**Supplemental Methods**

**Participants**

Participants (*N* = 119) were recruited through the Psychology Subject Pool at the University of Illinois at Chicago. All participants received course credit for completing the study. 20 participants were excluded for the following reasons: (a) 12 participants did not complete the second session, (b) two participants reported drug use, (c) three participants did not follow task instructions, (d) one participant’s data was not saved properly, and (e) two participants were outliers on one of the declarative memory tasks (see Analysis subsection for details). This resulted in *N* = 99 participants (58 women, 41 men, average age = 19.30 years, age range = 17 – 29 years) included in analysis.

***Power Analysis***

We conducted a power analysis to ensure that we have an adequate number of participants in the study. Specifically, we were interested in establishing sufficient power to observe convergent and discriminant validity. The power analysis was conducted using a Monte Carlo simulation procedure. Data was randomly generated from a hypothetical population correlation matrix where procedural memory tasks correlated with each other at a specified population value, and none of the procedural memory tasks correlated with declarative memory tasks (ρ= 0).1 Power was estimated from 10,000 simulations, where each simulation involved randomly generating data from the hypothetical correlation matrix, calculating an exploratory factor analysis, and deciding on model acceptability. We calculated power based on ρ = .40 for the procedural memory learning ability tasks. We chose this value for two reasons: 1) It was similar to the sizes of the correlations among some of our declarative memory tasks in a subset of the data analyzed prior to finishing data collection (*N* = 45 participants); 2) According to Cohen’s (1992) guidelines, ρ = .40 is a medium-to-large effect size. We determined that we could feasibly run a minimum of 80 participants. At ρ = .40, with 80 participants we determined that we would have 87% power, which was above a commonly-used threshold of 80%. In order to test at least 80 participants, we conducted this experiment over four semesters and had tested 119 by the end of the fourth semester.

**Materials**

***Procedural Memory Learning Ability***

Participants completed three assessments of procedural memory learning ability (Figure 1). The first assessment, the ASRT (Csabi et al., 2016), was based on the original task from J. H. Howard, Jr. and Howard (1997) and presented participants with a sequence learning task in which an item of the sequence consisted of a circle in a row of four circles being filled in by a dog’s head. The sequence in which the dog’s head appeared was a second-order pattern where patterned trials alternated with random trials. As such, participants might have seen a repeating sequence such as 3r1r4r2r, where the numbers correspond to the location of the dog’s head in the row of circles and ‘r’ represents a random location from one to four. The task instructed participants to press a key corresponding to the location of the dog’s head as quickly and accurately as possible. Participants pressed either the *z*, *c*, *b*, or *m* keys on a QWERTY keyboard (using their left and right middle and index fingers, respectively), with *z* corresponding to the leftmost circle position and *m* to the rightmost position. The task consisted of 20 blocks, with 85 trials per block (5 random trials followed by 80 alternating patterned and random trials, i.e., 3r1r4r2r, which consists of eight trials). Original research with this task provided evidence that participants were unable to describe the alternating pattern, suggesting the lack of explicit declarative knowledge (Howard, J. H., Jr. & Howard, 1997), so no data on possible explicit knowledge acquired during the task is reported here. Our measure of performance on the ASRT, as well as for the other two procedural memory learning ability tasks, was based on previous research that aimed to examine the role of procedural memory in L2. Thus, although alternative measures of performance are available for these tasks (as mentioned in the literature review), we only included measures that have been used in previous L2 work (see Table 1). Thus, performance on the ASRT was calculated following Faretta-Stutenberg and Morgan-Short (2018): The average reaction time on patterned trials was subtracted from the average reaction time on random trials, with a greater difference reflecting greater procedural learning.

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). The task assessed knowledge of probabilistic weather outcomes associated with combinations of cue cards. A combination of between 1 and 3 cue cards was presented all at once and that particular combination was associated with a certain probability of a fictional outcome of sunshine or rain. For example, a combination of a card with circles and a card with squares may have been associated with an 80% chance of sunshine. Participants were instructed to predict the weather based on a cue combination, and then the correct response appeared on the screen. Since the cards were probabilistically associated with a certain outcome, procedural learning could support the gradual process of acquisition. In this dual-task version, participants were also tasked with keeping track of the number of high tones that occurred during each trial. The high tones were pseudo-randomly interspersed with low tones, but participants were instructed to only count the high tones. This secondary task has been shown to impede the use of declarative memory on the weather prediction component of the task (Foerde et al., 2006). Thus, each trial consisted of making a weather prediction and counting the number of high tones. The task included a total of 320 dual-task trials divided into 8blocks, with 40 blocks per trial.2 Following previous L2 research (e.g., Faretta-Stutenberg & Morgan-Short, 2018), performance on the DT-WPT 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) for the probabilistic outcome associated with the cue combination. After completing the WPT, participants completed a debriefing measure of explicit knowledge. This task, called Cue Select, required participants to select a cue from the DT-WPT that was most likely to favor one of the weather outcomes (“sun”). Because Cue Select asks participants to explicitly consider how cues are associated with weather outcomes, performance on this task arguably reflects awareness of cue-weather associations on the task.

The final assessment of procedural memory was the Tower of London (Kaller et al., 2012). In this task, participants were presented with an initial configuration of colored circles that rested on pegs and were instructed to match a goal configuration. In producing the goal configuration, participants were constrained by being able to only move the topmost circle on each peg, and, when moved, the circle would fall to the lowest possible peg position. As such, participants were not able to move circles that had another circle on top of them, and they were not able to place circles in the middle of a peg that did not have any circles below the middle position. Participants were instructed to plan their sequence of moves before beginning the first move, and to do their best in completing the goal configuration in the stated number of moves, which began at three moves and increased to six moves by the end of the task. Participants completed one block each of four 3-move trials, eight 4-move trials, eight 5-move trials, and eight 6-move trials. A “trial” consisted of matching a single configuration to a goal configuration, and a “block” is defined as a sequence of trials containing the same number of moves required to match the goal configuration. Following Antoniou, Ettlinger, and colleagues (Antoniou et al., 2016; Ettlinger et al., 2014), participants repeated this task immediately after completing it for the first time.

Two measures of performance on this task have been used in previous L2 empirical work and were thus examined in our study. The first measure (TOL ImpBlock) examines the average percent change in initial think time (the duration from the presentation of a trial to the first move) from the beginning to the end of a block, 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; Ettlinger et al., 2014).

***Declarative Memory Tasks***

Participants completed three assessments of declarative memory (Figure 1): a verbal learning segment (Part V) of the Modern Language Aptitude Test (Carroll & Sapon, 1959), the Continuous Visual Memory Test (Trahan & Larrabee, 1988), and a recognition test called Declearn (Hedenius et al., 2013; Lukács et al., 2017). Part V of the Modern Language Aptitude Test (MLAT-V) assessed participants’ knowledge of English translations of 24 pseudo-Kurdish words. Participants initially studied the word-pair list, which contains the pseudo-Kurdish words along with their English translation equivalents, for two minutes. This was followed by a two-minute practice session in which participants produced the English translation equivalents of the pseudo-Kurdish words while being allowed to reference the word-pair list. Following this, participants proceeded to the 4-minute assessment section that instructed them to select the correct English translation of the 24 pseudo-Kurdish words that they studied. The format of this assessment was a five-option multiple choice test with 24 questions (one question per word). In this section, participants were not allowed to reference the word-pair list. Accuracy on the assessment was the performance measure for this task.

The second assessment of declarative memory was the Continuous Visual Memory Test (CVMT). Participants were presented with a series of images and tasked with answering whether or not they had previously seen the image. The images consisted of abstract designs presented on a computer screen, and thus measured nonverbal declarative memory ability. There were a total of seven “old,” or repeated target images that were presented seven times, and these were randomly interspersed with 63 “new,” or non-repeated, images. The randomized order was constant for all participants. Participants indicated whether they thought the item was new or old by either right- or left-clicking on the computer mouse, respectively. Performance on the Continuous Visual Memory Test was measured by computing the *d’* score, which is a measure from signal detection theory that accounts for response biases on recognition tasks (e.g., a bias to answer affirmatively when asked if a stimulus was previously shown).

The last assessment of declarative memory, the Declearn task (Hedenius et al., 2013), presented participants with a series of verbalizable and non-verbalizable shapes and then tested their recognition memory of the shapes. Non-verbalizable shapes were piloted to confirm low nameability of these items (Lukács et al., 2017). Declearn reflects aspects of declarative memory due to the item-based nature of the recognition memory task. Participants completed two phases: incidental encoding and recognition. The incidental encoding phase presented a series of 32 verbalizable, or “real” shapes, and 32 non-verbalizable, or “made-up” shapes, for a total of 64 shapes. In this phase, participants were tasked with responding whether a shape was real or made-up. Since participants were not instructed to remember the shapes and were not told there would be an assessment, this phase involves incidental encoding. After this phase, participants took a 10-minute break before beginning the recognition phase. The recognition phase presented participants with the 64 shapes from the encoding phase plus 64 novel shapes, for a total of 128 shapes. Half of the new shapes were real and half were made-up, so the proportion of real and made-up shapes was identical to that of the encoding phase. In the recognition phase, participants indicated whether or not the item was “new” or “old,” with “new” meaning the shape was not seen in the encoding phase and “old” meaning that the shape was seen in the encoding phase. In both phases, items were presented in a fixed pseudo-randomized order, with no more than three consecutive real or made-up objects. Performance was measured by computing *d’* scores for the recognition phase.

**Procedure**

Participants completed the study over two testing sessions scheduled on separate days. Both Sessions 1 and 2 lasted approximately 1.5 to 2 hours each, for a total of 3.5 to 4 hours of testing per participant with informed consent administered in the first session and the subject pool debriefing administered in the second session. In order 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 Table S1 for order of assessments).3 Additionally, a procedural memory task was always administered as the first task of each session. For time management, the Tower of London procedural memory assessment was administered in a separate session from the other two procedural memory assessments and as such was not counterbalanced with these two assessments. Lastly, in the second session (a) a measure of intelligence (a shortened version of Raven’s Advanced Progressive Matrices, Buffington, 2018; Raven, 1965) was administered but is not analyzed here;4 and (b) a language background questionnaire, including the LEAP-Q (Marian et al., 2007) and additional lab-specific questions, always occurred as the last task of the second session, unless there was time to administer it at the end of the first session.

|  |
| --- |
| **Table S1***Order of assessments* |
| Testing order |
| 1 | 2 | 3 | 4 |
| Session 1 |
| DT-WPT | DT-WPT | ASRT | ASRT |
| MLAT | CVMT | MLAT | CVMT |
| ASRT | ASRT | DT-WPT | DT-WPT |
| CVMT | MLAT | CVMT | MLAT |
| Session 2 |
| TOL | TOL | TOL | TOL |
| RAPM | Declearn | RAPM | Declearn |
| Declearn | RAPM | Declearn | RAPM |
| LBQ | LBQ | LBQ | LBQ |
| *Note*. Shows the partially counterbalanced order of assessments for the study. DT-WPT = Weather Prediction Task, MLAT = Part V of the Modern Language Aptitude Test, ASRT = Alternating Serial Reaction Time, CVMT = Continuous Visual Memory Test, TOL = Tower of London, RAPM = Raven’s Advanced Progressive Matrices, LBQ = Language Background Questionnaire. |

**Analysis**

Before analyses were run, data was cleaned by items and participants. For the ASRT, data was first cleaned by removing reaction times shorter than 100ms and longer than three standard deviations above the participant’s average reaction time and inaccurate trials were removed (total of 8.85% 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) vs. 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. In order 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)5 on the Spearman correlation matrix to test if the tasks factor into the expected procedural and declarative memory factors.

To perform the exploratory factor analysis, assumptions including multivariate normality and sampling adequacy (i.e., sufficiently large relationships among variables in the correlation matrix, examined the Kaiser-Meyer-Olkin Test) were evaluated. Principal Axis Factoring extraction was chosen because this extraction method does not depend on multivariate normality (Howard, M. C., 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 (Howard, M. C., 2016; Yong & Pearce, 2013). To conduct the analyses we used R version 3.6.0 (R Core Team, 2019) with the following packages: (a) psych, version 1.9.12.31 (Revelle, 2020); (b) GPArotation (Bernaards & Jennrich, 2005); nFactors version 2.4.1 (Raiche & Magis, 2020); (c) tidyr (Wickham & Henry, 2020); (d) corx version 1.0.2 (Conigrave, 2019); (e) dplyr (Wickham et al., 2020); and (f) GGally (Schloerke et al., 2020).

**Learning Over Time in Procedural Memory Learning Ability Assessments**

**Figure S1**

*ASRT Performance Over Time*

**

*Note.* Shows the ASRT score by block. The Random – Pattern score is calculated by averaging the reaction times separately for pattern and random trials within each block, and then subtracting the pattern reaction time from the random one. Higher scores indicate faster pattern reaction times and hence presumably more learning. Reaction times were cleaned before calculating averages (see Analysis subsection of the Method). Error bars show the standard error of the mean Random-Pattern Score within each block.

**Figure S2**

*WPT Performance Over Time*

**

*Note.* Shows the mean percent accuracy on WPT within each block. Error bars show the standard error of the mean accuracy within each block.

|  |
| --- |
|  |

|  |
| --- |
| **Table S2** |
| *TOL Performance Over Time* |
| Administration | Trial Time |  % Change Planning Time |
|  | 3 | 4 | 5 | 6 |  3 | 4 | 5 | 6 |
| Time 1 | 12.45 (4.79) | 15.79 (4.89) | 27.18 (8.36) | 29.80 (7.46) |  -18.56 (75.37) | 17.59 (63.38) | -32.81 (130.23) | -12.83 (108.95) |
| Time 2 | 8.14 (3.12) | 12.53 (3.92) | 21.46 (6.69) | 26.05 (7.40) |  -51.06 (112.49) | -9.06 (88.63) | -45.39 (123.59) | -26.68 (103.68) |
| *Note*: Trial times are the mean trial time, with the standard deviation in parentheses. All times are in seconds. % Change Planning is the percent change in initial think time, or the time from the presentation of the trial until the first move, from the beginning to the end of the block. Standard deviations are shown in parentheses. Blocks are denoted by the numbers 3-6, with each number referring to the minimum number of moves needed to solve each problem in that block. The TOL ImpBlock variable is calculated from the % Change Planning values in Time 1. The TOL ImpNormed variable is calculated from the Trial Times in Time 2. |

**Preliminary Considerations of Serial Reaction Time Task**

Although we did not include the Serial Reaction Time task (SRT; Hamrick, 2015; Lum et al., 2012; based on the task originally reported in Nissen & Bullemer, 1987) in the main study, we did administer the SRT to a subset of participants who participated in a follow-up study. We report this data as a preliminary investigation of the reliability and validity of the SRT. First, we briefly review theory and evidence on the reliability and validity of this task. According to some researchers, procedural memory is defined as learning and memory that takes place in the basal ganglia (Ullman et al., 2020). The SRT appears to reliably involve the basal ganglia, as suggested by a recent neuroanatomical meta-analysis (Janacsek et al., 2020). This evidence, at least under this definition of procedural memory, validates the SRT as a measure of procedural memory. Furthermore, other researchers have suggested that the SRT is a measure of implicit learning (Granena, 2013; Granena, 2019; Granena & Yilmaz, 2019; Suzuki & DeKeyser, 2015), a construct that may be related to procedural memory, which is thought to involve only learning of implicit information (Ullman et al., 2020). Research from this implicit learning perspective has found a version of the SRT similar to the one in the present study to be reliable, with a split-half coefficient of .79 in two studies (Granena, 2019; Granena & Yilmaz, 2019). Thus, theory (Ullman et al., 2020), research with an arguably related construct (implicit learning), and neuroimaging data (Janacsek et al., 2020) point towards the reliability and validity of the SRT. However, these conclusions do not necessarily imply that the SRT evidences convergent validity with other behavioral tasks of procedural memory learning ability.

Some behavioral research has explored the validity of the SRT as a measure of procedural memory (Parshina et al., 2018). Parshina et al. found that performance on the SRT did not correlate with performance on the ASRT *(r* = -.18), suggesting a lack of convergent validity. Of note, Parshina et al. used a different measure for the ASRT than the one used in the present study. Specifically, Parshina et al. measured performance on high vs. low-frequency triplets, whereas our measure compares pattern trials to random trials. Thus, further investigation of the reliability and validity of the SRT is important in order to see if these results on the SRT and ASRT hold with a different measure for the ASRT, as well as with a different sample, and to investigate the SRT’s relationships with other tasks claimed to measure procedural memory (i.e., TOL and DT-WPT).

As a preliminary investigation, we administered the SRT to a subset of participants in this study who participated in a follow-up session that took place after the sessions reported previously. Our research questions are the same as those from the main analyses: Is the SRT reliable? And, does the SRT show convergent and discriminant validity, considered respectively as patterning positively with other procedural memory learning ability assessments and *not* patterning positively with declarative memory learning ability assessments?

In the SRT, participants (*N* = 33) pressed a button on a game controller that corresponded to the location of a smiley face on the computer screen (see Hamrick, 2015 for additional details on the design and data analysis for this task). The locations followed a 10-item repeating pattern, although participants were not told about the pattern and were only instructed that it was a motor speed task. Learning was measured by subtracting the mean reaction time on the final patterned block from the mean reaction time on a subsequent block of random trials (i.e., the rebound score). The final random block is intended as a control block to assess how much participants are slowed down by random trials after seeing several blocks of the patterned sequence. Prior to analysis, for each participant data was cleaned by removing inaccurate trials, trials with reaction times slower than 100ms, and trials with reaction times greater than three standard deviations above the participant’s mean reaction time (mean reaction times were calculated separately for the patterned and random block). This resulted in removing 5.98% of the data. After this step, reaction times were log-transformed, following Hamrick (2015). Thus, the rebound scores reported below are based on log-transformed reaction times.

Descriptively, the SRT appears to be normally distributed, as evaluated according to the procedure described in the Results. There was also a significant learning effect, with the 95% confidence interval for the rebound score exceeding zero (*M =* .11, *SD* = .09, 95% CI: 0.08-0.14). For our research question on reliability, we calculated a measure that is conceptually similar1 to Cronbach’s alpha because learning was measured at the end of the task (see Analysis section for this rationale). Spearman-Brown reliability was .76. Next, to examine convergent and discriminant validity we only report correlation values2 as this subset of the data did not support a factor analysis. Regarding convergent validity, the SRT was positively associated with the ASRT, Spearman’s ρ = .38, *p* = .03. The SRT did not correlate significantly with any other procedural memory learning ability tasks (Spearman’s ρs < .22, *p*s > .23). Regarding discriminant validity, the SRT was not associated with any of the declarative memory learning ability tasks (Spearman’s ρs < .19, *p*s > .31).

Overall, this preliminary analysis suggests that the SRT exhibits acceptable reliability, consistent with the reliability values reported by Granena (2019) and Granena & Yilmaz (2019). Further, abilities on the SRT were associated with abilities on the ASRT, suggesting convergence on an underlying mechanism.3 The SRT also did not correlate with any of the declarative memory learning ability tasks, suggesting discriminant validity, although the smaller sample size in this analysis has lower power. 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. One interpretation that we believe is unlikely, though, is that more basic motor speed processes drive this relationship, as abilities on both the SRT and ASRT were measured by subtracting out performance on random trials, which includes participants’ baseline response speeds. Thus, individual differences in response speed are arguably not a component of the scores for SRT or ASRT. Taken together, given prior research on the SRT, its association with the ASRT in the present study, and its greater reliability than the ASRT, we cautiously suggest that the SRT is a suitable measure for procedural memory learning ability.

**Notes**

1 ρ is used here in the sense of a population value for the correlation statistic, not in the sense of Spearman’s rho.

2 There are 8 blocks with tones, and the 9th block does not have any tones.

3 Since 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 in order to ensure that participants were not fatigued from using one type of memory for a long period of time.

4 We included the Raven’s task in order to run a bifactor analysis, in which we predicted that Raven’s, as a measure of general intelligence, would capture general variance shared by all tasks. However, after running this analysis it became clear that there was not a general factor in this dataset. Thus, we decided that the exploratory factor analysis (see Analysis subsection and Results section) was the more appropriate analysis.

5 The measure is not exactly Cronbach’s alpha because this statistic cannot be calculated on continuous data, such as the reaction time data on the SRT. As such, we approximated Cronbach’s alpha by randomly splitting trials into two halves (without replacement) and calculating the correlation coefficient separately for each split (1,000 splits total). We then used the Fisher’s r to z transformation to normalize the data and then average them. Finally, the average correlation was transformed back to an r value using Fisher’s z to r and this final r value was used for the Spearman-Brown split-half reliability.

6 One participant in the SRT dataset was excluded from the main analyses for being an outlier on a declarative memory learning ability task, and as such these correlations are based on N = 32 participants.

7 On one level, these results appear to be inconsistent with the results from Parshina et al. (2018). However, given that we used different measures for the ASRT, we avoid strongly interpreting any differences between their results and ours.

**References**

Antoniou, M., Ettlinger, M., & Wong, P. C. M. (2016). Complexity, training paradigm design, and the contribution of memory subsystems to grammar learning. *PloS One, 11*(7), e0158812. 10.1371/journal.pone.0158812

Bernaards, C. A., & Jennrich, R. I. (2005). Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement, 65*, 676-696.

Buffington, J. (2018). *The reliability and validity of procedural memory assessments used in second language learning* (M.A. Thesis). Available from Dissertations & Theses @ CIC Institutions, ProQuest Dissertations & Theses A&I. (2194271302).

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Conigrave, J. (2019). *Corx: Create and format correlation matrices.*

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.* 10.3389/fnhum.2015.00709

Ettlinger, M., Bradlow, A. R., & Wong, P. C. M. (2014). Variability in the learning of complex morphophonology. *Applied Psycholinguistics, 35*(4), 807-831. 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*(1), 67-101. 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*(31), 11778-11783. 10.1073/pnas.0602659103

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*(6), 408-418. 10.1101/lm.45202

Granena, G. (2013). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning, 63*(4), 665-703. 10.1111/lang.12018

Granena, G. (2019). Cognitive aptitudes and L2 speaking proficiency. *Studies in Second Language Acquisition, 41*(2), 313-336. 10.1017/S0272263118000256

Granena, G., & Yilmaz, Y. (2019). Corrective feedback and the role of implicit Sequence‐Learning ability in L2 online performance. *Language Learning, 69*(S1), 127-156. 10.1111/lang.12319

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

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*(5), e63998. 10.1371/journal.pone.0063998

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185. 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*(4), 634-656. 10.1037/0882-7974.12.4.634

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*(1), 51-62.

Janacsek, 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*, 116387. 10.1016/j.neuroimage.2019.116387

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*(1), 46-53. 10.1037/a0025174

Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning & Memory, 1*(2), 106-120. 10.1101/lm.1.2.106

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science, 273*(5280), 1399-1402. 10.1126/science.273.5280.1399

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*(2), 202-220. 10.1177/1094428105284919

Lukács, Á, Kemény, F., Lum, J. A. G., & Ullman, M. T. (2017). Learning and overnight retention in declarative memory in specific language impairment. *PloS One, 12*(1), e0169474.

Lum, J. A., Conti-Ramsden, G., Page, D., & Ullman, M. T. (2012). Working, declarative and procedural memory in specific language impairment. *Cortex, 48*(9), 1138-54. <https://doi.org/>10.1016/j.cortex.2011.06.001

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language & Hearing Research, 50*(4), 940-967. 10.1044/1092-4388(2007/067)

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*(1), 56-72. 10.1017/S1366728912000715

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology, 19*(1), 1-32.

Nunally, J. C., & Bernstein, I. H. (1978). *Psychometric theory*. New York: McGraw-Hill.

Parshina, O., Obeid, R., Che, E. S., Ricker, T. J., & Brooks, P. J. (2018). SRT and ASRT: Similar tasks tapping distinct learning mechanisms? Paper presented at the *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 2205-2210.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Raiche, G., & Magis, D. (2020). *nFactors: Parallel analysis and other non graphical 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.*

Raven, J. C. (1965). *Advanced progressive matrices, sets I and II*. H. K. Lewis. (Distributed in the United States by The Psychological Corporation).

Revelle, W. (2020). *Psych: Procedures for psychological, psychometric, and personality research.*

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., & Crowley, J. (2020). *GGally: Extension to 'ggplot2.'*

Suzuki, Y., & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language Learning, 65*(4), 860-895. 10.1111/lang.12138

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

Trahan, D. E., & Larrabee, G. J. (1988). *Continuous visual memory test*. Psychological Assessment Resources.

Ullman, M. T., Earle, F. S., Walenski, M., & Janacsek, K. (2020). The neurocognition of developmental disorders of language. *Annual Review of Psychology, 71*(1), 389-417. 10.1146/annurev-psych-122216-011555

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *Dplyr: A grammar of data manipulation*

Wickham, H., & Henry, L. (2020). *Tidyr: Tidy messy data.* https://tidyr.tidyverse.org/

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*(2), 79-94.