

Methods Forum

THE RELIABILITY AND VALIDITY OF PROCEDURAL MEMORY ASSESSMENTS USED IN SECOND LANGUAGE ACQUISITION RESEARCH

Joshua Buffington *

University of Illinois at Chicago

Alexander P. Demos 

University of Illinois at Chicago


Kara Morgan-Short *

University of Illinois at Chicago

Abstract

Evidence for the role of procedural memory in second language (L2) acquisition has emerged in our field. However, little is known about the reliability and validity of the procedural memory measures used in this research. The present study ($N = 119$) examined the reliability and the convergent and discriminant validity of three assessments that have previously been used to examine procedural memory learning ability in L2 acquisition, the dual-task Weather Prediction Task (DT-WPT), the Alternating Serial Reaction Time Task (ASRT), and the Tower of London (TOL). Measures of declarative memory learning ability were also collected. For reliability, the DT-WPT and TOL tasks met acceptable standards. For validity, an exploratory factor analysis did not provide evidence for convergent validity, but the ASRT and the TOL showed reasonable discriminant validity with declarative memory measures. We argue that the ASRT may provide the purest engagement of procedural memory learning ability, although more reliable dependent measures for this task should

The research reported here came out of Joshua Buffington's MA thesis. We would like to acknowledge the following people for their feedback on this work: Susan R. Goldman, James W. Pellegrino, and Members of the Cognition of Second Language Acquisition Laboratory. We are also grateful for the UIC Award for Graduate Research for funding part of this study.

 The experiment in this article earned an Open Data badge for transparent practices. The materials are available at <https://osf.io/ux4qs/>.

* Correspondence concerning this article should be addressed to Joshua Buffington, Department of Psychology (m/c 285), University of Illinois at Chicago, 1007 W. Harrison St., Chicago, Illinois 60607-7137. E-mail: bfingtn2@uic.edu; or Kara Morgan-Short, Department of Hispanic and Italian Studies (m/c 315), University of Illinois at Chicago, 601 S. Morgan St., Chicago, Illinois 60607. E-mail: karams@uic.edu

be considered. The Serial Reaction Time task also appears promising, although we recommend further consideration of this task as the present analyses were post hoc and based on a smaller sample. We discuss these results regarding the assessment of procedural memory learning ability as well as implications for implicit language aptitude.

Recent considerations of aptitude for second language (L2) research include implicit language aptitude (e.g., Granena, 2013; Suzuki & DeKeyser, 2017), a multifaceted construct defined as “cognitive abilities that facilitate implicit learning and processing of an L2” (Granena, 2020, p. 7). From the various implicit learning and memory abilities (Reber, 2013; Squire & Dede, 2015), theoretical perspectives of L2 acquisition have posited a role for procedural memory and knowledge (DeKeyser, 2020; Paradis, 2009; Ullman, 2020). Given the implicit nature of procedural memory, understanding its role in L2 should be informative to perspectives on implicit language aptitude. Indeed, empirical research that has specifically investigated the role of procedural memory in L2 grammar suggests that individual differences in procedural memory learning ability are associated with L2 development at least at later stages of learning (see Buffington & Morgan-Short, 2019; Hamrick et al., 2018). This body of research has used various assessments of procedural memory learning ability but has yet to establish the reliability and validity of these assessments, which is critical for the internal validity of each study as well as for the robustness of conclusions that can be drawn across studies. Accordingly, in the present study we examine the reliability and validity of assessments commonly used in L2 studies that explicitly aimed to examine the role of procedural memory in L2. Although we conceptualized this study within a framework focused on procedural memory learning ability, the findings may also have important implications for implicit language aptitude (Granena, 2020) and implicit learning more generally (Kalra et al., 2019).

PROCEDURAL MEMORY AND ITS RELATION TO L2 ACQUISITION

Extensive behavioral, neurophysiological, and neuroimaging work has provided evidence for dissociable learning and memory systems (Eichenbaum, 2012; Gabrieli, 1998; Reber, 2013; Squire, 1994; Squire et al., 1994; Ullman, 2004, 2016; Ullman et al., 2020; Willingham et al., 2002). Here, we focus on procedural memory and use a general conceptualization of procedural memory gleaned from several perspectives (e.g., Eichenbaum, 2012; Gabrieli, 1998; Squire & Dede, 2015; Ullman et al., 2020). Specifically, procedural memory is considered to be one type of implicit learning and memory system that supports the acquisition of cognitive and motor skills and habits.¹ Procedural memory may be contrasted with other memory systems such as declarative memory, which supports the acquisition of facts and personal experiences (Eichenbaum, 2012; Gabrieli, 1998; Squire & Dede, 2015; Ullman et al., 2020).

Learning supported by procedural memory shows several cognitive and neuroanatomical characteristics. For one, learning in procedural memory is believed to be implicit, in that it does not involve conscious awareness (Ullman et al., 2020). Relatedly, learning in procedural memory is not supported by attention, at least for cognitive skills, and indeed attention may interfere with learning in procedural memory for cognitive skills (Foerde et al., 2006). The development of knowledge in procedural memory occurs gradually as opposed to the relatively fast acquisition of knowledge in declarative memory (Ullman

et al., 2020). Additionally, knowledge in procedural memory is typically encapsulated, meaning that it is generally inflexible with respect to the contexts in which it can be applied (Squire & Dede, 2015). Finally, neuroanatomically, procedural memory relies on a fronto-striatal circuit (including frontal cortex, the basal ganglia, and the thalamus) as well as other neural substrates such as the cerebellum, as compared to declarative memory, which relies on the medial temporal lobe and connected cortical regions (Eichenbaum, 2012; Ullman et al., 2020).

Three theoretical perspectives predict a role for procedural memory and knowledge in L2 acquisition: Skill Acquisition Theory (DeKeyser, 2020), Ullman's Declarative/Procedural Model (Ullman, 2020), and Paradis' Neurolinguistic Theory of Bilingualism (Paradis, 2009). Although there are important differences among these perspectives (Buffington & Morgan-Short, 2019; Morgan-Short & Ullman, *in press*), they agree in viewing procedural memory and knowledge as involved in the fluent production and comprehension of grammar in a second language, at least at higher levels of proficiency. Empirical research has begun to examine the association between procedural memory learning ability and L2 grammatical development (Antoniou et al., 2016; Ettliger et al., 2014; Faretta-Stutenberg & Morgan-Short, 2018; Hamrick, 2015; Morgan-Short et al., 2014; Morgan-Short et al., 2015; Pili-Moss et al., 2020; Suzuki, 2018). Indeed, a recent meta-analysis has shown that procedural memory learning ability is associated with L2 grammar abilities at higher, but not lower, levels of proficiency (Hamrick et al., 2018).

As this research expands to provide a more fine-grained understanding of the role of procedural memory in L2, the robustness and the interpretability of the findings will naturally depend on the validity and reliability of the tasks and measures used to assess procedural memory learning ability, among other factors. In examining the empirical research that has the explicitly stated purpose of testing the role of procedural memory in L2, we find a number of different tasks and measures (Table 1),² including different tasks that have been used by our research team.³ To date, four tasks have been used in L2 procedural memory research: the Alternating Serial Reaction Time task (ASRT; J. H. Howard & D. V. Howard, 1997), the dual-task version of the Weather Prediction Task (DT-WPT; Foerde et al., 2006; based on the single-task version, Knowlton et al., 1994; Knowlton et al., 1996), the Tower of London (TOL; Kaller et al., 2012; Shallice, 1982), and the Serial Reaction Time task (Hamrick, 2015; Lum et al., 2012; based on the task originally reported in Nissen & Bullemer, 1987). For some tasks, researchers have also used different measures of learning. For example, for the TOL, Ettliger et al. (2014) quantified improvement on the task as the overall trial solution time on a second administration of the task, whereas Morgan-Short et al. (2014) measured the change in initial think time between the first and last trials of each block on one administration of the task (and combined this measure with the measure from the DT-WPT to create a composite score). The use of different tasks and measures is not problematic in and of itself. To the extent that each task is a valid measure of the construct of interest, results that are reproduced across these tasks suggest generalizability of the findings. However, if a field of inquiry is built upon different tasks that have not been validated, then both the internal validity of each study and the overall validity of the conclusions drawn across studies cannot be known. In addition to validity, it is vital that the reliability of each task be established so that we know that the construct of interest is measured consistently. As inquiry about the role of procedural memory in L2 expands, it is important to consider

TABLE 1. Studies examining the relationship between L2 acquisition and procedural memory

Reference	Context of learning	L2	Procedural memory task	Measure(s) of procedural learning
Eitlinger et al., 2014	Passive and exposure-based	Artificial language: Morphophonology based on Shimakonde	TOL	Overall solution time (total time from presentation of trial to solution) on second administration of task
Antoniou et al., 2016	Passive and exposure-based	Artificial language: Shimakonde	TOL	Overall solution time (total time from presentation of trial to solution) on second administration of task
Morgan-Short et al., 2014	Implicit	Artificial language: Brocanto2	TOL DT-WPT	Initial think time (time from presentation of trial to first move) Accuracy on final block; combined TOL and DT-WPT into composite measure
Pili-Moss et al., 2020 ^a	Implicit	Artificial language: Brocanto2	TOL DT-WPT	Initial think time (time from presentation of trial to first move) Accuracy on final block; combined TOL and DT-WPT into composite measure
Morgan-Short et al., 2015	Implicit	Artificial language: Brocanto2	TOL DT-WPT	Initial think time (time from presentation of trial to first move) Accuracy on final block; combined TOL and DT-WPT into composite measure
Brill-Schuetz & Morgan-Short, 2014	Implicit and explicit	Artificial language: Brocanto2	DT-WPT ASRT	Accuracy on final block Average response time on patterned vs. random trials; Combined DT-WPT and ASRT into composite measure
Suzuki, 2018	Explicit	Miniature language based on Spanish: Supurango	TOL	Initial think time (time from presentation of trial to first move), Movement execution time (time from first move to solution of the trial), Overall solution time (total time from presentation of trial to solution)
Hamrick, 2015	Incidental	Semiartificial language: Syntax based on Persian	SRT	Rebound score (final pseudorandom RT—final pattern RT)
Tagarelli et al., 2016	Implicit and explicit	Semiartificial language: Syntax based on German	(A)SRT ^b	SRT: Rebound score (final pseudorandom RT—final pattern RT); ASRT: Sum of blocks that showed a learning effect, measured as faster RTs on pattern vs. random trials
Faretta-Stutenberg & Morgan-Short, 2018	Study-abroad and at-home / classroom	Spanish	DT-WPT	Accuracy on final block

TABLE 1. Continued

Reference	Context of learning	L2	Procedural memory task	Measure(s) of procedural learning
			ASRT	Average RT on patterned vs. random trials; Combined DT-WPT and ASRT into composite measure

Note: Shows previous studies that examined the role of procedural memory in L2. TOL = Tower of London; DT-WPT = DT Weather Prediction Task; ASRT = Alternating Serial Reaction Time; SRT = Serial Reaction Time.

^aPili-Moss et al. (2020) analyzed unreported data from the Morgan-Short et al. (2014) study.

^bFor technical reasons, some participants in Tagarelli et al. (2016) completed the ASRT, and others completed the SRT. Scores for both tasks were standardized for analysis.

evidence for the validity and reliability of the tasks being used to assess procedural memory learning ability.

EVIDENCE SUPPORTING THE RELIABILITY AND VALIDITY OF PROCEDURAL MEMORY TASKS

What is the evidence for the validity and reliability of the tasks that have been used to assess procedural memory learning ability in L2 research? Here we consider the most commonly used tasks: the ASRT, the DT-WPT, and the TOL (see Figure 1 for images of these tasks).⁴ Currently, there is limited evidence for the reliability of these tasks, especially for the specific measures that have been used in the L2 literature. Support for validity is largely based on (a) the design of the tasks, such as the instructions provided to participants, the design of the stimuli, and/or dual-task procedures (described in the following text and detailed in the “Methods” section), and (b) previous research that shows how performance on these tasks reflects the characteristics of procedural memory and may involve the engagement of the particular neural substrates tied to procedural memory. In the text that follows we review the characteristics of the task design and typical patterns of task performance for the ASRT, DT-WPT, and TOL as well as research findings indicating that neural substrates associated with procedural memory may underpin performance on these tasks.

ALTERNATING SERIAL REACTION TIME TASK

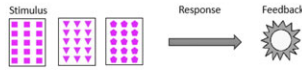
The ASRT is a sequence learning task in which an item of the sequence consists of a filled-in circle in a row of four circles (see Figure 1 for examples of circles filled in with dog heads). The sequence in which the circles are filled in follows a second-order pattern where patterned trials alternate with random trials. As such, participants see a repeating sequence such as 3r1r4r2r, where the numbers correspond to the location of the filled-in circle and “r” represents a random location. Participants are instructed to press a key corresponding to the location of the filled-in circle as quickly and accurately as possible. Learning is typically assessed by the difference in reaction times to target versus nontarget

Procedural Memory Tasks

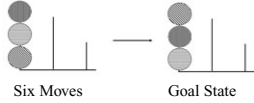
Alternating Serial Reaction Task



Weather Prediction Task



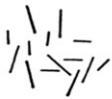
Tower of London



Declarative Memory Tasks

Continuous Visual Memory Test

New or old?



Declearn

Encoding

Real or made up?



Recognition

New or old?



MLAT, Part V

Study

- hij-draw
- naq-that
- sidqu-news

Test

- hij?
- A-frog
- B-fall
- C-cold
- D-draw
- E-book

FIGURE 1. Procedural and declarative tasks.

Note: Shows sample images from each of the procedural and declarative memory learning ability tasks used in the present study.

trials. For example, J. H. Howard and D. V. Howard (1997) examined reaction times on pattern versus random trials. Song, Howard, & Howard (2007) examined reaction times to highly frequent triplets of trials versus low-frequency triplets, and Nemeth et al. (2013) examined reaction times to high versus low frequency random triplets, among other measures. Regarding the reliability of the ASRT, a previous study demonstrated test-retest reliability ($r = .46, p = .04$) based on high- versus low-frequency triplets (Stark-Inbar et al., 2017). Within L2 research on procedural memory (Table 1), Faretta-Stutenberg and Morgan-Short (2018) reported that internal consistency reliability was acceptable (split-half reliability based on Spearman-Brown correlation = .921) based on all items (i.e., pattern and random items combined) split between the first and second halves of the task. However, additional analyses of the reliability for the specific measure of reliability based on the pattern versus random difference score is warranted.

Regarding the ASRT’s validity as a measure of procedural memory learning ability, previous research provides evidence for the use of procedural memory in acquiring the sequence in this task. In J. H. Howard and D. V. Howard (1997) learning was characterized by reaction times that gradually became faster on patterned trials than on random trials, despite the fact that participants were unable to accurately describe the regularity in the sequence, suggesting that their knowledge may have been implicit. Regarding the engagement of neural substrates on the ASRT, evidence appears to be consistent with the use of procedural memory (for review see J. H. Howard & D. V. Howard, 2013), although research has focused on triplet measures with the ASRT along with a related Triplet-Learning Task (J. H. Howard et al., 2008), perhaps in part because of the lack of clear neuroimaging contrasts with the pattern versus random items on this task. J. H. Howard and D. V. Howard (2013) review neuroimaging and neuropsychological studies on the ASRT, among other lines of research (e.g., genetic factors). Concerning neuroimaging, connections between the caudate nucleus, part of the basal ganglia, and the dorsolateral prefrontal cortex have been implicated in ASRT learning. Relatedly, neuropsychological

research points to ASRT learning deficits in patients who suffer from corticobasal syndrome, which involves degeneration in the basal ganglia. Overall, behavioral and neural research on the ASRT is suggestive of the use of procedural memory due to gradual, implicit learning that appears to be supported by the basal ganglia.

WEATHER PREDICTION TASK

In the WPT, participants are instructed to predict the “weather” (which is a choice between fictional “sunshine” or “rain”) based on a combination of cues (see [Figure 1](#) for example cards with geometric shapes serving as cues). Each combination of cues is associated with a certain probability of sunshine or rain. For example, a combination of a card with circles and a card with squares may be associated with an 80% chance of sunshine. As a secondary distractor task, participants are also tasked with keeping track of the number of high tones that occur during each trial (Foerde et al., 2006). Learning is typically assessed by examining whether participants’ predictions of “sunshine” or “rain” are optimal based on the probabilistic cue-outcome associations. For example, Foerde et al. counted optimal responses as accurate on a probe block administered after the task. Kalra et al. (2019) calculated the optimal response rate versus chance on each block during the task. Test-retest reliability on the learning measure used in the first and second administrations of the task (i.e., optimal response rate vs. chance on the third block of a four-block version of this task) showed that the two administrations were significantly correlated: $r = .39$, $p = .002$. (Kalra et al., 2019). Within L2 research on procedural memory ([Table 1](#)), Faretta-Stutenberg and Morgan-Short (2018) reported that the internal consistency of the WPT was acceptable (Cronbach’s alpha = .746) for performance on the final dual-task block. However, additional evidence for the internal consistency for this specific measure would further inform our understanding of its reliability.

Evidence from previous research suggests that the DT-WPT may be a valid assessment of procedural memory learning ability. Foerde and colleagues examined cognitive learning characteristics and neural substrates of performance on the DT-WPT as compared to a single-task WPT. Regarding cognitive characteristics, the dual-task version reduced the amount of declarative, explicit knowledge of cue-outcome associations compared to the single-task version (Foerde et al., 2006; Foerde et al., 2007), even as implicit knowledge of cue-outcome associations was evidenced. Neuroimaging results (Foerde et al., 2006) suggested that performance on the dual-task, but not the single-task, version was positively associated with activity in the striatum, whereas the single-task, but not the dual-task, version showed a positive relationship with activity in the medial temporal lobe. Neuropsychological evidence also suggests that WPT performance (for a single-task version) depends on neural substrates associated with procedural memory. Knowlton et al. (1996) showed that patients with Parkinson’s disease, which involves basal ganglia degeneration, were impaired relative to healthy controls on this task, whereas amnesic patients, who had damage to the medial temporal lobe, were unimpaired on the task despite a lack of memory for the training sessions.⁵ A role for procedural memory neural substrates in learning on the WPT is also supported by other studies (see Batterink et al., 2019 for review). Overall, these patterns of results suggest the use of procedural memory on the WPT, especially for the dual-task version.

TOWER OF LONDON

In the TOL, participants are instructed to match a goal configuration of colored circles that rest on pegs (Figure 1). In producing the goal configuration, participants are constrained by being able to move only the topmost circle on each peg, and, when moved, the circle will fall to the lowest possible peg position. Participants are instructed to plan their sequence of moves before beginning the first move, and to do their best in matching the goal configuration in the stated number of moves. Performance on the TOL can be quantified by accuracy and various reaction time measures. For example, Unterrainer et al. (2019) examined accuracy as the percentage of the correctly solved problems in the stated number of moves, whereas Unterrainer et al. (2003) examined reaction time measures, including the preplanning time (time from presentation of problem to first move, which is also described as “initial think time”; Morgan-Short et al., 2014) and movement execution time (time from first move to solution of the problem). Psychometric work with versions of the TOL that are similar to that administered by the present study have demonstrated acceptable levels of reliability specific to internal consistency for accuracy on a 32-item TOL (reliability estimates ranged from .691 to .828; Kaller et al., 2012) and on a 24-item TOL (overall reliability estimates ranged from .715 to .757; Unterrainer et al., 2019). In L2 research, analysis of the TOL has primarily been based on reaction time measures from the task. Thus, reliability should be examined for these measures.

Some research with the TOL suggests that it can be used as a valid measure of procedural memory learning ability, even as it has also been commonly used to examine processes in other cognitive areas, such as problem solving (i.e., planning ability, for example see Kaller et al., 2011). In research examining the cognitive characteristics of learning on the TOL, Ouellet et al. (2004) demonstrated that participants gradually improve in both accuracy and the time to complete this task over blocks of trials. They also demonstrated an improvement on goal configurations that were repeated compared to nonrepeated goal configurations, despite the fact that participants were unable to describe specific information about the repeated sequences, suggesting the use of implicit knowledge to solve these problems. As such, behavioral data indicate that participants may rely on procedural memory when acquiring the TOL due to the gradual improvement over time and use of implicit knowledge to solve the problems. In a neuroimaging study, Beauchamp et al. (2003) provided evidence that improved performance on the TOL is associated with activity in the fronto-striatal network, among other brain regions. The use of the fronto-striatal network on the TOL, along with early activation of other brain structures involved in procedural memory (e.g., cerebellum), is consistent with the use of procedural memory on this task. Lastly, neuropsychological evidence indicates that patients with Parkinson’s disease, which affects procedural memory neural circuits, show an impairment on the TOL that is associated with the severity of the disease (Owen et al., 1992). Also, whereas healthy controls rely on a fronto-striatal network when acquiring the TOL, patients with Parkinson’s disease show reduced activity in the fronto-striatal network and elevated activity in brain regions associated with declarative memory (Beauchamp et al., 2008; Dagher et al., 2001). A similar pattern of findings has also been evidenced in patients with obsessive compulsive disorder, which is associated with frontostriatal abnormality (van den Heuvel et al., 2005). Taken together, the pattern of

results across cognitive and neuroimaging evidence suggests that procedural memory may underlie learning on the TOL.

MOTIVATION AND RESEARCH QUESTIONS

In sum, previous research has provided some evidence for the reliability and validity of the ASRT, the DT-WPT, and the TOL as measures of procedural memory learning ability. However, regarding reliability, further examination is needed as the measures examined in previous research are largely not the same measures typically used in L2 research about procedural memory. Regarding validity, the overall pattern of evidence is consistent with the use of these tasks as measures of procedural memory learning ability. For one, the tasks show reasonable face validity (a subjective evaluation of whether the task captures the construct; Morling, 2015) in that all three tasks seem to involve gradual, implicit learning, which is characteristic of learning supported by procedural memory. In addition, some evidence for construct validity (how well the task measures the construct it claims to measure; Morling, 2015) is provided with neuroimaging and neuropsychological research that shows the engagement of neural substrates associated with procedural memory during task performance (although these neural substrates may also be engaged by other cognitive processes). However, very little evidence exists regarding the convergent and discriminant validity of these tasks, that is, whether the tasks are related to each other (as one might expect if they all measure a broad procedural memory learning ability construct) and whether they are unrelated to tasks that measure other constructs (such as declarative memory learning ability). Consequently, the current study directly examines the reliability and the convergent and discriminant validity of these tasks to increase our understanding of their overall reliability and validity in L2 research. Evidence of convergent validity of the tasks would establish that the tasks, though different, each reflect a broad construct of procedural memory learning ability. Interestingly, a recent study provided preliminary evidence that these tasks may not correlate positively with each other, suggesting that they may not measure the same cognitive ability, although a relatively low number of participants precludes strong conclusions (Buffington & Morgan-Short, 2018). Evidence of discriminant validity would establish that the tasks are not associated with other constructs such as declarative memory. Accordingly, we examined the following research questions:

RQ1 (reliability): Do assessments of procedural memory learning ability demonstrate internal consistency?

RQ2 (convergent validity): Do the assessments of procedural memory learning ability pattern positively together?

RQ3 (discriminant validity): Do assessments of procedural memory learning ability *not* pattern positively with assessments of declarative memory learning ability?

To test these research questions, we administered all three procedural memory learning ability assessments to participants in a within-subjects design. We also included three assessments of declarative memory learning ability to investigate discriminant validity.

METHOD

PARTICIPANTS

Participants ($N = 119$) received course credit for participation through the subject pool at the University of Illinois at Chicago. Twenty participants were excluded for reasons such as not following task instructions or extreme performance on a task (see Supplemental Methods for additional information). This left $N = 99$ participants (58 women, 41 men; average age = 19.30 years; age range = 17–29 years).

MATERIALS

PROCEDURAL MEMORY LEARNING ABILITY

Participants completed three assessments of procedural memory learning ability (Figure 1). For details on all tasks used in the present study, see Supplementary Materials. The first assessment, the ASRT (Csabi et al., 2016), was based on the original task from J. H. Howard and D. V. Howard (1997) and involved responding to the serial location of a target that followed an alternating pattern, such that every other trial was part of the pattern. Learning was measured by subtracting the reaction time on pattern trials from the reaction time on random trials. The second assessment of procedural memory was the dual-task version of the WPT (Foerde et al., 2006; based on the single-task version; Knowlton et al., 1994; Knowlton et al., 1996). Participants predicted a weather outcome (sunshine or rain) using cue cards that were probabilistically associated with the weather outcomes. The dual-task component was a tone-counting task, which has been shown to impede the use of declarative memory on the weather prediction component of the task (Foerde et al., 2006). Performance was measured by accuracy on the final dual-task block. Responses were scored as accurate if they matched the optimal, or expected, response (Gluck et al., 2002). After completing the WPT, participants completed a debriefing measure of explicit knowledge (Cue Select). The final assessment of procedural memory was the TOL (Kaller et al., 2012). In this task, participants were presented with a configuration of colored circles and instructed to move the circles to match a goal configuration within a specified number of moves. To measure improvement, participants repeated the task. Two dependent measures for this task have been used in previous L2 empirical work. The first measure (TOL ImpBlock) examines the average percent change in initial think time, or planning time, with a higher percent change representing a greater decrease in initial think time and presumably more procedural learning (used in Morgan-Short et al., 2014). The second measure (TOL ImpNormed) assesses the average total time to match a goal configuration on the second administration of the task, normalized relative to the other participants, as a measure of overall improvement on the task, which may or may not be specific to procedural memory (used in Antoniou et al., 2016; Ettliger et al., 2014).

DECLARATIVE MEMORY TASKS

Participants completed three assessments of declarative memory (Figure 1). The first assessment, Part V of the Modern Language Aptitude Test (MLAT-V; Carroll & Sapon,

1959), examined participants' memory for English translations of novel pseudo-Kurdish words following a brief study phase. The dependent measure was accuracy on a multiple-choice assessment. The second assessment of declarative memory was the Continuous Visual Memory Test (CVMT; Trahan & Larrabee, 1988). Participants were presented with a series of abstract images, some of which were repeated, and tasked with answering whether they had previously seen the image. Performance was measured with the d' score. The last assessment of declarative memory, the Declearn task (Hedenius et al., 2013), involved an incidental encoding phase in which participants made real/made-up judgments for images. Following this, they were tested on their recognition memory of the shapes. Recognition performance was measured with the d' score.

PROCEDURE

Participants completed the study over two testing sessions scheduled on separate days. Both sessions lasted approximately 1.5–2 hours each. To avoid fatigue effects of the cognitive tasks, the order of assessments was partially counterbalanced such that assessments in the same session were counterbalanced and no two procedural or declarative memory assessments occurred back to back (see Supplementary Materials for additional details).⁶

ANALYSIS

All data and analyses, including an R script that reproduces the results in the present study, are available through the Open Science Foundation at the following webpage: https://osf.io/ux4qs/?view_only=d5b9ed28ad7046f18d55201a087d6e46. Before analyses were run, data was cleaned by items and participants. For the ASRT, data was first cleaned by removing reaction times shorter than 100 ms and longer than three standard deviations above the participant's average reaction time and inaccurate trials were removed (total of 9.97% were removed). For DT-WPT, all trials in Block 8 were included except for those that had a 50% probability of sunshine or rain (7.5% of trials). For the CVMT, two participants were discovered as outliers, with performance more than three standard deviations below the group mean performance. This caused us to remove those entire cases from subsequent analyses, as factor analysis with missing data is not recommended (Tabachnick & Fidell, 2013). No data cleaning was required for any other tasks.

Details of all analyses are reported in the "Results," but here an overview and rationale for the "Results" section is provided. To examine our first research question, we calculated reliability for each dependent measure, with acceptable reliability defined at .70 or greater (Lance et al., 2006; Nunally & Bernstein, 1978). Generally, if the reliability for a dependent measure was below this threshold, we excluded it from analysis for RQs 2 and 3, although we made an exception for ASRT (see "Results," "RQ1: Reliability Analysis" subsection). Following the insightful critiques of two reviewers, we calculated reliability differently depending on whether the measure of learning ability was taken over the whole task (i.e., ASRT, TOL, and CVMT) versus at the end of the task (i.e., DT-WPT, MLAT, and Declearn). The reason for this is that learning should occur in each task, which leads to dependence among the test items and may degrade interpretations of reliability when it is calculated over the whole task. Thus, for measures of learning taken

over the whole task we calculated Spearman–Brown split-half reliability based on every other item in serial order. This approach controls for the serial dependence among items, as each half involves items across the whole task. For measures of learning taken at the end of the task, we calculated Cronbach’s alpha. To provide initial evidence regarding our second and third research questions about convergent and discriminant validity, respectively, a correlation matrix was computed using Spearman correlations among all assessments in the study. For our main analysis, we conducted an exploratory factor analysis (EFA)⁷ on the Spearman correlation matrix to test if the tasks factor into the expected procedural and declarative memory factors.

To perform the EFA, assumptions including multivariate normality and sampling adequacy (i.e., sufficiently large relationships among variables in the correlation matrix, examined with the Kaiser–Meyer–Olkin Test) were evaluated. Principal Axis Factoring extraction was chosen because this extraction method does not depend on multivariate normality (M. C. Howard, 2016). Following this, factor analysis was performed with a scree plot analysis with converging evidence from parallel analysis, optimal coordinates, and the acceleration factor (Horn, 1965; Raiche et al., 2006; Raiche & Magis, 2020). Factors were then rotated with an oblique rotation (quartimin), which allows the factors to correlate with each other (M. C. Howard, 2016; Yong & Pearce, 2013). To conduct the analyses, we used R version 3.6.0 (R Core Team, 2019; see Supplementary Materials for citations of packages used).

RESULTS

RQ 1: RELIABILITY ANALYSIS

The first step in the analysis was to evaluate RQ 1: Do assessments of procedural memory learning ability demonstrate internal consistency? Results of the reliability analysis are displayed in Table 2 and described in the following text. For thoroughness, descriptive statistics, including means, standard deviations, normality, and learning effects are also displayed in Table 2. Normality was evaluated with histograms, Q-Q plots, and the Shapiro–Wilk test. If the distribution of a measure was skewed based on visual inspection or the Shapiro–Wilk test was significant (indicating a departure from normality), then the measure received an “X” for normality, otherwise the measure received a check, indicating that we considered the measure to be normally distributed. Learning effects were evaluated with 95% confidence intervals on the average performance for each measure. If these intervals do not overlap with chance performance, this is taken as evidence that there was learning on the task. Learning effects are important in order to analyze learning at the group level. If group-level learning is not observed, this raises questions about the task’s validity as a measure of learning (e.g., there may be design issues with the task that prevent participants from learning). For procedural memory learning ability tasks, we also include plots of learning over time in Supplementary Materials.

ASRT evidenced normality and a learning effect that was significantly above chance performance. Chance performance was operationalized as no difference between the reaction times on pattern versus random trials, as measured by a confidence interval overlapping with 0. The reliability score of .42 for the ASRT did not meet our .70 threshold for acceptable reliability based on the pattern versus random difference score,

TABLE 2. Selection of measures to include in final analysis

	No. trials	M	SD	95% CI for average performance	Normality	Reliability
<i>Procedural memory tasks</i>						
ASRT (ms)	1,600	2.99	6.68	1.66–4.33*	✓	.42 ^{SA}
DT-WPT Block8 (%) ^b	37	59.15%	14.57%	56.25–62.06*	✓	.73 ^{CA}
<i>TOL</i>						
ImpBlock (%)	8	3.44%	41.51%	–4.83 to 11.72	×	–.12 ^{SB}
ImpNormed (standard score)	28	.51	.30	.45–.57 ^{NA}	×	.87 ^{SB}
<i>Declarative memory tasks</i>						
MLAT (accuracy)	24	15.32	5.18	14.29–16.36*	×	.84 ^{CA}
CVMT (<i>d'</i>)	96	1.32	0.73	1.18–1.47*	✓	.81 ^{SB}
Declearn (<i>d'</i>) ^c	128	1.17	0.73	1.02–1.31*	× ^d	.91 ^{CA}

Note: The units of the dependent measure for each task are shown in parentheses; NA = We note that the CI for TOL ImpNormed does not conceptually reflect a learning effect as scores on this measure do not have a baseline from which to evaluate learning (see main text); SB = Spearman–Brown split-half coefficient; CA = Cronbach’s alpha; CI = confidence interval; *above-chance learning effect.

^aThe observed split-half reliability for ASRT was negative. As such, instead of the Spearman–Brown correction we applied a correction suggested by Krus and Helmstadter (1993, eq. 15) for cases in which observed reliability is negative. We also investigated the source of this low reliability. Specifically, we looked at reliability separately for pattern and random items and then compared these reliabilities to the reliability of their difference score. We found that pattern and random items both have acceptable reliability, with values of .96 and .99, respectively. Thus, we suggest that the source of low reliability for ASRT may lie in taking the difference score between pattern and random items.

^bTo investigate if performance on DT-WPT Block8 might be associated with explicit knowledge developed during the task, we correlated the Block8 measure with the explicit debriefing task, Cue Select. There was a strong, positive correlation between these two tasks, $r(97) = .50, p < .001$. Thus, performance on the Block8 measure may rely, at least in part, on explicit knowledge, which would not be consistent with the exclusive use of procedural memory on this task.

^cFor Declearn, the dependent measure for reliability was accuracy, not *d'*, as using an accuracy measure allowed us to calculate Cronbach’s alpha, which is preferred over the Spearman–Brown split-half reliability because Cronbach’s alpha includes all possible split-half reliabilities, which is not done when calculating the Spearman–Brown split-half reliability.

^dThe Shapiro–Wilk normality test was positive for Declearn ($p = .04$), but visually the data appear to follow a normal distribution based on a histogram and Q-Q plot. So, the Declearn data could potentially be considered normally distributed.

although reliability for the pattern and random trials was acceptable (see Table 2, note a). Thus, overall, the ASRT appears to be an acceptable measure to include in the EFA.

DT-WPT task exhibited normality and a significant learning effect (with chance performance equal to 50% correct when guessing sunshine/rain). Reliability was acceptable according to the .70 threshold. Thus, this measure appears to be valid for use in the EFA.

For the TOL, we examined two dependent measures from previous L2 research. For the first measure, TOL ImpBlock, we included all trials at the beginning or end of a block from the first administration of the task (8 trials). This measure did not suggest a normal distribution and did not evidence a significant learning effect because the 95% confidence interval overlapped with chance performance (0% improvement).

Additionally, TOL ImpBlock did not show acceptable reliability ($< .70$). The second measure, TOL ImpNormed, involved performance on all trials of the second administration (28 trials). Descriptively, TOL ImpNormed does not appear to be normally distributed. Further, because there is no baseline to evaluate performance, there does not appear to be a reasonable way to measure a learning effect on TOL ImpNormed. In other words, the measure assesses learning on the second administration of the task, and thus no value represents chance learning, as would a value of zero if performance were compared between the first and second administrations. However, TOL ImpNormed did exhibit acceptable reliability. For this reason, TOL ImpNormed, but not TOL ImpBlock, was included in the EFA. Hereafter, TOL ImpNormed is referred to simply as TOL.

Next, we evaluated the measures for declarative memory. The first declarative memory measure is MLAT. This measure includes the 24 multiple-choice test items. Descriptively, the MLAT was not normally distributed but did evidence a learning effect (chance performance is 20% correct because each question has five options; expected number of correct answers by chance = 24 questions \times 20% = 4.8 questions). Reliability for the MLAT was above the acceptable threshold. Second, scores on the CVMT were based on 96 recognition trials. Descriptively, the CVMT showed both a learning effect (chance $d' = 0$) and normal distribution. Reliability was also acceptable. Third, the Declearn score includes the 128 trials on the recognition task. Descriptively, Declearn may not be normally distributed but participants did exhibit a learning effect (chance $d' = 0$). Reliability for Declearn was based on accuracy (see note c in Table 2) and was well above the minimum threshold. In sum, based on their reliability each of the memory measures appear to be valid for use in the EFA.

RQS 2 AND 3: CONVERGENT AND DISCRIMINANT VALIDITY

Correlations

As a preliminary investigation of our second and third research questions—convergent validity: do the assessments of procedural memory learning ability pattern positively together? and discriminant validity: do assessments of procedural memory learning ability *not* pattern positively with assessments of declarative memory learning ability?—we first examined a correlation matrix for all tasks. Results of this correlation matrix are displayed in Figure 2, which shows Spearman correlations among all our tasks. Spearman correlations, which measure the rank order or ordinal relationships among scores on the tasks, were used instead of Pearson correlations because the Spearman correlation can work with ordinal data, does not depend as strongly on normality, and can correct some deviations from linearity because it is based on ranks.

The results suggest a lack of correlations among procedural memory assessments, with no positive correlations and most of the correlations small to negligible in size. As such, the correlation matrix suggests that there may be a lack of convergent validity among the procedural memory learning ability assessments. Regarding discriminant validity, there appear to be three patterns of relevant preliminary results: (a) TOL showed a negative correlation with two of the declarative memory tasks (CVMT and Declearn) that was small-to-medium in size (Cohen, 1992); (b) ASRT did not correlate with any of the

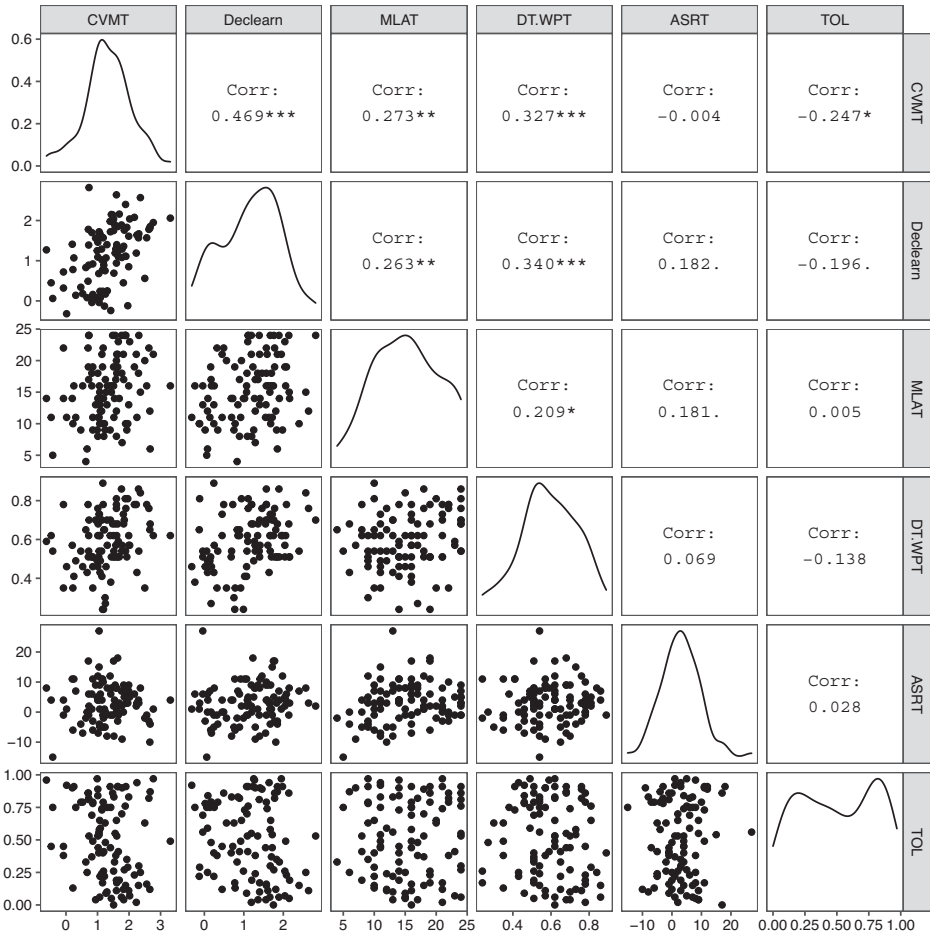


FIGURE 2. Intercorrelations for declarative and procedural memory tasks.

Note: Spearman correlations.

.*p* < .10; **p* < .05; ***p* < .01; ****p* < .001

declarative memory tasks; and (c) DT-WPT showed positive correlations with all three declarative memory tasks, with effect sizes ranging from small-to-medium to medium in size. These data suggest that TOL and ASRT may show discriminant validity with the declarative memory tasks, but DT-WPT could be patterning with declarative memory, which would be inconsistent with discriminant validity. Finally, we note that, as expected, all three declarative memory tasks (MLAT, CVMT, and Declearn) correlated significantly and positively with each other, with effect sizes ranging from small-to-medium to almost large. Overall, this preliminary analysis suggests that the three procedural memory learning ability assessments may not show convergent validity and that DT-WPT may not show discriminant validity, but in fact may pattern more strongly with the declarative memory learning ability tasks.

Exploratory Factor Analysis

For the final analysis of our second and third research questions we conducted an EFA. The overall Kaiser–Meyer–Olkin Score (KMO) score was .68 (Table 3), which is above the minimum recommended threshold. Individual tasks also had acceptable KMO scores (>.60) except for ASRT. Some researchers recommend dropping tasks with low KMO scores prior to conducting the EFA (M. C. Howard, 2016). However, as all tasks are motivated based on previous literature, and because there are a low number of tasks in the EFA to begin with, we retained all the tasks for the EFA.

Next, we created a scree plot that includes supporting evidence from parallel analysis, optimal coordinates, and the acceleration factor to decide how many factors to include in the EFA (Figure 3). Most of the results suggest two factors, and as this is consistent with two factors representing procedural and declarative memory, we conducted the EFA with two group factors.

Results from this EFA are reported in Table 4. The two factors accounted for 31% cumulative variance, with 23% and 8% for Factor 1 and 2, respectively. These factors correlated weakly at .24. Factor loadings were interpreted according to the following criteria from M. C. Howard (2016): (a) satisfactory factor loadings should load onto a

TABLE 3. Kaiser–Meyer–Olkin scores for the exploratory factor analysis

Overall	CVMT	Declearn	MLAT	DT-WPT	ASRT	TOL
.68	.67	.68	.69	.77	.52	.67

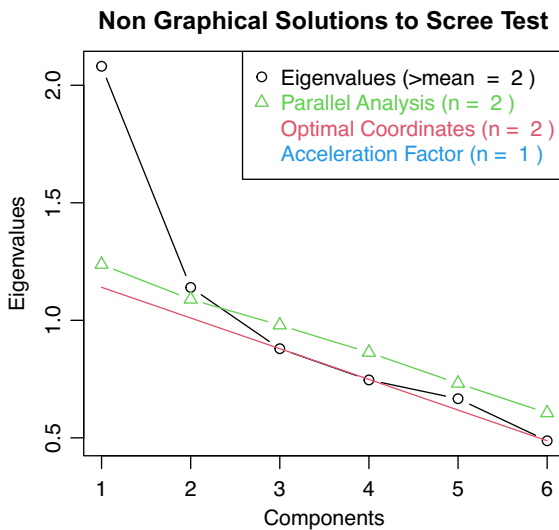


FIGURE 3. Scree plot for EFA.

Note: Shows scree plot along with converging evidence from parallel analysis, optimal coordinates, and the acceleration factor (Raiche & Magis, 2020).

TABLE 4. Factor loadings for exploratory factor analysis

	Task	F1	F2	h2	u2
<i>Declarative memory tasks</i>	CVMT	0.77	-0.10	0.57	0.43
	Declearn	0.61	0.21	0.48	0.52
	MLAT	0.27	0.33	0.22	0.78
<i>Procedural memory tasks</i>	DT-WPT	0.45	0.09	0.23	0.77
	ASRT	-0.03	0.50	0.24	0.76
	TOL	-0.35	0.18	0.13	0.87
	SS Loadings	1.37	.45		
	% variance	23%	8%		

Note: Shows results of EFA with percent variance explained by the model. F1 = first factor; F2 = second factor; h2 = communality (sum of squared factor loadings), or total variance in each task that is accounted for by F1 and F2; u2 = variance in each task not accounted for by the factors, or the task’s uniqueness; bolding indicates loadings (>.40); % variance = proportion of total variance accounted for by the factor.

primary factor at a value of .40 or greater; (b) any alternate loadings for a task should not exceed 0.30; and (c) the difference between the primary and alternate factor loadings for a task should be at least 0.20. Three tasks clearly loaded onto Factor 1: CVMT, Declearn, and DT-WPT. DT-WPT was not expected to load with CVMT and Declearn, but this is consistent with the correlation matrix. One task clearly loaded onto Factor 2: ASRT. TOL did not meet threshold for inclusion in Factor 2, but it was trending to a negative loading on Factor 1. MLAT did not load clearly onto either factor.

Given the loadings, Factor 1 suggests a declarative memory learning ability factor, as both Declearn and CVMT loaded onto this factor. Factor 2 had only one task loading with it, ASRT, but none of the other procedural memory learning ability tasks loaded onto this factor and as such it does not clearly suggest a factor of procedural memory learning ability. Overall, the EFA does not appear to suggest two factors of procedural and declarative memory, although they clearly form one interpretable factor that may represent declarative memory learning ability.

DISCUSSION

Regarding our first research question—do assessments of procedural memory learning ability demonstrate internal consistency?—our results suggest that measures used for the DT-WPT and TOL tasks showed acceptable levels of internal consistency reliability (>.70). The ASRT pattern versus random measure fell below this threshold, although responses to the pattern and random trials themselves were reliable. For our second research question—do the assessments of procedural memory learning ability pattern positively together?—results from our correlation matrix did not provide evidence that these tasks correlate with each other, and the EFA did not show evidence that the tasks loaded onto the same factor. The ASRT showed an acceptable loading on the second factor, but none of the other procedural memory learning ability tasks also loaded onto this factor. Thus, our results suggest that the procedural memory learning ability tasks in the present study do not seem to measure a broad procedural memory construct. Finally, regarding our third research question—do assessments of procedural memory learning

ability *not* pattern positively with assessments of declarative memory learning ability?—the correlation matrix showed no positive correlations with declarative memory learning ability tasks for ASRT and TOL. The DT-WPT correlated positively with all three declarative memory learning ability tasks. Also, the DT-WPT, but not the ASRT or the TOL, loaded onto the first factor in the EFA that also included Declearn and CVMT. This factor seems to be indicative of a declarative memory learning ability factor, as both Declearn and CVMT showed acceptable loadings onto this factor with no cross-loadings onto the second factor. Overall, then, for the procedural memory learning ability tasks we found (a) evidence of reliability, except for the ASRT pattern versus random difference measure; (b) a lack of evidence for convergent validity among the tasks; and (c) evidence consistent with discriminant validity as compared to declarative memory learning ability for two of the tasks (ASRT and TOL, but not for DT-WPT).

Although previous L2 research has not systematically addressed these reliability and validity questions about procedural memory learning ability tasks, our findings can be interpreted in light of some findings from L2 research and from related research in cognitive psychology. First, regarding reliability, our results complement previous DT-WPT and TOL results (see literature review in the preceding text) in that their findings of acceptable internal consistency and test-retest reliability are extended to internal consistency for the specific measures used in prior L2 procedural memory work. However, the pattern versus random difference measure of ASRT used in the present study appears to have low internal consistency reliability even as internal consistency for the pattern and random trials was high (see Table 2, note a). This finding is not consistent with previous reliability results for this task although those findings were either based on the trials themselves or on another ASRT measure (Faretta-Stutenberg & Morgan-Short, 2018; Stark-Inbar et al., 2017). Interestingly, Trafimow (2015) proposes that high true correlations between tests and low deviation ratios, or ratios of the variances between tests, likely lowers the reliability of the difference score. Results for the ASRT are consistent with this claim: The true correlation between pattern and random scores was essentially perfect, at 1.02 (raw $r = .99$), and the deviation ratio was also low, based on the values supplied by Trafimow, at .99. This observation may account for why the pattern and random scores are separately reliable although their difference score has low reliability. Overall, then, our reliability results encourage confidence in the DT-WPT and TOL, insofar as these tasks appear consistent in their measurement, but reliability for the ASRT pattern versus random measure used here is low.

Second, regarding convergent validity, no patterns of association were found among the DT-WPT, ASRT, and TOL tasks. This finding was somewhat surprising given the patterns of evidence reviewed in the preceding text suggesting that performance on these tasks reflects characteristics of procedural memory, including the gradual acquisition of knowledge on the tasks, the implicit nature of this knowledge, and the engagement of neural substrates associated with procedural memory. However, the results seem consistent with research within the broader domain of implicit learning and memory. For example, Godfroid and Kim (2021) also did not find a positive relationship between ASRT and TOL when using different dependent measures than those in the current study. Further, Godfroid and Kim found that performance on these tasks loaded onto different factors in an EFA. Studies within cognitive psychology have also generally not found evidence of convergent validity among various implicit and statistical learning tasks

(Gebauer & Mackintosh, 2007; Kalra et al., 2019; Siegelman & Frost, 2015) with some claims that implicit learning is not a unitary ability or mechanism. However, Kalra et al. found a good fit for a one-factor model for three specific tasks (probabilistic classification, serial reaction time, and implicit category learning, but not for artificial grammar learning) for which learning seems to depend on the basal ganglia rather medial temporal-lobe structures. Arguably, these tasks might represent the more specific construct of procedural memory given their reliance on the basal ganglia. Interestingly, two of our tasks are similar to those used in Kalra et al.: Our DT-WPT is also a probabilistic classification task, and our ASRT is a variant for the serial reaction time task. Why then do we not find convergent validity for our three procedural memory learning tasks, which also seem to rely on the basal ganglia? The answer may partly lie in the specific task characteristics and dependent measures. For example, Kalra's probabilistic classification task was comprised of 262 single-task trials across four blocks with the dependent measure of learning being taken on the third block, whereas our DT-WPT was comprised of 320 dual-task trials across eight blocks with the dependent measure of learning being taken on the eighth block. Future research will be needed to determine whether convergent validity can be evidenced across procedural memory learning ability tasks with further consideration of their specific characteristics and dependent measures of learning. However, convergent validity might be somewhat difficult to find for procedural memory tasks as the neuro-cognitive substrates of procedural memory seem to be somewhat segregated into discrete circuits that may be involved in separate aspects of behavior (Middleton & Strick, 2000).

Finally, regarding discriminant validity, our results for the ASRT and TOL are consistent with these tasks showing discriminant validity as neither task patterned with assessments of declarative memory learning abilities. However, the DT-WPT appears to pattern with the declarative memory learning ability tasks, which generally seems inconsistent with evidence from previous research that the DT-WPT engages procedural memory (Foerde et al., 2006). Although procedural memory may always be involved at some level for this task (as evidenced by striatal activity that did not differ during training in single- versus dual-task conditions; Foerde et al., 2006, p. 11780), the task conditions seem to modulate the "relative contribution" of the declarative and procedural memory systems (p. 11781). If participants in the current study did not fully engage in the tone-counting task that created the dual-task condition, their reliance on declarative memory may not have been effectively modulated. Unfortunately, we do not have a measure of their performance on the tone counting task and are not able to assess how well they engaged in the dual-task condition. Overall, then, the present results suggest that the DT-WPT may, at least in some cases, engage declarative memory learning abilities, which we interpret to mean that the task should not be considered a process-pure measure of procedural memory learning ability unless researchers can ensure that recourse to declarative memory is effectively blocked. For the ASRT and TOL, our results suggest that they do not recruit declarative memory learning abilities in a detectable manner and can be used as individual differences measures independent of declarative memory.⁸

IMPLICATIONS FOR FUTURE L2 RESEARCH

The present results have implications for how researchers examine the role of procedural memory learning abilities in L2 acquisition. Although we do not have evidence that the

ASRT, DT-WPT, and TOL tasks jointly capture a broad construct of procedural memory, given the overall pattern of evidence presented in the literature suggesting their reliance on procedural memory, we consider the case for each task as a measure of procedural memory learning ability, depending on the measure used and the theoretical motivation for the task.

The case for the DT-WPT may be the weakest given that, in our results, it loaded on a factor that we interpret as a declarative memory learning ability factor. However, the stimuli represent probabilistic cue-outcome associations, which may be subserved by prediction-based learning in procedural memory (Ullman et al., 2020). Also, the task has been shown to engage procedural memory substrates, and a different version of the probabilistic classification task has been associated with other tasks that engage the basal ganglia (Kalra et al., 2019). Thus, we do not interpret our results as negating the engagement of procedural memory on this task. However, in future research, we suggest that the DT-WPT not be considered a process-pure measure of procedural memory learning abilities, unless researchers can ensure that declarative memory learning abilities are not recruited in the task. Perhaps using a dependent measure earlier in the task, as in Kalra et al., when declarative memory processes may not have begun to support learning (see note 5) would provide a purer measure of procedural memory learning ability.

Although the TOL did not show associations with declarative memory learning ability in our study, the task may also not be ideal as a process-pure measure of procedural memory learning ability. The “change in initial think time” TOL measure used in previous L2 research may arguably exhibit face validity regarding reduced reliance on declarative strategies before initiating a move, but this measure showed very low levels of reliability in the current study. The more general measure of overall improvement showed acceptable levels of reliability but may not show high face validity as it is an overall measure of a task that has generally been associated with planning abilities (e.g., Kaller et al., 2012; Unterrainer et al., 2019) and skill acquisition (e.g., Beauchamp et al., 2003; Ouellet et al., 2004). Although the task shows engagement of fronto-striatal networks tied to procedural memory, neural areas thought to reflect other processes involved in skill acquisition also seem to be engaged, at least early on during the task (Beauchamp et al., 2003). Thus, if research is theoretically motivated to examine a skill acquisition perspective of L2 (DeKeyser, 2020), this might be considered a valid task (depending on the measure being used). In this case, researchers might want to consider the version of the TOL used in previous skill acquisition work by (e.g., Beauchamp et al., 2003; Beauchamp et al., 2008; Ouellet et al., 2004) or the updated TOL-F used in planning research, for which psychometric examinations have been conducted (e.g., Kaller et al., 2012; Unterrainer et al., 2019). However, if a researcher is examining procedural memory from a neurobiologically based perspective (Ullman, 2020), then the TOL may not reflect a process-pure measure of procedural memory learning ability.

The ASRT seems to be a likely candidate as a valid and relatively process-pure measure of procedural memory learning ability. Learning on the ASRT is gradual, seemingly implicit, and has been shown to involve striatal activity (J. H. Howard & D. V. Howard, 2013). In the current study, the ASRT was not associated with declarative memory learning ability and was not correlated with the TOL, which may not be a process-pure task. Furthermore, the pattern versus random measure has been shown to reflect second-order statistical dependencies (J. H. Howard & D. V. Howard, 1997), which are likely to

rely on procedural memory (Ullman et al., 2020). Thus, out of the tasks examined in the current study, the ASRT might be interpreted as the most valid, process-pure task of procedural memory learning ability. However, future research should carefully consider other ASRT measures that may demonstrate higher reliability than the pattern versus random difference score examined here. Interestingly, measures based on triplets of items have been argued to specifically represent probabilistic sequence learning (i.e., high- vs. low-frequency triplets; Song et al., 2007) and statistical learning (i.e., high- vs. low-frequency random triplets; Nemeth et al., 2013). Researchers may also consider different versions of this task, such as the Triplet-Learning Task (J. H. Howard et al., 2008), for which motor learning aspects are removed, and the Serial Reaction Time task (Nissen & Bullemer, 1987), which has been widely used in cognitive psychology and for which learning relies on the basal ganglia (Janacsek et al., 2020). (See “Limitations” and Supplementary Materials for more information.) In all cases, researchers (including our own research group) should provide a clear theoretical motivation as well as validity and reliability considerations for the choice of task and dependent measure.

Because procedural memory learning ability is a domain-general implicit learning ability, the results of our study also have implications for growing and important considerations of implicit language aptitude in L2 acquisition (e.g., Granena, 2013; Suzuki & DeKeyser, 2017). Granena (2020) states that implicit aptitude is not a unitary construct and considers the distinctions between implicit learning and implicit memory as well as between declarative and nondeclarative memory systems. Regarding implicit learning and memory, the measures used in the current study arguably can be considered measures of learning, as the measures were based on continuous data throughout the learning task for the ASRT, the second administration of the TOL, and on the final learning block for the DT-WPT. The results of our EFA are generally consistent with findings that measures of implicit learning do not seem to reflect a broad construct (e.g., Gebauer & Mackintosh, 2007; Godfroid & Kim, 2021; Kalra et al., 2019; Siegelman & Frost, 2015; cf. Kalra et al. when artificial grammar learning was excluded from analysis). As pointed out by others (e.g., Godfroid & Kim, 2021; Granena, 2020; Siegelman & Frost, 2015), differences among tasks such as modality of the stimuli (e.g., auditory or visual), specificities of the stimuli (e.g., being based on letters or figures), the type of relationship among the stimuli (e.g., deterministic, probabilistic), whether motor learning is involved, and the point at which the dependent measure is taken (during learning or after) may all play a role in the lack of convergence of tasks onto a single construct.

Regarding declarative and nondeclarative memory systems, all nondeclarative memory systems, including the procedural memory system (Ullman et al., 2020), are characterized by implicit learning and memory (Squire & Dede, 2015). Although there are patterns of evidence for implicit learning and the engagement of procedural memory for all the tasks considered in the current study (see literature review), the ASRT may arguably be the task that provides the most process-pure measure of procedural memory learning ability and nondeclarative memory more generally. However, as emphasized by Granena (2020, p. 39), it is vitally important for future research to continue to examine the reliability and construct validity of nondeclarative, implicit learning and memory tasks. To the extent that research establishes the reliability and validity of procedural memory learning ability tasks, as in the current study, the findings will be informative to L2 implicit learning aptitude research (as procedural memory learning ability is one type of implicit learning

ability). Similarly, to the extent that the tasks validated in the implicit aptitude research are specific to the procedural memory and learning system (Ullman et al., 2020), implicit learning research that uses those specific tasks will be relevant to research on procedural memory learning ability (see note 3). However, it is important to emphasize that not all tasks used to assess implicit learning and memory will be relevant to procedural memory learning ability, as other types of implicit learning and memory, for example, priming, depend on separate neural systems.

We offer a brief note regarding the implications of our results on assessing declarative memory learning ability, which was not the focus of the current study. Although all three declarative memory learning ability measures correlated positively with each other, in the EFA, only the Declearn and CVMT tasks showed satisfactory factor loadings onto the factor that we interpreted as declarative memory learning ability. Thus, researchers examining the role of declarative memory in future research may consider using one of these two tasks rather than the MLAT-V. An advantage of using the Declearn or CVMT is that they are nonverbal tasks, so associations between performance on these tasks and L2 can be more strongly attributed to domain-general abilities in the declarative memory and learning system.

LIMITATIONS

The strength of our conclusions regarding the validity of these procedural memory learning ability tasks should be considered in light of the study's limitations. First, the current study did not include the Serial Reaction Time task (SRT), which was used to study the role of procedural memory learning abilities in L2 in Hamrick (2015) and partially in Tagarelli et al. (2016). We chose not to include this task in the main study because (a) it has not been as commonly used in studies that examine the role of procedural memory in L2 (see Table 1), and (b) given our within-subjects design, there may have been transfer effects between the SRT and the ASRT, as participants learn sequences in both these tasks and learning one sequence may have effects on learning a subsequent sequence. However, in a smaller follow-up study, we administered the SRT to a subset of participants ($N = 33$) who came in for a third session that was not originally planned as part of the main study (see Supplementary Materials for a description of the follow-up study, analyses, and results regarding the SRT). Results from the follow-up study indicated that the reliability of the SRT was acceptable (Spearman-Brown reliability was .76). The correlational analyses between the SRT, the three procedural memory learning ability tasks, and the three declarative memory learning ability tasks administered in the main study revealed a significant relationship between the SRT and the ASRT (Spearman's $\rho = .38, p = .03$). No other correlations were statistically significant. These findings serve as preliminary evidence for convergent validity between the ASRT and the SRT and discriminant validity between the SRT and the tasks of declarative memory learning ability. Thus, we tentatively suggest that procedural memory learning ability may drive the relationship between the SRT and ASRT, although we acknowledge the existence of alternative interpretations. More research into the reliability and validity of the SRT as a task of procedural memory learning ability is needed. Such research will also be quite informative to work with implicit learning, as the SRT is a task that is commonly used in that line of research as well.

A second limitation of our study concerns the lack of L2 data in the present work. In future research, it will be important to examine the relative predictive validity of these tasks, including the SRT, for L2. Such work is important because it would be useful to know if some of these procedural memory learning ability tasks are more strongly associated with L2 learning than others. Future work should also consider test-retest reliability for the tasks to establish the stability of the measure over time. Finally, a further limitation of our study is that we examined the procedural (and declarative) memory learning ability construct using an EFA with a relatively small cognitive battery of tests, although the tests that were included were specifically motivated by the L2 literature. More correlated tests within each domain might have allowed for the constructs to emerge if the underlying connection to the construct exists but is weak.

CONCLUSION

In conclusion, we investigated the reliability and validity of procedural memory learning ability tasks that have been used to study L2 acquisition. We found evidence of reliability, except for the ASRT pattern versus random difference measure. Convergent validity was not evidenced in our results, suggesting that these measures do not jointly capture a broad construct of procedural memory. Discriminant validity was evidenced for the ASRT and TOL, but not for the DT-WPT, which may engage declarative memory learning abilities. In interpreting our results in light of prior research, we suggest that, although each task seems to recruit procedural memory learning abilities to some extent, the ASRT may provide a relatively process-pure measure, although different measures of the ASRT and other variants of sequence-learning tasks should be considered. Indeed, in a post-hoc follow-up study, we examined the SRT and found acceptable reliability for the task and a positive relationship between the ASRT and SRT. These results have implications both for future research that examines the role of procedural memory and may also have implications for research about implicit language aptitude and implicit learning more generally. However, future research on reliability and validity for the tasks used in these lines of research is needed.

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit <http://dx.doi.org/10.1017/S0272263121000127>.

NOTES

¹The use of the term *procedural memory* has not been defined consistently. For example, it has been used more broadly regarding learning and memory that is nondeclarative, i.e., “revealed in the absence of conscious recollection or verbal reflection” (Eichenbaum et al., 1994, p. 457). It has also been used more specifically to refer to “learning and memory that relies on the basal ganglia (BG) and associated circuitry” (Ullman et al., 2020, p. 391), which does not comprise all types of nondeclarative learning and memory. We attempt to characterize procedural memory in a way that is largely consistent with the different conceptualizations of the term, such that our characterization should represent the overlapping parts of a Venn diagram with circles representing various uses of *procedural memory*.

²To our knowledge, the original study to examine the role of declarative and procedural memory in L2 was Carpenter (Carpenter, 2008; Carpenter et al., 2009). Motivated by this work, Morgan-Short et al. (2014) adopted the tasks and measures used in Carpenter (the DT-WPT for procedural memory; Modern Language Aptitude Test, part V and Continuous Visual Monitoring Task for declarative memory). A second procedural memory task, the TOL, was adopted from Ettlenger et al. (2014).

³Note that some empirical research not included in Table 1 may also be relevant to the role of procedural memory in L2 (e.g., Granena, 2013; Suzuki & DeKeyser, 2017). For example, Granena (2013) examines individual differences in sequence learning ability and L2. To the extent that sequence learning ability as measured by the Serial Reaction Time task is subserved by procedural memory, the results are highly relevant to the question of the role of procedural memory in L2. However, we do not include these studies in our review because we specifically considered tasks used in studies with the explicitly stated purpose of examining procedural memory in L2.

⁴We do not consider the Serial Reaction Time task (Nissen & Bullemer, 1987) in this review as this task has only been used fully in one previous study with the explicitly stated purpose of examining the role of procedural memory in L2 (Hamrick, 2015).

⁵These results hold for learning on up to 50 trials of the WPT. Because this study employed a single-task version of the WPT, with extended training (100–150 trials) declarative memory can support learning and Parkinson's disease patients approached the performance level of healthy control participants after extended training.

⁶Because the assessments are all rather different from each other, practice effects are not expected, so the order of assessments was primarily intended to balance out potential fatigue effects. The additional constraint of nonconsecutive procedural or declarative memory assessments was included to ensure that participants were not fatigued from using one type of memory for a long period.

⁷EFA was the planned analysis because as a first pass we wanted to see what structure the data might reveal. Alternative analyses such as CFA might be more appropriate for testing how well our observed data fits a theoretical model of declarative and procedural memory. However, we chose not to report a CFA as the primary analysis because after seeing the correlation matrix it was clear that two latent structures were not present, which could lead to model convergence failures. Also, CFA has a more involved set of theoretical assumptions that this dataset cannot meet (see Bollen, 2002). We did however run a CFA and we had to remove the WPT task to make the model converge (also using weighted least squares approach because the matrix was not positive definite). We found poor fit, CFI = .86, RMSEA = .114.

⁸Interestingly, the ASRT, TOL, and the DT-WPT also did not correlate with performance on the Raven's Advanced Progressive Matrices, $rs < .17$, $ps > .10$ (see Supplementary Materials for information about this task). The three measures of declarative memory learning abilities did correlate with performance on the Raven's ($rs > .27$, $ps < .006$).

REFERENCES

- Antoniou, M., Ettlenger, M., & Wong, P. C. M. (2016). Complexity, training paradigm design, and the contribution of memory subsystems to grammar learning. *PLoS One*, *11*, Article e0158812. <https://doi.org/10.1371/journal.pone.0158812>
- 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>
- Beauchamp, M. H., Dagher, A., Aston, J., & Doyon, J. (2003). Dynamic functional changes associated with cognitive skill learning of an adapted version of the Tower of London task. *NeuroImage*, *20*, 1649–1660.
- Beauchamp, M. H., Dagher, A., Panisset, M., & Doyon, J. (2008). Neural substrates of cognitive skill learning in Parkinson's disease. *Brain and Cognition*, *68*, 134–143.
- Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *53*, 605–634. <https://doi.org/10.1146/annurev.psych.53.100901.135239>
- Brill-Schuetz, K. & Morgan-Short, K. (2014). The role of procedural memory in adult second language acquisition. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*, 260–265. <https://escholarship.org/content/qt0dc7958r/qt0dc7958r.pdf>

- Buffington, J., & Morgan-Short, K. (2018). Construct validity of procedural memory tasks used in adult-learned language. *Proceedings of the Annual Conference of the Cognitive Science Society*, 40, 1422–1427. <https://cogsci.mindmodeling.org/2018/papers/0276/0276.pdf>
- Buffington, J., & Morgan-Short, K. (2019). Declarative and procedural memory as individual differences in second language aptitude. In Z. E. Wen, P. Skehan, A. Biedroń, S. Li & R. L. Sparks (Eds.), *Language aptitude: Advancing theory, testing, research and practice* (pp. 215–237). Routledge.
- Carpenter, H., Morgan-Short, K., & Ullman, M. T. (2009). *Predicting L2 using declarative and procedural memory assessments: A behavioral and ERP investigation*. Presented at the Georgetown University Round Table, Washington, DC.
- Carpenter, H. S. (2008). *A behavioral and electrophysiological investigation of different aptitudes for L2 grammar in learners equated for proficiency level* [Unpublished doctoral dissertation]. Georgetown University.
- Carroll, J. B., & Sapon, S. M. (1959). *Modern language aptitude test*. Psychological Corporation.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159.
- Csabi, E., Benedek, P., Janacsek, K., Zavecz, Z., Katona, G., & Nemeth, D. (2016). Declarative and non-declarative memory consolidation in children with sleep disorder. *Frontiers in Human Neuroscience*, 9, Article 709. <https://doi.org/10.3389/fnhum.2015.00709>
- Dagher, A., Owen, A. M., Boecker, H., & Brooks, D. J. (2001). The role of the striatum and hippocampus in planning: A PET activation study in Parkinson's disease. *Brain*, 124, 1020–1032.
- DeKeyser, R. M. (2020). Skill acquisition theory. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition* (3rd ed., pp. 83–104). Routledge.
- Eichenbaum, H. (2012). *The cognitive neuroscience of memory: An introduction* (2nd ed.). Oxford University Press.
- Eichenbaum, H., Otto, T., & Cohen, N. J. (1994). Two functional components of the hippocampal memory system. *Behavioral and Brain Sciences*, 17, 449–472. <https://doi.org/10.1017/S0140525X00035391>
- Ettlinger, M., Bradlow, A. R., & Wong, P. C. M. (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>
- Foerde, K., Knowlton, B. J., & Poldrack, R. A. (2006). Modulation of competing memory systems by distraction. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 11778–11783. <https://doi.org/10.1073/pnas.0602659103>
- Foerde, K., Poldrack, R. A., & Knowlton, B. J. (2007). Secondary-task effects on classification learning. *Memory & Cognition*, 35, 864–874. <https://doi.org/10.3758/BF03193461>
- Gabrieli, J. D. E. (1998). Cognitive neuroscience of human memory. *Annual Review of Psychology*, 49, 87–115. <https://doi.org/10.1146/annurev.psych.49.1.87>
- Gebauer, G. F., & Mackintosh, N. J. (2007). Psychometric intelligence dissociates implicit and explicit learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 34–54. <https://doi.org/10.1037/0278-7393.33.1.34>
- Gluck, M. A., Shohamy, D., & Myers, C. (2002). How do people solve the “weather prediction” task? Individual variability in strategies for probabilistic category learning. *Learning & Memory*, 9, 408–418. <https://doi.org/10.1101/lm.45202>
- Godfroid, A., & Kim, K. M. (2021). The contributions of implicit-statistical learning aptitude to implicit second-language knowledge. *Studies in Second Language Acquisition*. Advance online publication. <https://doi.org/10.1017/s0272263121000085>
- 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. (2020). *Implicit language aptitude*. Cambridge University Press. <https://doi.org/10.1017/9781108625616>
- 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>
- Hamrick, P., Lum, J. A. G., & Ullman, M. T. (2018). Child first language and adult second language are both tied to general-purpose learning systems. *Proceedings of the National Academy of Sciences of the United States of America*, 115, 1487–1492. <https://doi.org/10.1073/pnas.1713975115>

- Hedenius, M., Ullman, M. T., Alm, P., Jennische, M., & Persson, J. (2013). Enhanced recognition memory after incidental encoding in children with developmental dyslexia. *PLoS One*, *8*, Article e63998. <https://doi.org/10.1371/journal.pone.0063998>
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185. <https://doi.org/10.1007/BF02289447>
- 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>
- Howard, J. H., Jr., & Howard, D. V. (2013). Aging mind and brain: Is implicit learning spared in healthy aging? *Frontiers in Psychology*, *4*, Article 817. <https://doi.org/10.3389/fpsyg.2013.00817>
- Howard, J. H., Jr., Howard, D. V., Dennis, N. A., & Kelly, A. J. (2008). Implicit learning of predictive relationships in three-element visual sequences by young and old adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1139–1157. <https://doi.org/10.1037/a0012797>
- Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction*, *32*, 51–62.
- Janacek, K., Shattuck, K. F., Tagarelli, K. M., Lum, J. A. G., Turkeltaub, P. E., & Ullman, M. T. (2020). Sequence learning in the human brain: A functional neuroanatomical meta-analysis of serial reaction time studies. *NeuroImage*, *207*, Article 116387. <https://doi.org/10.1016/j.neuroimage.2019.116387>
- 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.
- Kaller, C. P., Unterrainer, J. M., & Stahl, C. (2012). Assessing planning ability with the Tower of London task: Psychometric properties of a structurally balanced problem set. *Psychological Assessment*, *24*, 46–53. <https://doi.org/10.1037/a0025174>
- Kalra, P. B., Gabrieli, J. D. E., & 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>
- Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science*, *273*, 1399–1402. <https://doi.org/10.1126/science.273.5280.1399>
- Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning & Memory*, *1*, 106–120. <https://doi.org/10.1101/lm.1.2.106>
- Krus, D. J., & Helmstadter, G. C. (1993). The problem of negative reliabilities. *Educational and Psychological Measurement*, *53*, 643–650. <https://doi.org/10.1177/0013164493053003005>
- Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*, 202–220. <https://doi.org/10.1177/1094428105284919>
- Lum, J. A., Conti-Ramsden, G., Page, D., & Ullman, M. T. (2012). Working, declarative and procedural memory in specific language impairment. *Cortex*, *48*, 1138–1154. <https://doi.org/10.1016/j.cortex.2011.06.001>
- Middleton, F. A., & Strick, P. L. (2000). Basal ganglia and cerebellar loops: Motor and cognitive circuits. *Brain Research Reviews*, *31*, 236–250. [https://doi.org/10.1016/S0165-0173\(99\)00040-5](https://doi.org/10.1016/S0165-0173(99)00040-5)
- Morgan-Short, K., Deng, Z., Brill-Schuetz, K. A., Faretta-Stutenberg, M., Wong, P. C. M., & Wong, F. (2015). A view of the neural representation of second language syntax through artificial language learning under implicit contexts of exposure. *Studies in Second Language Acquisition*, *37*, 383–419. <https://doi.org/10.1017/S0272263115000030>
- Morgan-Short, K., Faretta-Stutenberg, M., Brill-Schuetz, K., Carpenter, H., & Wong, P. C. M. (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>
- Morgan-Short, K., & Ullman, M. T. (in press). Declarative and procedural memory in second language learning: Psycholinguistic considerations. In A. Godfroid & H. Hopp (Eds.), *The Routledge handbook of second language acquisition and psycholinguistics*. Routledge.
- Morling, B. (2015). *Research methods in psychology: Evaluating a world of information* (2nd ed.). W. W. Norton & Company.
- Nemeth, D., Janacek, K., & Fiser, J. (2013). Age-dependent and coordinated shift in performance between implicit and explicit skill learning. *Frontiers in Computational Neuroscience*, *7*, Article 147. <https://doi.org/10.3389/fncom.2013.00147>
- Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, *19*, 1–32.

- Nunnally, J. C., & Bernstein, I. H. (1978). *Psychometric theory*. McGraw-Hill.
- Ouellet, M., 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.
- Owen, A. M., James, M., Leigh, P. N., Summers, B. A., Marsden, C. D., Quinn, N. a., Lange, K. W., & Robbins, T. W. (1992). Fronto-striatal cognitive deficits at different stages of Parkinson's disease. *Brain*, 115, 1727–1751.
- Paradis, M. (2009). *Declarative and procedural determinants of second languages*. John Benjamins Publishing Company.
- Pili-Moss, D., Brill-Schuetz, K. A., Faretta-Stutenberg, M., & Morgan-Short, K. (2020). Contributions of declarative and procedural memory to accuracy and automatization during second language practice. *Bilingualism: Language and Cognition*, 23, 639–651. <https://doi.org/10.1017/S1366728919000543>
- Raiche, G., & Magis, D. (2020). *nFactors: Parallel analysis and other nongraphical solutions to the Cattell scree test*.
- Raiche, G., Riopel, M., & Blais, J. G. (2006). *Nongraphical solutions for the Cattell's scree test*. Paper presented at the annual meeting of the Psychometric Society.
- R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- Reber, P. J. (2013). The neural basis of implicit learning and memory: A review of neuropsychological and neuroimaging research. *Neuropsychologia*, 51, 2026–2042. <https://doi.org/10.1016/j.neuropsychologia.2013.06.019>
- Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 298, 199–209. <https://doi.org/10.1098/rstb.1982.0082>
- Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, 81, 105–120. <https://doi.org/10.1016/j.jml.2015.02.001>
- Song, S., Howard, J. H., Jr., & Howard, D. V. (2007). Implicit probabilistic sequence learning is independent of explicit awareness. *Learning and Memory*, 14, 167–176. <https://doi.org/10.1101/lm.437407>
- Squire, L. R. (1994). Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. In D. Schacter & E. Tulving (Eds.), *Memory systems 1994* (pp. 203–231). MIT Press.
- Squire, L. R., & Dede, A. J. O. (2015). Conscious and unconscious memory systems. *Cold Spring Harbor Perspectives in Biology*, 7, Article a021667. <https://doi.org/10.1101/cshperspect.a021667>
- Squire, L. R., Hamann, S., & Knowlton, B. J. (1994). Dissociable learning and memory systems of the brain. *Behavioral and Brain Sciences*, 17, 422–423. <https://doi.org/10.1017/S0140525X00035330>
- Stark-Inbar, A., Raza, M., Taylor, J. A., & Ivry, R. B. (2017). Individual differences in implicit motor learning: Task specificity in sensorimotor adaptation and sequence learning. *Journal of Neurophysiology*, 117, 412–428. <https://doi.org/10.1152/jn.01141.2015>
- Suzuki, Y. (2018). The role of procedural learning ability in automatization of L2 morphology under different learning schedules. *Studies in Second Language Acquisition*, 40, 923–937. <https://doi.org/10.1017/S0272263117000249>
- Suzuki, Y., & DeKeyser, R. M. (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>
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.
- 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>
- Trafimow, D. (2015). A defense against the alleged unreliability of difference scores. *Cogent Mathematics*, 2, Article 1064626. <https://doi.org/10.1080/23311835.2015.1064626>
- Trahan, D. E., & Larrabee, G. J. (1988). *Continuous visual memory test*. Psychological Assessment Resources.
- Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, 92, 231–270. <https://doi.org/10.1016/j.cognition.2003.10.008>
- Ullman, M. T. (2016). The declarative/procedural model: A neurobiological model of language learning, knowledge, and use. In G. Hickok, & S. L. Small (Eds.), *Neurobiology of language* (pp. 953–968). Elsevier Inc. <https://doi.org/10.1016/B978-0-12-407794-2.00076-6>

- Ullman, M. T. (2020). The declarative/procedural model. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition* (3rd ed., pp. 128–161). Routledge. <https://doi.org/10.4324/9780429503986-7>
- Ullman, M. T., Earle, F. S., Walenski, M., & Janacek, K. (2020). The neurocognition of developmental disorders of language. *Annual Review of Psychology*, *71*, 389–417. <https://doi.org/10.1146/annurev-psych-122216-011555>
- Unterrainer, J. M., Rahm, B., Kaller, C. P., Wild, P. S., Münzel, T., Blettner, M., Lackner, K., Pfeiffer, N., & Beutel, M. E. (2019). Assessing planning ability across the adult life span in a large population-representative sample: Reliability estimates and normative data for the Tower of London (TOL-F) task. *Journal of the International Neuropsychological Society*, *25*, 520–529. <https://doi.org/10.1017/S1355617718001248>
- Unterrainer, J. M., Rahm, B., Leonhart, R., Ruff, C. C., & Halsband, U. (2003). The Tower of London: The impact of instructions, cueing, and learning on planning abilities. *Cognitive Brain Research*, *17*, 675–683. [https://doi.org/10.1016/S0926-6410\(03\)00191-5](https://doi.org/10.1016/S0926-6410(03)00191-5)
- Van den Heuvel, O. A., Veltman, D. J., Groenewegen, H. J., Cath, D. C., van Balkom, A. J. L. M., van Hartkamp, J., Barkhof, F., & van Dyck, R. (2005). Frontal-striatal dysfunction during planning in obsessive-compulsive disorder. *Archives of General Psychiatry*, *62*, 301–310. <https://doi.org/10.1001/archpsyc.62.3.301>
- Willingham, D. B., Salidis, J., & Gabrieli, J. D. E. (2002). Direct comparison of neural systems mediating conscious and unconscious skill learning. *Journal of Neurophysiology*, *88*, 1451–1460. <https://doi.org/10.1152/jn.2002.88.3.1451>
- Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, *9*, 79–94.