modality-based differences in statistical learning aptitude are relevant to adult L2 learners insofar as learners experience a mix of written and spoken input that may shift according to their instructed or naturalistic learning environments. For instance, Kim and Godfroid (2019, experiment 2) reported an advantage for visual over auditory input in the L2 acquisition of implicit knowledge of syntax by college-educated adults. While results of correlation research are best interpreted cumulatively, across different research studies, the medium-strong ASL-VSL correlation in the present study is consistent with the view (Arciuli,

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

2017; Frost et al., 2015; Siegelman et al., 2017b) that statistical learning is a domain-general process that is not uniform across modalities.

Seen in this light, it is interesting that the other assessment of statistical learning in the visual modality, the ASRT, showed no correlation with the VSL (see Table 5). Both tests use nonverbal material to assess an individual's ability to extract transitional probabilities from visual input. The ASRT has an added motor component, which may have contributed to the lack of convergence between the two measures. Additionally, VSL and ASRT may not have correlated because of *when* learning was assessed. Learning on the ASRT was tracked *online*, during the task, as a reaction time improvement (speed up) over training. In the ASL and VSL, however, assessment of learning took place *offline*, in a separate multiple-choice test that came after the training phase. It has been argued that the conscious reflection involved in offline tasks may confound the largely implicit learning that characterizes statistical learning (Christiansen, 2019). Online measures of implicit-statistical learning such as the ASRT, however, may be able to capture learning with a higher resolution (Siegelman et al., 2017b) and a better signal-to-noise ratio. Although more research is needed to evaluate these claims, our results support the superiority of online measurement. Using structural equation modeling, we confirmed the predictive validity of the ASRT for predicting implicit grammar knowledge in a sample of advanced L2 speakers (see the next section on RQ2 for further discussion). Conversely, the VSL or ASL did not have predictive validity for L2 implicit grammar knowledge in this study, potentially because the two measures of statistical learning allowed for participants' conscious involvement on the posttests. To investigate this result in more depth, researchers could reexamine the predictive validity of the ASL and VSL for specific grammar structures in our test battery such as embedded questions or third-person *-s*, which contain a clear, patterned regularity that lends itself well to statistical learning.

Lastly, the ASL, VSL, and ASRT were unrelated to the TOL. The TOL task finds its origin in research on planning and executive function (Shallice, 1982) and was used in a modified form, as a measure of cognitive skill learning, in Ouellet et al. (2004). Because TOL measures the effects of practice, it can be regarded as a measure of skill acquisition (Ouellet et al., 2004) and is assumed to reflect procedural learning (Ouellet et al., 2004) and provide a measure of individual differences in procedural memory ability (e.g., Antoniou et al., 2016; Buffington & Morgan-Short, 2018; Buffington et al., 2021; Ettlinger et al., 2014; Morgan-Short et al., 2014). The contributions of procedural memory to implicit-statistical learning are complex (Batterink et al., 2019; Williams, 2020). Batterink et al. (2019) reported that "a common theme that emerges across implicit learning and statistical learning paradigms is that there is frequently interaction or competition between the declarative and nondeclarative [e.g., procedural] memory systems of the brain…. Even in paradigms that have been specifically designed to isolate 'implicit learning' per se, healthy learners completing these tasks may show behavioral evidence of having acquired both declarative and nondeclarative memory" (p. 485, our addition in brackets). This interaction between declarative and nondeclarative memory in implicit learning tasks could explain the lack of convergent validity between TOL and the other measures of implicit-statistical learning aptitude; that is, measures of implicit-statistical learning may draw on multiple memory systems including, but not limited to, procedural memory. Our results are consistent with Buffington and Morgan-Short (2018) and Buffington et al. (2021), who also reported a lack of correlation between the

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

ASRT and TOL in two samples of university-level research participants ($r = -.03$, $n = 27$ and $r = .03$, $n = 99$).

The TOL does not involve patterned stimuli like the other three measures in this study, but focuses instead on an individual's improvement (accuracy gains or speed up) in solving spatial problems as a result of practice. The lack of predictive validity for implicit knowledge in advanced L2 speakers creates a need for further research into the learning processes and memory systems engaged by the TOL. TOL is indeed measuring practice, but our results, in addition to those of Buffington and colleagues (2021), do not support the claim that such practice effects reflect an individual's procedural memory learning ability. Further research into the construct validity of the TOL will be necessary. To facilitate future validation efforts, it would be helpful to standardize the use of the TOL task in L2 research. Multiple task versions, with and without repeating trials, as well as with accuracy scores versus with reaction times, are currently used in parallel in SLA, which renders comparisons of results across studies difficult (compare Antoniou et al., 2016; Buffington & Morgan-Short, 2018; Ettlinger et al., 2014, who used a task version with repeating trials, with Morgan-Short et al., 2014; Pili-Moss et al., 2019; Suzuki, 2017, who used a non-repeating version of the task). To this point, Kaller and colleagues (2016) published the TOL-F, an accuracy-based version of the TOL with improved psychometric properties that is still new in L2 research but could be of great value to achieve greater standardization in the field.

On balance, our results suggest that the findings for implicit-statistical learning aptitude do not generalize beyond the measure with which they were obtained. Future researchers will therefore need to continue treating different tests of implicit-statistical learning aptitude as noninterchangeable. For maximum generalizability, it will be important to continue using a multitest approach as exemplified in the present study. Including multiple tests of implicit-statistical learning aptitude will ensure proper representation of the substantive domain and may help researchers steer clear of confirmation bias. Over time, it will also enable researchers to refine their understanding of the different dimensions of implicit-statistical learning aptitude (Siegelman et al., 2017a) and come to a more nuanced understanding of these dimensions' roles, or nonroles, in different L2 learning environments, for learners of different ages and education levels, and with different target structures. Our call for a multitest approach echoes common practice in explicit learning aptitude research, where researchers routinely administer a battery of different tests to language learners to measure their aptitudes (see Kalra et al., 2019; Li, 2015, 2016).

ONLY TIMED, ACCURACY-BASED TESTS SUPPORTED AS MEASURES OF IMPLICIT KNOWLEDGE (RQ2)

This study was conducted against the background of an ongoing debate about how best to measure L2 learners' implicit knowledge. Measures of implicit-statistical learning aptitude can inform the construct validity of different tests—timed, accuracy-based tests and reaction time tasks—by revealing associations of aptitude with these hypothetical measures of implicit knowledge (DeKeyser, 2012; Granena, 2013). The results of this study support the predictive validity of implicit-statistical learning aptitude (ASRT) for performance on timed language tests, affirming the validity of timed, accuracy-based tests as measures of implicit knowledge (Ellis, 2005). Similar support for the validity of reaction-

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

time–based tests was lacking (cf. Suzuki & DeKeyser, 2017), which emphasized that our understanding of reaction-time measures of linguistic knowledge is still at an early stage.

We find these results to be intriguing. The two reaction-time tasks in the study, WMT and SPR, rely on the same mechanism of grammatical sensitivity (i.e., slower responses to ungrammatical than grammatical sentences) to capture an individual's linguistic knowledge. It has been assumed, often without much challenge, that grammatical sensitivity on reaction-time tests operates outside the participants' awareness, and hence may represent the participants' linguistic competence or implicit knowledge (for a critical discussion of this assumption, see Godfroid, 2020; Marsden et al., 2018). But in spite of the underlying similarity between the two tasks, performance on the SPR and the WMT correlated weakly, $rs = .178$, $p = .098$ (see Figure 1), and the two tasks loaded poorly onto the implicit knowledge factor in the CFA/SEM analysis (SPR, Std. Est. = 0.225; WMT, Std. Est. = 0.054). This indicates that current models of L2 linguistic knowledge do not account well for participants' performance on reaction-time tasks.

The construct validity of reaction time measures of linguistic knowledge cannot be separated from the instrument reliability. Compared to the accuracy-based tasks in the study, learners' performance on the WMT and SPR (the two reaction time tasks) was somewhat less reliable (see Table 4 for a comprehensive review on the validity and reliability of the nine linguistic measures). This has been a fairly consistent observation for reaction time measures, and in particular reaction time difference measures used in individual differences research (e.g., Draheim et al., 2019; Hedge et al., 2018; Rouder & Haaf, 2019), such as the grammatical sensitivity scores calculated for SPR and WMT in this study. Draheim et al. (2019) pointed out that researchers who work with reaction time difference measures often see one task "dominate" a factor, with other measures loading poorly onto the same factor. This is exactly what happened in the three-factor SEM model, where the implicit knowledge factor accounted perfectly for participants' SPR performance, but did not explain much variance in WMT scores. The three-factor model was abandoned for a simpler, two-factor SEM model, but that model did not account well for either reaction-time measure (see Figure 2 and Appendix S3 in online Supplementary Materials). These results suggest that reaction-time tests of linguistic knowledge are not a homogeneous whole (either inherently or because of lack of internal consistency), in spite of their shared methodological features. Therefore, given the current state of affairs, claims about their construct validity ought to be refined to the level of individual tests, for instance WMT or SPR separately, rather than reaction time measures as a whole.

To illustrate, we performed a post-hoc correlation analysis of the ASRT with WMT and SPR separately. We found that the ASRT correlated significantly and positively with the WMT (Spearman rank, $rs = .335$, $p = .002$), mirroring the global result for implicit knowledge (i.e., the latent variable, which was also predicted by the ASRT). SPR did not correlate with the ASRT ($rs = -.027$, $p = .804$) or with other measures of implicit-statistical learning aptitude. These results suggest that at the individual-test level, the WMT has some characteristics of a measure of implicit knowledge, consistent with earlier findings from Granena (2013) and Suzuki and DeKeyser (2015). No such evidence for SPR was obtained in this study.

Last but not least, our results revealed a significant association between implicit-statistical learning aptitude (the ASRT) and a latent factor that included four timed, accuracy-based tests (TWGJT, TAGJT, EI, OP). This supported the validity of these

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

measures as implicit knowledge tests (Ellis, 2005). Successful performance on the timed, accuracy-based measures requires fast and accurate processing of targeted grammatical knowledge. The ASRT, however, is an entirely nonlinguistic (nonverbal) task that requires fast and accurate motor responses from participants. To obtain a high aptitude score on the ASRT, participants need to speed up over time as they induce the repeating patterns in the motor sequence. One possible account for the ASRT-implicit knowledge relationship, therefore, is that both measures rely on participants' procedural memory (also see Buffington et al., 2021). On this account, the ASRT derives its validity as a predictor of implicit knowledge because it taps into the same neural substrate as implicit knowledge of language does, namely procedural memory. Similarly to procedural memory representations, implicit knowledge takes time to develop. This may explain why in previous studies, as in the present one, the SRT and ASRT predicted performance in proficient or near-native L2 learners (Granena, 2013; Linck et al., 2013; Suzuki & DeKeyser, 2015; but see Suzuki & DeKeyser, 2017; Tagarelli et al., 2016) or predicted collocational knowledge in L1 speakers and not L2 speakers (Yi, 2018). For researchers who may not have the resources to include multiple measures of implicit-statistical learning, the SRT or ASRT may thus be the best, single-test option to gain insight into the nature of learner processes or linguistic outcomes (also see Kaufman et al., 2010, who referred to the SRT as "the best measure of implicit learning currently available," p. 325).

CONCLUSION

We examined the contributions of implicit-statistical learning aptitude to implicit L2 grammar knowledge. Our results are a part of an ongoing, interdisciplinary research effort, designed to uncover the role of domain-general mechanisms in first and second language acquisition. Implicit-statistical learning aptitude was found to differ along multiple dimensions, suggesting a need for caution when generalizing results from a specific test (e.g., ASRT) to the larger theoretical constructs of implicit learning, statistical learning, and procedural memory because results may be specific to the test with which they were obtained, and the theoretical constructs may not be unitary in nature.

We also adduced support for the validity of timed, accuracy-based knowledge tests (i.e., OP, EI, timed auditory/written GJTs) as measures of implicit knowledge, supporting their use in the language classroom, language assessment, and lab-based language research to assess implicit grammar knowledge. Reaction time measures (i.e., SPR, word monitoring) currently do not enjoy the same level of validity evidence, in spite of their widespread use in lab-based research.

Despite its contributions, this study had some limitations that must be considered when interpreting the results. First, our participants were highly heterogeneous in their L1s, language learning contexts, and length of residence in an English-speaking country. Nearly half of our participants were Chinese, who may have had a jagged profile of explicit and implicit knowledge. Differences in L1 background could invite possible transfer effects (both positive and negative) across the tasks and structures. This study would also have benefited from a larger sample size, both for the EFA and the SEM. Lastly, it will be crucial to establish a good test-retest reliability for the different measures of implicit-statistical learning aptitude in future research (see Kalra et al., 2019;

https://doi.org/10.1017/S0272263121000085 Published online by Cambridge University Press

Siegelman & Frost, 2015) to show that these aptitude measures can serve as stable individual differences measures that preserve rank order between individuals over time.

Nonetheless, the results of this study help reconcile different theoretical positions regarding the measurement of L2 implicit knowledge by affirming the validity of timed, accuracy-based tests. They also point to the validity and reliability of reaction-time measures as an important area for future research. We would very much welcome other researchers to advance this research agenda and hope that the test battery developed for this project will help contribute to this goal.

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S0272263121000085.

NOTES

[1]We have chosen to adopt the term "implicit-statistical learning" based on Conway and Christiansen (2006), Perruchet and Pacton (2006), Reber (2015), Christiansen (2019), and Rebuschat and Monaghan (2019), in which these authors make arguments for combining the two approaches of implicit learning and statistical learning into one phenomenon due to their similar focus, ancestry, and use of artificial languages.

[2]Although knowledge represented in procedural memory is implicit (inaccessible to awareness), both declarative and procedural memory underlie implicit knowledge, suggesting procedural memory and implicit knowledge are related but not isomorphic (Batterink et al., 2019; Ullman, 2020; Williams, 2020).

[3]The full covariance matrix with error covariances for each figure is available from the authors upon request.

REFERENCES

Antoniou, M., Ettlinger, M., & Wong, P. C. M. (2016). Complexity, training paradigm design, and the contribution of memory subsystems to grammar learning. *PLoS ONE*, *11*, 1–20. https://doi.org/10.1371/journal.pone.0158812.

Arciuli, J. (2017). The multi-component nature of statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*, 1–9. https://doi.org/10.1098/rstb.2016.0058.

Armstrong, B. C., Frost, R., & Christiansen, M. H. (2017). (Eds.). New frontiers for statistical learning in the cognitive sciences [thematic issue]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*. https://doi.org/10.1098/rstb.2016.0047

Batterink, L. J., Paller, K. A., & Reber, P. J. (2019). Understanding the neural bases of implicit and statistical learning. *Topics in Cognitive Science*, *11*, 482–503. https://doi.org/10.1111/tops.12420.

Bowles, M. A. (2011). Measuring implicit and explicit linguistic knowledge: What can heritage language learners contribute? *Studies in Second Language Acquisition*, *33*, 247–271. https://doi.org/10.1017/S0272263110000756.

Brooks, P. J., & Kempe, V. (2013). Individual differences in adult foreign language learning: The mediating effect of metalinguistic awareness. *Memory & Cognition*, *41*, 281–296. https://doi.org/10.3758/s13421-012-0262-9.

Buffington, J., Demos, A. P., & Morgan-Short, K. (2021, forthcoming). The reliability and validity of procedural memory assessments used in second language learning. *Studies in Second Language Acquisition.*

Buffington, J., & Morgan-Short, K. (2018). Construct validity of procedural memory tasks used in adult-learned language. In C. Kalish, M. Rau, J. Zhu, & T. T. Rogers (Eds.), *Proceedings of the 40th annual conference of the cognitive science society* (pp. 1420–1425). Cognitive Science Society.

Christiansen, M. H. (2019). Implicit statistical learning: A tale of two literatures. *Topics in Cognitive Science*, *11*, 468–481. https://doi.org/10.1111/tops.12332.

Conway, C. M., & Christiansen, M. H. (2006). Statistical learning within and between modalities: Pitting abstract against stimulus-specific representations. *Psychological Science*, *17*, 905–912. https://doi.org/10.1111/j.1467-9280.2006.01801.x.

DeKeyser, R. M. (2003). Implicit and explicit learning. In C. J. Doughty & H. M. Long (Eds.), *The handbook of second language acquisition* (pp. 312–348). Blackwell.

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

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin*, *145*, 508. https://doi.org/10.1037/bul0000192.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*, 141–172. https://doi.org/10.1017/S0272263105050096.

Ellis, R. (2009). Measuring implicit and explicit knowledge of a second language. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 31–64). Multilingual Matters.

Ellis, R., & Loewen, S. (2007). Confirming the operational definitions of explicit and implicit knowledge in Ellis (2005): Responding to Isemonger. *Studies in Second Language Acquisition*, *29*, 119–126. http://www.jstor.org/stable/44488647.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, *27*, 464–491. https://doi.org/10.1093/applin/aml001.

Ettlinger, M., Bradlow, A. R., & Wong, P. C. (2014). Variability in the learning of complex morphophonology. *Applied Psycholinguistics*, *35*, 807–831. https://doi.org/10.1017/S0142716412000586.

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*, 67–101. https://doi.org/10.1177/0267658316684903.

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*, 1128–1153. https://doi.org/10.1037/bul0000210.

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences*, *19*, 117–125. https://doi.org/10.1016/j.tics.2014.12.010.

Godfroid, A. (2020). *Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide*. Routledge.

Godfroid, A., Kim, K. M., Hui, B., & Isbell, D. (2018). *Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide*. Routledge.

Godfroid, A., Loewen, S., Jung, S., Park, J., Gass, S., & Ellis, R. (2015). Timed and untimed grammaticality judgments measure distinct types of knowledge. Evidence from eye-movement patterns. *Studies in Second Language Acquisition*, *37*, 269–297.

Granena, G. (2013). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning*, *63*, 665–703. https://doi.org/10.1111/lang.12018.

Granena, G. (2019). Cognitive aptitudes and L2 speaking proficiency: Links between LAMA and HI–LAB. *Studies in Second Language Acquisition*, *41*, 313–336. https://doi.org/10.1017/S0272263118000256.

Gutiérrez, X. (2013). The construct validity of grammaticality judgment tests as measures of implicit and explicit knowledge. *Studies in Second Language Acquisition*, *35*, 423–449. https://www.jstor.org/stable/26336217.

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

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*, 1166–1186. https://doi.org/10.3758/s13428-017-0935-1.

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

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55. https://doi.org/10.1080/10705519909540118.

Kaller, C. P., Debelak, R., Köstering, L., Egle, J., Rahm, B., Wild, P. S., Blettner, M., Beutel, M. E., & Unterrainer, J. M. (2016). Assessing planning ability across the adult life span: Population-representative and age-adjusted reliability estimates for the Tower of London (TOL-F). *Archives of Clinical Neuropsychology*, *31*, 148–164.

Kaller, C. P., Rahm, B., Köstering, L., & Unterrainer, J. M. (2011). Reviewing the impact of problem structure on planning: A software tool for analyzing tower tasks. *Behavioural Brain Research*, *216*, 1–8. https://doi.org/10.1016/j.bbr.2010.07.029.

Kalra, P. B., Gabrieli, J. D., & Finn, A. S. (2019). Evidence of stable individual differences in implicit learning. *Cognition*, *190*, 199–211. https://doi.org/10.1016/j.cognition.2019.05.007.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition*, *116*, 321–340. https://doi.org/10.1016/j.cognition.2010.05.011.

Kim, J., & Nam, H. (2017). Measures of implicit knowledge revisited: Processing modes, time pressure, and modality. *Studies in Second Language Acquisition*, *39*, 431–457. https://doi.org/10.1017/S0272263115000510.

Kim, K. M., & Godfroid, A. (2019). Should we listen or read? Modality effects in implicit and explicit knowledge. *The Modern Language Journal*, *103*, 648–664.

Li, S. (2015). The associations between language aptitude and second language grammar acquisition: A meta-analytic review of five decades of research. *Applied Linguistics*, *36*, 385–408. https://doi.org/10.1093/applin/amu054.

Li, S. (2016). The construct validity of language aptitude. *Studies in Second Language Acquisition*, *38*, 801–842. https://doi.org/10.1017/S027226311500042X.

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., Smith, B. K., Bunting, M. F., & Doughty, C. J. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning*, *63*, 530–566. https://doi.org/10.1111/lang.12011.

Marsden, E., Thompson, S., & Plonsky, L. (2018). A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*, *39*, 861–904. https://doi.org/10.1017/S0142716418000036.

McDonough, K. I. M., & Trofimovich, P. (2016). The role of statistical learning and working memory in L2 speakers' pattern learning. *The Modern Language Journal*, *100*, 428–445. https://doi.org/10.1111/modl.12331.

Misyak, J. B., & Christiansen, M. H. (2012). Statistical learning and language: An individual differences study. *Language Learning*, *62*, 302–331. https://doi.org/10.1111/j.1467-9922.2010.00626.x.

Monaghan, P., & Rebuschat, P. (2019). (Eds.). Aligning implicit learning and statistical learning: Two approaches, one phenomenon [special issue]. *Topics in Cognitive Science*, 11. https://doi.org/10.1111/tops.12364

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

Ouellet, M. C., Beauchamp, M. H., Owen, A. M., & Doyon, J. (2004). Acquiring a cognitive skill with a new repeating version of the Tower of London task. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *58*, 272–288. https://doi.org/10.1037/h0087450.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*, 233–238. https://doi.org/10.1016/j.tics.2006.03.006.

Pili-Moss, D., & Brill–Schuetz, K. A., Faretta-Stutenberg, M., & Morgan-Short, K. (2019). Contributions of declarative and procedural memory to accuracy and automatization during second language practice. *Bilingualism: Language and Cognition*, *23*, 1–13. https://doi.org/10.1017/S1366728919000543.

Core Team, R. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/.

Reber, A. S. (2015). Foreword. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages*. John Benjamins.

Rebuschat, P., & Monaghan, P. (2019). Editors' introduction: Aligning implicit learning and statistical learning: Two approaches, one phenomenon. *Topics in Cognitive Science*, *11*, 459–467. https://doi.org/10.1111/tops.12438.

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin and Review*, *26*, 452–467.

Shallice, T. (1982). Specific impairments of planning. Philosophical Transactions of the Royal Society of London. B. *Biological Sciences*, *298*, 199–209. https://doi.org/10.1098/rstb.1982.0082.

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017a). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*, 1–. https://doi.org/10.1098/rstb.2016.0059.

Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition*, *177*, 198–213. https://doi.org/10.1016/j.cognition.2018.04.011.

Siegelman, N., Bogaerts, L., & Frost, R. (2017b). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavioral Research Methods*, *49*, 418–432.

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, *81*, 105–120.

Skehan, P. (2015). Foreign language aptitude and its relationship with grammar: A critical overview. *Applied Linguistics*, *36*, 367–384. https://doi.org/10.1093/applin/amu072.

Spada, N., Shiu, J. L. J., & Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies. *Language Learning*, *65*, 723–751. https://doi.org/10.1111/lang.12129.

Suzuki, Y. (2017). The optimal distribution of practice for the acquisition of L2 morphology: A conceptual replication and extension. *Language Learning*, *67*, 512–545. https://doi.org/10.1111/lang.12236.

Suzuki, Y., & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language Learning*, *65*, 860–895. https://doi.org/10.1111/lang.12138.

Suzuki, Y., & DeKeyser, R. (2017). The interface of explicit and implicit knowledge in a second language: Insights from individual differences in cognitive aptitudes. *Language Learning*, *67*, 747–790. https://doi.org/10.1111/lang.12241.

Tagarelli, K. M., Ruiz, S., Vega, J. L. M., & Rebuschat, P. (2016). Variability in second language learning: The roles of individual differences, learning conditions, and linguistic complexity. *Studies in Second Language Acquisition*, *38*, 293–316. https://doi.org/10.1017/S0272263116000036.

Ullman, M. T. (2020). The declarative/procedural model: A neurobiologically motivated theory of first and second language. In B. VanPatten, G. Keating, & S. Wulff (Eds.), *Theories in second language acquisition: An introduction* (3rd ed., pp. 128–161). Routledge.

Ullman, M. T., Earle, F. S., Walenski, M., & Janacsek, K. (2020). The neurocognition of developmental disorders of language. *Annual Review of Psychology*, *71*, 389–417. doi:10.1146/annurev-psych-122216-011555.

Vafaee, P., Suzuki, Y., & Kachinske, I. (2017). Validating grammaticality judgment tests: Evidence from two new psycholinguistic measures. *Studies in Second Language Acquisition*, *39*, 59–95. https://doi.org/10.1017/S0272263115000455.

Williams, J. N. (2020). The neuroscience of implicit learning. *Language Learning*, *70*, 255–307. https://doi.org/10.1111/lang.12405.

Yi, W. (2018). Statistical sensitivity, cognitive aptitudes, and processing of collocations. *Studies in Second Language Acquisition*, *40*, 831–856. https://doi.org/10.1017/S0272263118000141.

Zhang, R. (2015). Measuring university-level L2 learners' implicit and explicit linguistic knowledge. *Studies in Second Language Acquisition*, *37*, 457–486. https://doi.org/10.1017/S0272263114000370.