Retrieval Practice Promotes Learning of Turkish as a Foreign Language: A Computer-Assisted Language Learning Study

Maya C. Rose[1,2] (mrose4@gradcenter.cuny.edu)
Patricia J. Brooks[1,2] (patricia.brooks@csi.cuny.edu)
Arshia Lodhi[2] (arshia.lodhi@cix.csi.cuny.edu)
Angela Cortez[2] (angela.cortez@csi.cuny.edu)

[1]Department of Educational Psychology, The Graduate Center, City University of New York
[2]Department of Psychology, College of Staten Island, City University of New York

Abstract

Variation in second language acquisition is evident from earliest stages. This study examined effects of learning tasks (retrieval practice, comprehension, verbal repetition) on comprehension of Turkish as a new language. Undergraduates (N = 156) engaged with Turkish spoken dialogues in a computer-assisted language learning session via Zoom, with learning tasks manipulated between-subjects. Participants completed pre/posttests assessing comprehension of Turkish number and case marking, a vocabulary test, and open-response questions gauging explicit awareness. The retrieval-practice group showed highest performance overall, after controlling for significant effects of nonverbal ability and pretest. For comprehension of number/case marking, the comprehension group performed comparably to the retrieval-practice group. For vocabulary comprehension, the verbal-repetition group performed comparably to the retrieval-practice group. Differential performance associated with learning tasks indicates benefits of testing and production and aligns with transfer-appropriate processing. As predicted by the noticing hypothesis, explicit awareness of number and case marking correlated with comprehension accuracy.

Keywords: second language acquisition; Turkish; miniature language; testing effect; retrieval practice; noticing hypothesis

Introduction

Learning a new language is difficult for many individuals, yet manageable for others. Variation in second language (L2) learning outcomes is associated with input conditions and aptitude (Dörnyei, 2005; Granena et al., 2016). Miniature language learning paradigms provide a feasible method for studying individual differences in learners' grasp of linguistic patterns (e.g., gender agreement, case marking) at the earliest stages of learning (Kempe & Brooks, 2016), and may be implemented online via a computer-assisted language learning (CALL) protocol. The current study used a miniature version of Turkish, an agglutinative language featuring vowel harmony and allomorphic variation, to explore effects of testing and speech production on the acquisition of nominal morphology (number and case marking), vocabulary, and metalinguistic awareness.

Role of Testing and Production in L2 Acquisition

Research indicates that repeated testing in the form of recognition or recall-based tests enhances learning to a greater extent than simply restudying the information (Karpicke & Aue, 2015; Marsh et al., 2007). Even if learners struggle to retrieve the appropriate information from memory, testing can serve to consolidate memory (Karpicke & Roediger, 2008; Rowland, 2014). Using a miniature artificial language, Hopman and MacDonald (2018) observed that adults who practiced retrieving phrases from memory outperformed those who practiced matching pictures with corresponding phrases on a subsequent grammar comprehension test. Similar benefits of retrieval practice were observed in a German L2 classroom setting (Keppenne et al., 2021). In these L2 studies, the researchers did not include a condition where participants were asked to repeat phrases, as opposed to retrieve them from memory. Hence, the studies did not clearly distinguish benefits of testing via retrieval practice from benefits of overt production (i.e., verbal repetition). In the context of learning L2 vocabulary, researchers have started to disentangle effects of testing and production, with findings suggesting that retrieval practice promotes vocabulary learning more than verbal repetition practice (e.g., Akifumi, 2016; Kang et al., 2013). The current study extends this line of research to L2 grammar learning.

Retrieval practice may promote L2 acquisition by providing opportunities for learners to register discrepancies between what they have heard and what they can produce on their own. According to Schmidt's (1990) noticing hypothesis, attention is a necessary prerequisite for encoding L2 input. As the learner processes L2 input, they become explicitly aware of features of the language, leading them to engage in deeper processing. This, in turn, promotes retention and use of those features in the learner's own L2 productions (Leow, 2018). Swain and Lapkin (1995) proposed the output hypothesis, arguing that the act of producing L2 utterances may lead the learner to recognize problems, which may trigger further linguistic processing and noticing of L2 features. Both of these theories predict that metalinguistic awareness and L2 grammatical knowledge will develop in tandem. What is less clear is whether L2 production in the context of verbal repetition (as opposed to retrieval practice) is sufficient to trigger noticing.

Individual Differences in Aptitude

In a miniature language study, Brooks and Kempe (2013), found that language-learning aptitude correlated with explicit

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

awareness and accuracy in producing Russian gender agreement and case marking. Regression models indicated that the effect of aptitude on production accuracy was indirect and mediated by awareness. That is, learners with higher aptitude were more likely to notice patterns, which in turn predicted their accuracy in applying the patterns to their own productions. The current study sought to replicate this finding in relation to L2 comprehension after a single CALL session.

Language learning aptitude is a broad construct that accounts for individual differences in L2 outcomes in specified learning contexts (e.g., college classrooms; CALL studies). Aptitude has been assessed using measures of nonverbal ability (Brooks et al., 2017) and various indices of memory, including phonological short-term memory (Ellis, 1996), verbal working memory (Miyake & Friedman, 1998), and declarative and procedural memory (Hamrick, 2015). In a psychometric investigation of a foreign language aptitude test, Grigorenko et al. (2000) found that scores on the Culture Fair intelligence test (an indicator of nonverbal ability) loaded onto an intelligence-related factor, while the scores on the verbal declarative component of the Modern Language Aptitude Test (Carroll & Sapon, 1959) loaded onto a separate language-specific factor. Due to time constraints, the current study used only the Culture Fair test as a measure of aptitude.

Research Objectives and Hypotheses

Using a miniature version of Turkish embedded in a CALL protocol, we manipulated learning tasks in a between-subjects design to elucidate effects of testing and production at the outset of L2 learning. Drawing on research on retrieval practice and the testing effect, it was hypothesized that learners asked to retrieve the Turkish inflected nouns from memory would exhibit higher accuracy on comprehension tests as compared to learners instructed to verbally repeat the nouns. We also manipulated the modality of testing by including a group that engaged in comprehension practice, i.e., multiple-choice testing with no production component. In accordance with the noticing hypothesis, we expected to find a strong correlation between metalinguistic awareness and Turkish comprehension accuracy, with nonverbal ability (Culture Fair scores) associated with individual differences in learning outcomes. Given established relations between language background and L2 learning (Foucart & Frenck-Mestre, 2011), we included number of prior languages as a covariate in analyses.

Method

Participants

College students between the ages of 18 and 30 years were recruited from a psychology department subject pool at an open-admission Hispanic-serving public university. The students received research participation credit for completing the 2-hour session on Zoom. Students with prior knowledge of Turkish or Turkic languages (e.g., Uzbek) were excluded. The sample comprised 156 students (93 females, 61 males, 1 non-binary, 1 did not disclose), aged 18 to 28 years (*M* = 19.5, *SD* = 2.0). Race/ethnicity was self-reported as follows: 35.9% White, 24.4% Black/African American, 23.1% Hispanic/Latinx, 12.2% Middle Eastern, 10.3% Asian (categories were not mutually exclusive). Students were randomly assigned to CALL conditions: *comprehension* (*n* = 52*), verbal repetition* (*n* = 52), and *retrieval-practice* (*n* = 52).

Turkish Miniature Language Materials

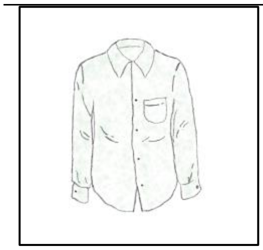
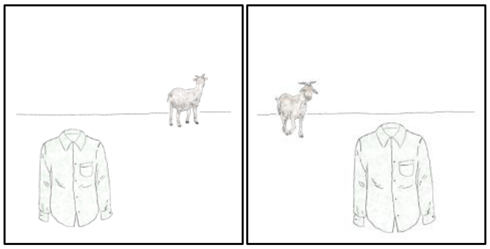
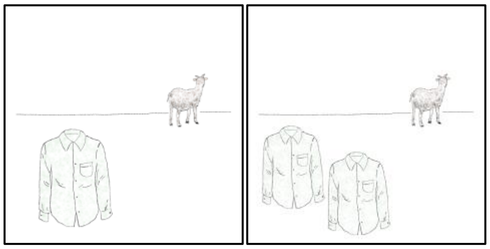
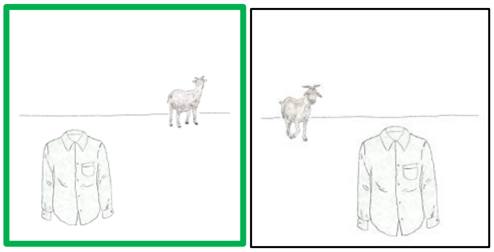
Noun Vocabulary and Dialogues. The Turkish vocabulary consisted of 36 nouns ending in *–ek*, *–ak*, or *–a* in the nominative case. Each noun appeared in four dialogues referencing a goat coming towards or going away from one or two objects. In each dialogue, a question was asked by a male speaker and answered by a female speaker. Table 1 presents the Turkish questions and representative answers, consisting of Turkish nouns inflected for case (dative [to] or ablative [from]) and number (singular or plural). Across dialogues, the subject (*keçi* [goat]) was held constant. Of the 144 question-answer dialogues (36 nouns x 2 case x 2 number): 36 were used in the pretest, training, and posttest; 72 were used for training only; and 36 items were used for posttest only. Each noun, case, and number occurred roughly equal numbers of times across pretest, training, and posttest trials. Dialogues were presented auditorily with pictures of the goat and corresponding object(s); see Table 2 for example trials for each CALL condition. Turkish orthography and English translations were not available at any point.

Table 1: Examples of Turkish question-answer dialogues with singular and plural nouns in ablative and dative case.

Ablative Question:	
Keçi nereden geliyor? [Goat where-from coming?]	
Ablative Answers:	
Singular Object	*gömlek*[1]*-ten* [shirt-ABL]
	bardak-tan [cup-ABL]
	araba-dan [car-ABL]
Plural Object	*gömlek-ler-den* [shirt-PL-ABL]
	bardak-lar-dan [cup-PL-ABL]
	araba-lar-dan [car-PL-ABL]
Dative Question:	
Keçi nereye gidiyor? [Goat where-to going?]	
Dative Answers:	
Singular Object	*gömle-ğe* [shirt-DAT]
	barda-ğa [cup-DAT]
	araba-ya [car-DAT]
Plural Object	*gömlek-ler-e* [shirt-PL-DAT]
	bardak-lar-a [cup-PL-DAT]
	araba-lar-a [car-PL-DAT]

Note: ABL = ablative [from], DAT = dative [to], PL = plural; [1] nouns end in *–ek*, *–ak*, or *–a* and exhibit allomorphic variation when inflected for case and number.

Table 2: Example trials for each CALL training condition.

Comprehension	**Verbal Repetition**	**Retrieval-Practice**
Instructions: Listen to the dialog and use the right or left arrow key to select the picture matching what the woman says.	Instructions: Listen to the dialog and repeat the woman's answer to the question.	Instructions: Listen to the question and answer it in Turkish aloud.
 Introduction: Female voice: *gömlek* [shirt]	 Introduction: Female voice: *gömlek* [shirt]	 Introduction: Female voice: *gömlek* [shirt]
 Case trial: Male voice: *Keçi nereden geliyor?* [Where is the goat coming from?] Female voice: *gömlekten* [from the shirt]  Number trial: Male voice: *Keçi nereden geliyor?* [Where is the goat coming from?] Female voice: *gömlekten* [from the shirt]	 Male Voice: *Keçi nereden geliyor?* [Where is the goat coming from?] Female voice: *gömlekten* [from the shirt]	 Male Voice: *Keçi nereden geliyor?* [Where is the goat coming from?]
 Feedback (case trial): Answer is replayed as correct picture is shown: Female voice: *gömlekten* [from the shirt]	 Feedback: Answer is replayed and participant repeats the Turkish inflected word a second time to advance to the next trial: Female voice: *gömlekten* [from the shirt]	 Feedback: Answer is played and participant repeats the correct Turkish inflected word to advance to the next trial: Female voice: *gömlekten* [from the shirt]

CALL Procedure

CALL tasks were programmed in PsychoPy and run online on the Pavlovia platform. Participants completed tasks in the following order: pretest, training, posttest, vocabulary test.

Pretest. The pretest consisted of one block of *comprehension* trials (18 case, 18 number trials); see Table 2.

Training. Training consisted of three blocks of 36 trials (evenly distributed over the four case/number combinations). Participants were presented with instructions at the start of each block; trials followed procedures shown in Table 2. Participants in the *comprehension* condition completed three blocks of comprehension trials. Participants in the *verbal repetition* condition completed three blocks of verbal repetition trials. Participants in *retrieval-practice* completed one block of verbal repetition trials and two blocks of retrieval-practice trials. Note that on comprehension and verbal repetition trials, participants heard the Turkish noun three times, while on retrieval-practice trials, they heard the noun twice and attempted to generate it once on their own.

Posttest. The posttest consisted of one block of *comprehension* trials (36 case, 36 number trials). Half of the trials were identical to the pretest ("old" items); the other half were reserved for the posttest ("new" items).

Vocabulary Test. Participants were asked to choose from sets of four pictures the one matching each Turkish noun (36 trials, no feedback) using designated keys. Nouns were presented auditorily in the nominative case; see Table 3.

Table 3: Example trial for CALL vocabulary test.

Instructions	Pictures Presented	
Choose the picture that matches the Turkish word: Q (top left) P (top right) A (bottom left) L (bottom right) Female voice: *gömlek* [shirt]		

Metalinguistic Awareness. Following the vocabulary test, participants completed a questionnaire (adapted from Brooks & Kempe, 2013) asking them to describe what they noticed about the Turkish words and how they indicated the direction of the goat and the number of objects. Responses were scored for awareness of case and number marking. For case, 0 indicated no awareness, 1 indicated mention of word endings (e.g., "The word for away sounded like *tan* and the word for moving towards sounded like *ya*."), and 2 indicated mention of allomorphic variation in the endings (e.g., "I noticed that an English sounding *a* was for moving toward, and an English sounding *tan* or *dan* was for moving away."). For number, 0 indicated no awareness, 1 indicated mention of the plural marker (e.g., "When there were two objects, a *lar* sound was included in the word"), and 2 indicated mention of allomorphic variation (e.g., "*Larra* or *lerre* meant it was plural"). Two trained assistants independently scored 60% of responses. Percent agreement was 91.2% for number (*kappa* = .82) and 85.3% for case (*kappa* = .78) After establishing reliability, the assistants coded remaining responses together.

Language Learning Aptitude Assessments

Language Background Questionnaire. Using a Qualtrics form, participants listed languages studied in school, spoken at home, or learned abroad. For each language, they reported proficiency in domains of reading, writing, listening, and speaking (6-point Likert scale; 1 = *very poor*, 6 = *excellent*); proficiency was calculated as the average of the four domains. Participants reported an average proficiency of 3.31 (*SD* = 1.47, *Range* = 0–6) in their best language other than English. Participants reported knowing an average of 2.7 languages (including English; *SD* = 0.8, *Range* = 1–5). Total languages varied across CALL conditions, $F(2, 153) = 3.60$, $p = .030$, $\eta_p^2 = .04$. Participants in the *comprehension* condition knew more languages than the *verbal repetition* condition ($M = 2.92$ vs. 2.52), $p = .024$. The *retrieval-practice* condition ($M = 2.77$) did not differ from other conditions, *p's* > .23. Due to the group difference, we included total languages as a control variable in all analyses.

Nonverbal Ability. A computerized version of Test 1 (Series) and Test 2 (Classification) of the Culture Fair Intelligence Test, Scale 3, Form A (Cattell & Cattell, 1973) was administered via Qualtrics. Series problems involved selecting an abstract geometric pattern from six alternatives to complete the series. Classification problems asked participants to identify which two of five patterns were alike. Problem difficulty increased as each test progressed. Participants were told to complete as many problems as possible in the allotted time (3 minutes for 13 problems in Test 1; 4 minutes for 14 problems in Test 2). Before each test, participants completed several example problems with feedback. Scores were calculated as the number of problems answered correctly across Tests 1 and 2 ($M = 11.6$, $SD = 2.8$, Range = 4–18). Scores did not differ across CALL conditions, $F(2, 153) = 0.85$, $p = .43$, $\eta_p^2 = .01$).

Zoom Recordings

At the start of the Zoom session, the participant shared their screen. The research assistant (RA) confirmed that no other programs were running and started recording the session. The RA remained present to ensure that participants completed tasks as instructed.

Table 4: Mean scores (percentage correct) for each language learning outcome for the full sample and CALL conditions.

	Full Sample (*N* = 156)		Comprehension (*n* = 52)	Verbal Repetition (*n* = 52)	Retrieval-Practice (*n* = 52)
Pretest	*M* (SD)	*Range*	*M* (SD)	*M* (SD)	*M* (SD)
Total score	56.7% (12.2)	27.8–94.4%	57.3% (14.0)	57.2% (12.6)	55.7% (9.6)
Case trials	58.0% (16.2)	27.8-100%	60.3% (17.4)	58.5% (17.2)	55.1% (13.8)
Number trials	55.4% (14.1)	16.7-94.4%	54.3% (15.3)	55.8% (12.8)	56.2% (14.2)
Posttest					
Total score	80.7% (17.9)	40.3-100%	83.5% (17.2)	74.4% (18.7)	84.3% (16.2)
Case trials (old)	82.8% (20.0)	27.8-100%	86.9% (18.9)	75.5% (22.0)	86.0% (17.2)
Case trials (new)	82.2% (20.4)	27.8-100%	85.9% (18.9)	77.4% (20.4)	83.2% (21.2)
Number trials (old)	78.4% (19.8)	22.2-100%	80.8% (18.8)	70.9% (21.5)	83.4% (17.0)
Number trials (new)	79.6% (21.2)	33.3-100%	80.4% (22.2)	73.7% (21.6)	84.7% (18.7)
Vocabulary Test	66.0% (19.8)	11.1-100%	54.6% (19.3)	71.1% (18.3)	72.2% (16.7)
Metalinguistic Awareness	1.60 (1.17)	0–4	1.65 (1.12)	1.37 (1.17)	1.79 (1.21)

Results

Descriptive statistics for the language learning outcomes are shown in Table 4. All outcomes were correlated suggesting stable individual differences; see Table 5. Performance on case/number and old/new trials was similar. Consequently, for brevity, we report analyses conducted on total scores at pretest and posttest.

Table 5: Correlations across outcome measures (*N* = 156).

	Pretest	Posttest	Vocabulary
Pretest			
Posttest	.42*		
Vocabulary	.23*	.32*	
Metalinguistic awareness	.29*	.57*	.25*

*Bonferroni corrected α = .00833

Analysis of Covariance Analyses

We used ANCOVAs to examine effects of CALL conditions and covariates on each outcome measure; note that assumptions of normality were met. All proportions were arcsine transformed; analyses of raw scores yielded nearly identical results. We first examined performance on the pretest assessing comprehension of Turkish case and number marking prior to the CALL training blocks. Note that each pretest comprehension trial included feedback, allowing participants to learn as trials progressed. At pretest, CALL conditions did not differ, $F(2, 151) = 0.62$, $p = .538$; $\eta_p^2 = .01$. Individual differences at pretest were associated with nonverbal ability (Culture Fair scores), $F(1, 151) = 7.97$, $p = .005$, $\eta_p^2 = .05$.

Next, we examined performance on the vocabulary test; see Table 6 for ANCOVA results. Performance varied significantly by CALL condition. Post-hoc tests indicated higher accuracy in the *retrieval-practice* and *verbal repetition* conditions than in the *comprehension* condition; see Table 4 for mean scores for each condition. Individual differences in vocabulary test scores were associated with two of the covariates: nonverbal ability (Culture Fair scores) and pretest total scores.

Table 6: ANCOVA predicting vocabulary scores.

Variable	df	F (η_p^2)
CALL condition	(2, 150)	15.24*** (.17)
Pretest total score	(1, 150)	8.14** (.05)
Culture Fair score	(1, 150)	6.56** (.04)
Total languages	(1, 150)	0.24 (.00)
Overall model	(5, 150)	10.42*** (.26)

****p* < .001, ***p* < .01

We then examined performance on the posttest assessing comprehension of Turkish case and number marking after CALL training; see Table 7 for ANCOVA results. Posttest accuracy varied by CALL condition. Post-hoc tests indicated higher scores in the *retrieval-practice* and *comprehension* conditions than in the *verbal repetition* condition; see Table 4 for mean scores for each condition. Individual differences in posttest accuracy were associated with three of the covariates: nonverbal ability (Culture Fair scores), pretest total scores, and vocabulary test scores, but not with language background (total number of languages).

The last ANCOVA model examined metalinguistic awareness (i.e., total scores for case and number awareness, *Range* = 0 to 4); see Table 7 for results. The pattern of results matched what was observed for the posttest assessing comprehension of Turkish case and number marking, though the effect of CALL condition became significant only after controlling for covariates. Variation in metalinguistic awareness was associated with nonverbal ability, pretest scores (Culture Fair), and vocabulary test scores, but not with language background (total languages).

Table 7: ANCOVA predicting posttest scores (case and number trials combined) and metalinguistic awareness.

Variable	df	Posttest Case/Number F (η_p^2)	Metalinguistic Awareness F (η_p^2)
CALL condition	2, 149	9.26*** (.11)	3.46* (.04)
Pretest total score	1, 149	23.46*** (.14)	14.06*** (.09)
Vocabulary score	1, 149	10.74** (.07)	9.07** (.06)
Culture Fair score	1, 149	10.05** (.06)	14.80*** (.09)
Total languages	1, 149	3.33 (.02)	1.40 (.01)
Overall model	6, 149	15.56*** (.39)	11.81*** (.32)

***p < .001, **p < .01, *p < .05

Mediation Analysis

Following Brooks and Kempe (2013), we used the PROCESS macro (Hayes, 2017) to find out whether the effect of nonverbal ability (Culture Fair) on posttest accuracy was indirect, i.e., mediated by metalinguistic awareness. CALL condition was entered as a moderator and pretest total scores and total languages were entered as covariates in the analysis. Nonverbal ability predicted metalinguistic awareness, t(148) = 2.67, p = .008, and metalinguistic awareness predicted posttest accuracy, t(151) = 9.88, p < .001. The direct effect of nonverbal ability on posttest accuracy was not significant, t(151) = 1.03, p = .307. Instead, there was a significant indirect effect of nonverbal ability on posttest accuracy that was mediated by metalinguistic awareness, B = 0.04 (bootstrap SE .01), p < .001, percentile bootstrap 95% CI [0.02, 0.05]. CALL condition did not moderate these effects. Figure 1 depicts the mediation model with coefficients for each path.

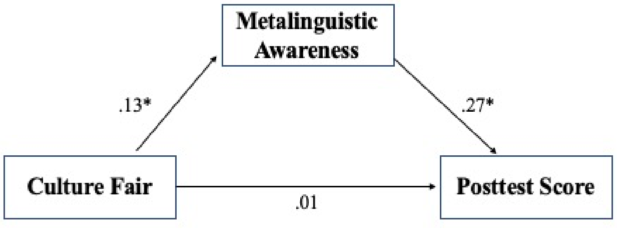


Figure 1: Mediation model illustrating relationship between nonverbal ability (Culture Fair), metalinguistic awareness, and posttest score; *p < .001.

Discussion

This study examined the impact of CALL learning tasks on comprehension of Turkish as a new language. Findings revealed differential effects of testing and production. Overall, participants who engaged in retrieval practice exhibited better comprehension than their counterparts in the other CALL conditions, despite having heard the Turkish nouns fewer times during training (i.e., when attempting to retrieve the words from memory). Notably, the *retrieval-practice* condition was superior to the *verbal repetition* condition on the posttest assessing comprehension of number and case marking, indicating that simply repeating the Turkish inflected words was not as effective as retrieval practice in triggering noticing of the grammatical markers. On the other hand, the *comprehension* condition performed similarly to the *retrieval-practice* condition on the posttest. This contrasts with Hopman and MacDonald (2018) and Keppenne et al. (2021) who found better performance with production-based tests and suggests that either recall- or recognition-based tests may promote L2 grammar learning. The strong performance of the comprehension condition in the current study might also be attributed to transfer appropriate processing (Morris et al., 1977), as the posttest was identical in format to the comprehension training.

On the vocabulary test, the *retrieval-practice* and *verbal repetition* conditions performed comparably (and superior to the *comprehension* condition). This suggests that overt production rather than testing served to strengthen lexical representations, adding to other research demonstrating benefits of production for L2 learning (Dahlen & Caldwell-Harris, 2013; Forrin et al., 2012). Together, the results suggest separable roles for testing and production in strengthening linguistic representations. Future multi-session studies should examine whether benefits of testing and production on L2 learning persist over time.

Metalinguistic awareness correlated moderately (r = .57) with posttest accuracy—in line with the noticing hypothesis that adult L2 learning depends on awareness (Schmidt, 1990). Yet awareness was rather limited (M score = 1.6 out of 4): Participants almost never used linguistic terms to describe Turkish grammatical patterns (e.g., *suffix*, *case marker*), but rather listed “words” they had noticed (e.g., *lar* for two objects). Indeed, some participants exhibited high posttest scores without expressing any awareness, suggesting that it was possible to learn aspects of Turkish grammar implicitly (see also Reber, 1967; Rogers, 2017). In tests of mediation, aptitude (nonverbal ability) had an indirect effect on posttest accuracy but a direct effect on metalinguistic awareness, suggesting that variation in L2 outcomes related to aptitude might stem from differences in awareness (Brooks & Kempe, 2013). Under this view, aptitude assists in the development of explicit awareness, which in turn enhances accuracy in L2 comprehension. Future work should include declarative memory measures, which load onto the language-related component of aptitude (Grigorenko et al., 2000), to explore further how individual differences in aptitude contribute to L2 comprehension and metalinguistic awareness.

Acknowledgements

This research was supported by a NSF Doctoral Dissertation Research Improvement Grant awarded to PB and MR, a CUNY Graduate Center Doctoral Student Research Grant awarded to MR, and a PSC-CUNY Research Award to PB.

References

Akifumi, Y. (2016). The effects of receptive and productive word retrieval practice on second language vocabulary learning. *KATE Journal*, *30*, 139–152.

Brooks, P. J., & Kempe, V. (2013). Individual differences in adult foreign language learning: The mediating effect of metalinguistic awareness. *Memory & Cognition, 41*, 281–296.

Brooks, P. J., Kwoka, N., & Kempe, V. (2017). Distributional effects and individual differences in L2 morphology learning. *Language Learning*, *67*(1), 171–207.

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

Cattell, R. B., & Cattell, A. K. S. (1973). *Measuring intelligence with the Culture Fair tests.* Institute for Personality and Ability Testing.

Dahlen, K., & Caldwell-Harris, C. (2013). Rehearsal and aptitude in foreign vocabulary learning. *The Modern Language Journal*, *97*(4), 902–916.

Dörnyei, Z. (2014). *The psychology of the language learner: Individual differences in second language acquisition*. Routledge.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition*, *18*(1), 91–126.

Forrin, N. D., MacLeod, C. M., & Ozubko, J. D. (2012). Widening the boundaries of the production effect. *Memory & Cognition*, *40*, 1046–1055.

Foucart, A., & Frenck-Mestre, C. (2011). Grammatical gender processing in L2: Electrophysiological evidence of the effect of L1–L2 syntactic similarity. *Bilingualism: Language and Cognition*, *14*(3), 379–399.

Granena, G., Jackson, D. O., & Yilmaz, Y. (Eds.). (2016). *Cognitive individual differences in second language processing and acquisition*. John Benjamins.

Grigorenko, E. L., Sternberg, R. J., & Ehrman, M. E. (2000). A theory-based approach to the measurement of foreign language learning ability: The Canal-F theory and test. *The Modern Language Journal*, *84*(3), 390–405.

Hamrick, P. (2015). Declarative and procedural memory abilities as individual differences in incidental language learning. *Learning and Individual Differences*, *44*, 9–15.

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Publications.

Hopman, E. W. M., & MacDonald, M. C. (2018). Production practice during language learning improves comprehension. *Psychological Science, 29*(6) 961–971.

Kang, S. H. K., Gollan, T. H., & Pashler, H. (2013). Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic Bulletin and Review*, *20*(6), 1259–1265.

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review*, *27*(2), 317–326

Karpicke, J. & Roediger, H. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968.

Kempe, V., & Brooks, P. J. (2016). Miniature natural language learning in L2 acquisition research. In G. Granena, D. O. Jackson, & Y, Yilmaz (Eds.), *Cognitive individual differences in second language processing and acquisition*. John Benjamins.

Keppenne, V., Hopman, E. W., & Jackson, C. N. (2021). Production-based training benefits the comprehension and production of grammatical gender in L2 German. *Applied Psycholinguistics, 42*(4), 907–936.

Leow, R. P. (2018). Noticing hypothesis. In J. I. Liontas (Ed.), *The TESOL Encyclopedia of English Language Teaching*. John Wiley & Sons, Inc.

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review*, *14*(2), 194–199.

Miyake, A., & Friedman, N. P. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. F. Heal & L. E. Bourne, Jr. (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention*. Erlbaum.

Morris, C. D., Bransford, J. D. & Franks, J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*(6), 855-863.

Rogers, J. (2017). Awareness and learning under incidental learning conditions. *Language Awareness*, *26*(2), 113–133.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463.

Schmidt, R. W. (1990). The role of conscisousness in second language learning. *Applied Linguistics*, *11*(2), 129–158.

Swain, M. & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics*, *16*(3), 371–391.