Studies in Second Language Acquisition, **43** (2021), 663–676
doi:10.1017/S0272263120000704

Research Report

PROBING THE CONSTRUCT VALIDITY OF LLAMA_D AS A MEASURE OF IMPLICIT LEARNING APTITUDE

INCIDENTAL INSTRUCTIONS, CONFIDENCE RATINGS, AND REACTION TIME

Yuichi Suzuki *

Kanagawa University

Abstract

A subtest of the LLAMA test battery (LLAMA_D) has been proposed as a potential test of implicit learning aptitude. To improve its construct validity, in the present study, the original LLAMA_D (a) instructions for incidental learning were modified, and (b) confidence ratings of test responses and (c) reaction time (RT) measurements were added. This revised LLAMA_D was administered along with the other LLAMA subtests (LLAMA-B, -E, and -F). Unconscious knowledge that may (not) result from the exposure was assessed through the relationship between the accuracy/RT and confidence ratings. The results suggest that LLAMA_D accuracy largely reflects conscious retrieval of previously heard sound sequences. However, an index derived from the LLAMA_D RT measure (coefficient of variance) was associated with an aspect of oral fluency, which is presumably dependent on proceduralization. Several recommendations are proposed to redesign and extend LLAMA_D as a potential aptitude test for proceduralization.

INTRODUCTION

Renewed interest in aptitude in the context of second language (L2) acquisition has led to several recent studies exploring this concept (Granena et al., 2016; Wen et al., 2019). As

The experiment in this article earned an Open Materials badge for transparent practices. The materials are available at https://www.iris-database.org/iris/app/home/detail?id=york%3a938863&ref=search

This study was supported by Grant-in-Aid for Scientific Research (KAKENHI) from Japan Society for the Promotion of Science (JSPS). I would like to express my sincere gratitude to Prof. Paul Meara for his generous encouragement of this project. I am grateful to my RAs, Atsushi Miura, Misaki Kuratsubo, and Miyu Koyama for their assistance in data collection and coding.

* Correspondence concerning this article should be addressed to Yuichi Suzuki, Department of Cross-Cultural Studies, Kanagawa University, 3-27-1, Rokkakubashi, Kanagawa-ku, Yokohama-shi, Kanagawa 221-8686, Japan. E-mail: szky819@kanagawa-u.ac.jp

an aptitude test battery is an integral part of aptitude research, the LLAMA test is increasingly being used, in part because it is language independent and can be administered to all individuals who can read the Roman alphabet. Furthermore, it is freely available, unlike other high-stake and secured aptitude tests such as the Modern Language Aptitude Test (MLAT) (Carroll & Sapon, 1959; Sasaki, 2012) and High-Level Language Aptitude Battery (Hi-LAB) (Linck et al., 2013). Using the LLAMA test, a number of key findings have been generated in various domains of L2 learning, including but not limited to the role of aptitude in L2 acquisition by children and adults (e.g., Abrahamsson & Hyltenstam, 2008; Granena, 2013b) and aptitude–treatment interactions (e.g., Kourtali & Révész, 2019; Suzuki & DeKeyser, 2017a; Yilmaz & Grañena, 2016). As LLAMA was originally developed for exploratory and non-high-stakes purposes (Meara, 2005), it is necessary to assess its reliability and validity. Such investigations have recently commenced, and potential and limits are gradually emerging (Bokander & Bylund, 2019; Granena, 2013a, 2019; Rogers et al., 2017; Yalçın et al., 2016).

Although the LLAMA test was developed based on the MLAT (Meara, 2005), it also includes a new subcomponent, LLAMA_D, which is a sound-recognition test claimed to be a potential measure of aptitude for implicit learning (Granena, 2013a, 2019). While traditional aptitude measures like MLAT predominantly focus on explicit learning, whereby test-takers are expected to consciously remember and analyze linguistic materials, LLAMA_D may have the potential to serve as an implicit learning aptitude measure, something that is urgently needed in the L2 field (DeKeyser, 2019; Skehan, 2019).

LLAMA_D AS A POTENTIAL MEASURE OF APTITUDE FOR IMPLICIT LEARNING

Previous research has consistently indicated that the three subcomponents of the LLAMA test (B, E, and F) tend to correlate with each other, LLAMA_D is not related to any of these measures (Artieda & Muñoz, 2016; Bokander & Bylund, 2019; Granena, 2013a, 2019). This is to be expected, given that the former three subtests involve the deliberate study phase of linguistic stimuli. However, LLAMA_D comprises solely of a simple exposure phase where test-takers listen to 10 sound sequences. Granena (2013a, 2019) proposed an intriguing hypothesis that LLAMA_D can capture individual differences in some aspects of implicit memory. According to Granena (2019), "learning conditions created by the test are closer to implicit induction (i.e., acquiring patterns unintentionally through exposure) than explicit induction (i.e., figuring out rules and relations)" and test-takers respond to the test prompts based on "feelings of familiarity involving fast, automatic processes, rather than on conscious recollection involving slow, controlled search processes" (pp. 315−316).

This proposal is theoretically supported by the findings yielded by research on recognition memory in the context of psychology (Eichenbaum et al., 2007; Wang & Yonelinas, 2012; Yonelinas, 2002). Recognition memory is often measured by a judgment task, similarly to the LLAMA_D test procedure, in which test-takers are required to determine whether an item has been presented to them previously or not. Two distinct cognitive mechanisms are stipulated to underlie recognition memory—*recollection* and *familiarity-based recognition.* The first requires conscious retrieval of an episodic event with detailed contextual information, whereas the second merely denotes the feeling that

https://doi.org/10.1017/S0272263120000704 Published online by Cambridge University Press

an item has been seen previously without explicitly recalling detailed contextual information (Granena, 2019).

NEW AND OLD LLAMA_D ITEMS

The LLAMA_D test consists of 10 familiar items (all of which have been presented to test-takers during the exposure phase) and 20 new items. In the test phase, test-takers have to determine whether they have previously heard the word (old items) or not (new items). Results yielded by a recent large-scale validation study conducted by Bokander and Bylund (2019) revealed that old and new items generally loaded onto different factors, suggesting that further exploration of the distinction between familiar and new stimuli is warranted in the current study.

SUBJECTIVE MEASURES OF CONSCIOUSNESS: CONFIDENCE RATINGS

Several measures have been proposed for distinguishing conscious and unconscious knowledge using subjective ratings during test performance (Rebuschat, 2013). In a series of studies focusing on implicit learning, Dienes and colleagues have demonstrated that unconscious knowledge can be acquired, for example, by learning grammar of an artificial language (Dienes, 2012; Dienes & Perner, 1999; Dienes & Scott, 2005). Two subjective methods—source attribution and confidence ratings—are typically used to dissociate the conscious from unconscious knowledge. The latter is of particular interest to the current study, as the participants were instructed to report their confidence levels (e.g., highly confident, somewhat confident, not confident) for each response to a task, such as the forced-choice judgment task. When required to provide responses based on conscious knowledge, learners typically demonstrate higher accuracy for higher-confidence items. In contrast, when tests tap into unconscious knowledge, above-chance level of accurate performance can be attained even for lower-confidence items. These findings, combined with the negative correlation between confidence level and accuracy reported by Scott and Dienes (2010), support the existence of unconscious knowledge. In sum, accuracy and confidence are positively correlated when responses are based on conscious knowledge, whereas no such relationship or even a negative relationship is expected for those tapping into unconscious knowledge. To explore the unconscious knowledge that may or may not be acquired from the LLAMA_D, in the present study, participants were required to report their confidence levels for each test item.

LLAMA_D INSTRUCTIONS

The instructions provided to test-takers influence the type of learning (e.g., intentional or incidental) induced (Williams & Paciorek, 2015). When the instructions direct learners' attention to a linguistic form that should be memorized, they are more likely to learn intentionally. However, when the instructions are more incidental (e.g., directing attention to nontarget features) during the learning phase, learners would be less likely to engage in conscious learning.

The LLAMA_D instructions provided in the LLAMA test manual state "your task is to listen carefully to these words. In the test phase of the program, you will hear these words

https://doi.org/10.1017/S0272263120000704 Published online by Cambridge University Press

alongside other words that you have not heard before" (Meara, 2005, p. 9). In a recent study, Saito (2017) explored the possibility of modifying the wording to render these instructions more incidental. Thus, the modified LLAMA_D instructions merely stated that participants should focus on the sound in the exposure phase, without any indication that they would be tested later. This incidental instruction format was adopted "to respond to a concern raised by Granena (2013b): LLAMA_D might encourage test takers to use conscious and intentional learning strategies if they were informed that their memory was to be tested after the listening session" (Saito, 2017, p. 689). Using the incidental instructions may allow LLAMA_D to serve as an implicit learning aptitude test. As this potential has not been fully explored yet, this gap in the extant knowledge has motivated the current study.

UTILIZING RT IN LLAMA_D AND ITS RELATIONSHIP WITH UTTERANCE FLUENCY

The most comprehensive aptitude test battery, Hi-LAB, consists of two tests—serial-reaction time (SRT) and available long-term memory (ALTM) tasks—aimed at assessing implicit learning aptitude (Linck et al., 2013). The two tests presumably tap into different types of implicit (nondeclarative) memory (Granena, 2019), whereby SRT assesses the sequence learning ability, and ALTM measures the primability. In both aptitude tests, however, reaction time (RT) is captured to infer the implicit (unconscious) behavior (e.g., Nissen & Bullemer, 1987). Mere fast processing speed, however, does not guarantee that a cognitive process is implicit (e.g., Moors, 2016) because implicit processes should entail a lack of awareness. By contrast, a different aspect of nondeclarative memory, that is, proceduralization, does not require the unawareness criterion. A simpler criterion like efficient information processing may suffice for capturing proceduralization. It thus may be worthwhile incorporating RT in the LLAMA_D test and investigating its potential value for assessing aptitude for proceduralization or even for implicit learning.

To assess the predictive validity of LLAMA_D, examining one facet of speaking proficiency—utterance fluency—is useful because fluent speech requires efficient linguistic encoding that is typically associated with proceduralization of the linguistic knowledge (De Jong et al., 2013; Kahng, 2014; Kormos, 2006). Of particular interest for the current investigation, Granena (2019) found that implicit memory (measured by the composite score of the LLAMA_D accuracy and ALTM scores) was a significant predictor of L2 utterance fluency (i.e., pruned speech rate per minute). In the present study, utterance fluency (as an indicator of procedural and/or implicit knowledge) was utilized to investigate the extent to which RT, as well as accuracy score, in the LLAMA_D test can predict utterance fluency measures.

CURRENT STUDY

The goal of this exploratory study is to scrutinize the construct validity of LLAMA_D as a measure of implicit learning aptitude. To achieve this aim, the original LLAMA_D was modified in the following respects: (a) changing instructions for incidental learning, (b) adding confidence ratings of test responses, and (c) measuring reaction time of test responses. The current investigation cannot be considered as a validation study of the original LLAMA_D test; the three aforementioned modifications were made to the original test to explore its potential and limits. Furthermore, the LLAMA_D scores were

https://doi.org/10.1017/S0272263120000704 Published online by Cambridge University Press

compared to an L2 learning outcome that is presumably tied to implicit and/or procedural learning in the L2 acquisition context. L2 utterance fluency measures (i.e., speed and breakdown fluency) were used as dependent variables representing implicit knowledge or at least proceduralization underlying L2 speaking (De Jong et al., 2013; Kahng, 2014; Kormos, 2006). The following three research questions were addressed:

1. Do the old and new items of LLAMA_D assess different aspects of sound recognition ability?
2. Are confidence ratings associated with accuracy and RT measures of LLAMA_D?
3. Is utterance fluency (proceduralization in speaking) associated with LLAMA_D accuracy and RT measures?

METHODS

PARTICIPANTS

The study sample comprised 59 Japanese native speakers, all of whom were students attending a private Japanese university (aged 18−22 years). Their English proficiency was estimated to range from A2 (elementary) to B1 (intermediate) level on the Common European Framework of Reference for Languages (CEFR) benchmark. They all took the LLAMA test individually in a quiet office.

INSTRUMENTS

Modified LLAMA_D. The original LLAMA_D was modified and programmed using the presentation software DMDX (Forster & Forster, 2003). In line with the strategy adopted by Saito (2017), in the LLAMA_D exposure phase, participants were instructed to check the sound volume before they listened to 10 new words, to encourage incidental learning. In the subsequent (unannounced) recognition test, they were required to indicate as quickly as possible whether or not a word they heard was present in the exposure phase. To assess unconscious knowledge that may or may not result from prior exposure, confidence ratings were added to the test (e.g., Rebuschat, 2013). The unannounced (surprise) test consisted of 30 items (10 of which were old and 20 were new), each of which required participants to make a decision (i.e., old vs. new) and provide a confidence rating on a four-point scale (i.e., not confident at all, slightly confident, very confident, 100% confident), following the design adopted by Norman and Price (2015). Once the participant selected a confidence rating, he/she was presented with the next test item. It is important to note that two modifications to the instructions (i.e., checking the sound volume and responding as quickly as possible) might influence participants' response behaviors, and that the need to rate one's confidence in the given response might detract attention and induce memory decay due to longer testing time.

All accuracy values recorded by the program were used for further analysis. A mean RT was computed only for correct responses because RTs associated with incorrect responses may not reflect the same underlying cognitive processes. In addition to the speed measure (RT), the coefficient of variance (CV) was calculated for RTs and served as a processing efficiency index. The CV, which is computed as the ratio of participants' mean SD of RT and mean RT, has been widely used in cognitive psychology research when studying

https://doi.org/10.1017/S0272263120000704 Published online by Cambridge University Press

behavioral patterns of healthy individuals as well as brain-injured patients (Segalowitz & Segalowitz, 1993). Following N. S. Segalowitz and Frenkiel-Fishman's (2005) rationale, RTs and CV can capture different dimensions of LLAMA_D performance, which allows us to examine the potential links with speaking fluency measures more thoroughly. The former simply reflects processing speed, and the latter indicates stability of processing after correcting for processing speed.

LLAMA_B. LLAMA_B consists of a learning and a test phase and assesses the ability to learn vocabulary in written form. In the 2-minute learning phase, participants were told to remember 20 words associated with stimulus pictures. In the subsequent test phase, they were required to choose a correct picture for each of these 20 words.

LLAMA_E. LLAMA_E is a sound–symbol correspondence task comprising of a 2-minute learning phase, during which participants learned to connect alphabet-like symbols with different sounds, and a test phase, consisting of 20 two-choice questions.

LLAMA_F. LLAMA_F assesses grammatical inferencing ability. In the 5-minute learning phase, participants were expected to infer grammatical rules by studying 20 sentences and corresponding pictures. In the test phase, they were required to respond to forced two-choice questions. As the LLAMA_F subtest yielded relatively lower reliability in previous research, in the present study, 15 additional items were added in an attempt to increase the reliability (these additional items are available in the IRIS database; see Suzuki & DeKeyser, 2017b for a similar approach). These additional items were created based on the existing items, and the correlations of this extended LLAMA_F with the other subtests were very similar to the ones reported in previous studies (see Appendix C in the Online Supplementary File). Note that the LLAMA subtests other than LLAMA_D were included in this study to check whether the current LLAMA test results are consistent with the previous findings (i.e., descriptive statistics, reliability, and correlations).

Speaking test. The speaking test was administered to current study participants to measure utterance fluency, which is presumably dependent on proceduralization of L2 knowledge. The test was an oral narrative story task based on a six-frame cartoon. The students received the following instructions: "Yesterday, you saw an event depicted in the six-frame cartoon on the next page. You are going to explain the story to a friend who doesn't know the story in three minutes." The procedure involved (a) a 3-minute planning and (b) a 3-minute oral narration. In the planning stage, participants saw a cartoon (prompt) along with several guiding questions and a useful vocabulary list (13 English–Japanese word pairs) that facilitated content generation. During the oral narration, participants were presented with the prompt only and were expected to narrate the story in 3 minutes without the aid of the guiding questions or useful vocabulary list. While all 59 students took part in this test, due to problems with the recording device, only 50 valid datasets were obtained and were submitted to fluency analysis.

The speech data were transcribed and analyzed using Praat software (Boersma & Weenink, 2016). Two objective fluency measures were derived: (a) articulation rate (the number of syllables per minute of speech, excluding pauses) as a measure of speed fluency and (b) mid-clause pause duration (mean duration of mid-clause filled and unfilled pauses) as a measure of breakdown fluency. Mid-clause pauses were used rather than clause-final pauses, because the former are more directly related to linguistic processing difficulties (and thus L2 proceduralization indices), whereas the latter is believed to be more

associated with content planning, including nonlinguistic processes (Kahng, 2014, 2017; Lambert et al., 2017; Suzuki, 2020).

RESULTS

PRELIMINARY ANALYSES

Descriptive statistics and reliability indexed by Cronbach's alpha pertaining to all LLAMA tests assessed in the present study are provided in Appendices A and B in the Online Supplementary File. Although in their large-scale LLAMA test validation study Bokander and Bylund (2019) reported the lowest reliability of all subtests for the original LLAMA_D (.54), in the current study, an even lower Cronbach alpha (.20) for the modified LLAMA_D was obtained. When LLAMA_D items were divided into old and new, their respective reliability scores increased to .55 and .47.

Further preliminary analyses were conducted to check the compatibility with the previous research findings on the construct validity of LLAMA_D. Two exploratory principal component analyses were conducted. The obtained results broadly confirmed that LLAMA_D loaded onto a distinct component separately from the other three LLAMA subtests (B, E, and F). The full results are reported in Appendix C in the Online Supplementary File.

NEW AND OLD LLAMA_D ITEMS

To explore the (old and new) item characteristics of the LLAMA_D subtest, a two-component principal component analysis of the 30 items was conducted. Consistent with Bokander and Bylund's (2019) findings, new and old items clustered separately. There were 14 items (7 old and 7 new) with $\geq .30$ component loadings, whereby all new items clustered together, and all old items were categorized as the different component.

Mean and SD values of old and new categories (15 items for each category) are presented in Appendix D in the Online Supplementary File. Paired-samples *t*-tests revealed no significant differences in accuracy between the old and new items, $t(58) = -0.21$, $p = .83$, Cohen's $d = -0.03$. Moreover, RT was significantly faster for the old compared to the new items with a medium effect size, $t(57) = -4.50$, $p < .001$, Cohen's $d = -0.59$. CV was also smaller, but only marginally significantly so, for the old items than for the new items with a small effect size, $t(57) = -1.87$, $p = .07$, Cohen's $d = -0.25$.

CONFIDENCE RATING OF LLAMA_D

As many participants did not use all eight categories of responses on the LLAMA_D test (2 [old/new] $\times$ 4 [four confidence levels]), 2 $\times$ 4 repeated-measures ANOVA could not be run (see Appendix E in the Online Supplementary File for the complete dataset). To overcome this limitation, the four confidence levels were collapsed into two categories (high vs. low), in line with the approach adopted by Dienes and Scott (2005). After the four confidence levels were collapsed into two categories (high vs. low), a

https://doi.org/10.1017/S0272263120000704 Published online by Cambridge University Press

TABLE 1. Results of repeated-measures ANOVA (accuracy data)

	Mean square	F	p	η^2_p
Item Type	0.14	2.19	0.15	0.04
Confidence	2.49	62.08	< .001	0.57
Item Type × Confidence	1.32	36.33	< .001	0.44

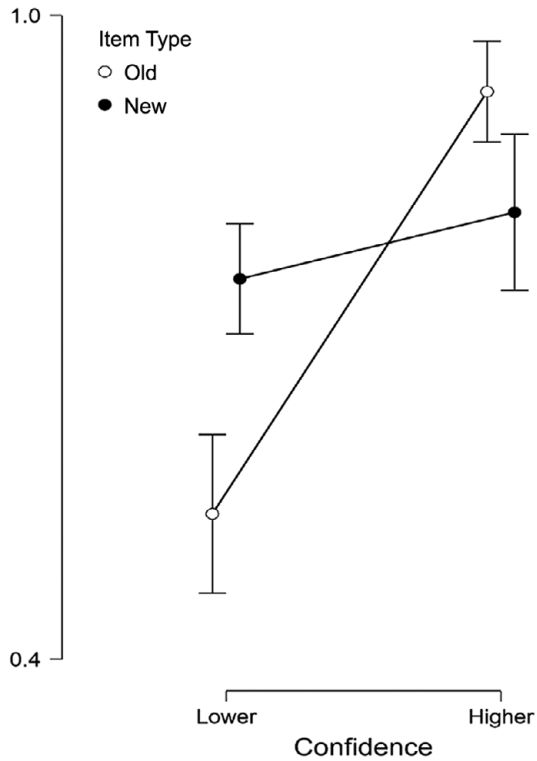

FIGURE 1. Accuracy rates of LLAMA_D (Item Type × Confidence).
Note: The error bars indicate 95% confidence intervals.

2 × 2 repeated-measures ANOVA was conducted on the available accuracy data ($n = 48$). The assumption of sphericity was met. The analysis yielded a significant effect of Confidence with a large effect size (see Table 1). This means that the LLAMA_D accuracy scores were positively correlated with confidence, suggesting that the participants demonstrated their conscious knowledge of sound sequence. Intriguingly, an interaction between Item Type and Confidence was also significant with a large effect size.

The Item Type × Confidence interaction is illustrated in Figure 1 where the accuracy rates (%) of old and new items are plotted for responses for which the participants reported higher and lower confidence. For the old items, the accuracy rate was higher for the higher-confidence items ($M = 93\%$, $SD = 12\%$) than the lower-confidence items

https://doi.org/10.1017/S0272263120000704 Published online by Cambridge University Press

TABLE 2. Results of repeated-measures ANOVA (RT data)

	Mean square	*F*	*p*	η^2_p
Item Type	51,600	0.14	0.71	0.00
Confidence	23,401,310	74.13	<.001	0.66
Item Type × Confidence	3,321	0.01	0.92	0.00

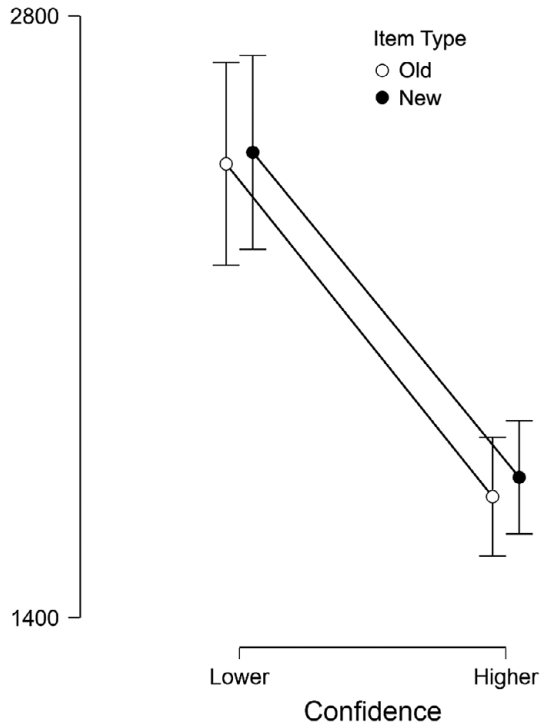

FIGURE 2. RT of LLAMA_D (item type × confidence).
Note: The error bars indicate 95% confidence intervals.

($M = 54\%$, $SD = 27\%$), $t(47) = 10.02$, $p < .001$, $d = 1.45$. For the new items, there was no significant difference between the higher-confidence items ($M = 82\%$, $SD = 27\%$) and the lower-confidence items ($M = 75\%$, $SD = 17\%$), $t(47) = 10.02$, $p = .13$, $d = -0.22$.

Similar to accuracy data analysis, a 2 × 2 repeated-measures ANOVA was conducted on RT data ($n = 40$). The assumption of sphericity was met. As shown in Table 2, once again, the effect of Confidence was significant with a large effect size, suggesting that responses in which students were more confident were provided more rapidly. Neither the effect of Item Type nor the interaction with Confidence was significant. Figure 2 illustrates that, irrespective of item type, the responses in which participants felt more confident (Old Item: $M = 1{,}682$, $SD = 456$; New Item: $M = 1{,}727$, $SD = 392$) were provided more rapidly than those in which the students were less confident (Old Item: $M = 2{,}456$, $SD = 820$; New Item: $M = 2{,}482$, $SD = 718$).

https://doi.org/10.1017/S0272263120000704 Published online by Cambridge University Press

TABLE 3. Pearson's correlations between LLAMA_D scores and fluency measures

	Accuracy		RT		CV	
	Old	New	Old	New	Old	New
Articulation rate	.02	.01	−.05	.10	−.11	.20
Mid-clause pause duration	.04	−.20	.14	−.06	.35*	−.05

*$p < .05$

RELATIONSHIP BETWEEN LLAMA_D AND UTTERANCE FLUENCY MEASURES

To examine the association of LLAMA subtest scores with objective fluency measures, correlation coefficients were computed. None of the LLAMA subtest scores (LLAMA_B, LLAMA_E, and LLAMA_F, as well as accuracy, RT and CV of LLAMA_D) were significantly correlated with articulation rate or mid-clause pause duration (see Appendix F in the Online Supplementary File). However, when the scores were recomputed separately for old and new LLAMA_D items (see Table 3), a significant positive relationship was found between mid-clause pause duration and CV of old items ($r = .35$, $p = .01$). While no correction was made to the p values for these multiple correlation coefficients to avoid reaching too conservative a decision, the magnitude of the correlation coefficient was interpreted to be reasonably meaningful, given Li's (2016) meta-analysis on the association between aptitude and L2 speaking skill ($r = .37$).

DISCUSSION

To improve the construct validity of LLAMA_D, three research questions were posed and the findings are summarized here. First, old items not only clustered separately from the new items but also were responded to faster and more stably. Second, higher confidence ratings were associated with higher accuracy and faster RTs, suggesting that learners were applying conscious (explicit) knowledge in the test. Last, old item recognition efficiency was related to an aspect of oral fluency (i.e., mid-clause pause duration).

THE NEED FOR CLARIFYING THE APTITUDE CONSTRUCT MEASURED BY LLAMA_D

Based on the positive relationship between confidence ratings and test performance (Dienes, 2012; Dienes & Perner, 1999; Dienes & Scott, 2005), the validity of LLAMA_D as an implicit aptitude measure is challenged, countering the arguments put forth by Granena (2013a, 2019). Item recognition presumably requires (conscious) recollection of previously encountered stimuli (Eichenbaum et al., 2007; Wang & Yonelinas, 2012; Yonelinas, 2002). Conscious recognition memory is supported by parts of the medial temporal lobe, including the hippocampus, which is closely related to the declarative memory system (Eichenbaum et al., 2007). If LLAMA_D measures conscious recognition memory, it can simply be a measure of declarative–explicit rather than implicit memory.

By contrast, the CV for RT may be worthy of further exploration as a potential measure of proceduralization (consolidating linguistic constructions for more fluent use). Proceduralization is often equated with implicit learning (e.g., Skehan, 2019),

https://doi.org/10.1017/S0272263120000704 Published online by Cambridge University Press

whereby it would indicate speed and efficiency of L2 knowledge and skill usage. However, these phenomena are distinct, as implicit learning should take place without conscious awareness, whereas proceduralization does not necessarily require the unawareness criterion. Consequently, instead of focusing on aptitude for implicit learning in the narrow sense of its definition, it is argued that an aptitude test like LLAMA_D can be seen as a measure of a memory-based, procedural system (Buffington & Morgan-Short, 2019; Henke, 2010). This characterization of proceduralization is compatible with one of the skills examined in this study, that is, oral fluency. The current findings indicate that CV was significantly, albeit weakly, related to mid-clause pause duration in L2 speech, which presumably reflects proceduralization of linguistic formulations, such as lexical retrieval and grammatical encoding (De Jong et al., 2013; Kahng, 2014; Kormos, 2006). LLAMA_D may, therefore, tap into the ability to *accurately* recognize a sequence of words, as well as the ability to do it quickly and, perhaps more importantly, *efficiently*, which may play a pivotal role in L2 proceduralization. Owing to the nature of this exploratory study, this interpretation of this modified LLAMA_D as a measure of aptitude for proceduralization remains speculative. However, given the dearth of aptitude measurements for proceduralization and its importance for L2 learning (DeKeyser, 2019; Skehan, 2019), it is hoped that this tentative interpretation can stimulate future pursuit of this line of thinking.

FUTURE DIRECTIONS: A PROPOSAL FOR REDESIGNING LLAMA_D FOR MEASURING PROCEDURALIZATION APTITUDE

In the current study, the impact of several modifications to the original LLAMA_D was explored. To extend LLAMA_D as a proceduralization aptitude measure, four key issues must be overcome. First, the incidental instructions (e.g., Saito, 2017) resulted in a much lower internal consistency of LLAMA_D compared to that reported in previous research using the original instructions. Because the instructions for incidental learning seem to lower the test reliability, instructions for intentional learning are preferable. Theoretically, the intentional instructions are also compatible with the idea of proceduralization conceived as a deliberate learning (DeKeyser, 2015).

Second, although *old* item recognition accuracy was significantly correlated with confidence ratings, the same did not apply for *new* items. Consistent with the results reported by Bokander and Bylund (2019), systematic differences between old and new items were detected using multiple methods (confidence rating, reliability analysis, principal component analysis). These findings suggest that familiar and new items tap into distinct underlying abilities. Thus, rather than using a composite score, scores should be reported separately for new and old items, provided that a sufficient number of test items is available to ensure high internal consistency. Possibly, signal detection theory (Green & Swets, 1966) can be applied to compute the number of "yes" responses to new items (i.e., "false alarm" rate), as this was purported to serve as a useful index of familiarity memory, as opposed to recognition memory (see Yonelinas, 2002 for more details).

Third, the current study findings indicate that RT and CV can potentially be used as useful indicators of individual differences in aptitude for proceduralization. As RT is utilized in the existing implicit learning aptitude tests (SRT and ALTM tasks), the

https://doi.org/10.1017/S0272263120000704 Published online by Cambridge University Press

usefulness of RT measures for aptitude test construction should be explored further, given its extensive use in psycholinguistic L2 research (Jiang, 2011; Suzuki, 2017).

Last, researchers can consider increasing the number of exposure and test items; a greater number of test items would increase test reliability. Because proceduralization requires repeated practice and long-term L2 use, it may also be worth creating multiple learning sessions (Skehan, 2019), instead of a single 10-item exposure phase.

CONCLUSIONS

The findings yielded by this exploratory study cast doubt on the validity of LLAMA_D as a measure of implicit aptitude. It is therefore proposed that LLAMA_D be construed as a measure of aptitude for proceduralization in the domain of L2 acquisition (i.e., consolidating linguistic constructions for more fluent L2 use). Moreover, four design features of LLAMA_D need to be modified to improve and refine the target construct of the test. Further development and expansion of the LLAMA_D test is therefore warranted.

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S0272263120000704.

REFERENCES

Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition*, *30*, 481–509.

Artieda, G., & Muñoz, C. (2016). The LLAMA tests and the underlying structure of language aptitude at two levels of foreign language proficiency. *Learning and Individual Differences*, *50*, 42–48.

Boersma, P., & Weenink, D. (2016). Praat: Doing phonetics by computer. Version 6.0.14. http://www.praat.org/

Bokander, L., & Bylund, E. (2019). Probing the internal validity of the LLAMA language aptitude tests. *Language Learning*, *Advanced Access*, *70*, 11–47.

Buffington, J., & Morgan-Short, K. (2019). Declarative and procedural memory as individual differences in second language aptitude. In Z. Wen, P. Skehan, A. Biedroń, S. Li, & R. L. Sparks (Eds.), *Language aptitude: Advancing theory, testing, research and practice* (pp. 215–237). Routledge.

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

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, *34*, 893–916.

DeKeyser, R. M. (2015). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed., pp. 94–112). Routledge.

DeKeyser, R. M. (2019). The future of aptitude research. In Z. Wen, P. Skehan, A. Biedron, S. Li, & R. L. Sparks (Eds.), *Language aptitude: Advancing theory, testing, research, and practice* (pp. 317–329). Routledge.

Dienes, Z. (2012). Conscious versus unconscious learning of structure. In P. Rebuschat & J. Williams (Eds.), *Statistical learning and language acquisition* (Vol. 1, pp. 337–364). Mouton De Gruyter.

Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences*, *22*, 735–808.

Dienes, Z., & Scott, R. (2005). Measuring unconscious knowledge: Distinguishing structural knowledge and judgment knowledge. *Psychological Research*, *69*, 338–351.

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience*, *30*, 123–152.

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, *35*, 116–124.

Granena, G. (2013a). Cognitive aptitudes for L2 learning and the LLAMA language aptitude test. In G. Granena, & Long, M. H. (Ed.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 105–130). John Benjamins.

Granena, G. (2013b). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning*, *63*, 665–703.

Granena, G. (2019). Cognitive aptitudes and L2 speaking proficiency: Links between LLAMA and Hi-LAB. *Studies in Second Language Acquisition*, *41*, 313–336.

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

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley.

Henke, K. (2010). A model for memory systems based on processing modes rather than consciousness. *Nature Reviews Neuroscience*, *11*, 523–532.

Jiang, N. (2011). *Conducting reaction time research in second language studies*. Routledge.

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning*, *64*, 809–854.

Kahng, J. (2017). The effect of pause location on perceived fluency. *Applied Psycholinguistics*, *39*, 569–591.

Kormos, J. (2006). *Speech production and second language acquisition*. Routledge.

Kourtali, N.-E., & Révész, A. (2019). The roles of recasts, task complexity, and aptitude in child second language development. *Language Learning*, *Advanced Access*, *70*, 179–218.

Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition*, *39*, 167–196.

Li, S. (2016). The construct validity of language aptitude: A meta-analysis. *Studies in Second Language Acquisition*, *38*, 801–842.

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

Meara, P. M. (2005). LLAMA language aptitude tests: The manual. *Lognostics.* http://www.lognostics.co.uk/tools/llama/llama_manual.pdf

Moors, A. (2016). Automaticity: Componential, causal, and mechanistic explanations. *Annual Review of Psychology*, *67*, 263–287.

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

Norman, E., & Price, M. C. (2015). Measuring consciousness with confidence ratings. In M. Overgaard (Ed.), *Behavioral methods in consciousness research* (pp. 159–180). Oxford University Press.

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, *63*, 595–626.

Rogers, V., Meara, P. M., Barnett-Legh, T., Curry, C., & Davie, E. (2017). Examining the LLAMA aptitude tests. *Journal of the European Second Language Association*, *1*, 49–60.

Saito, K. (2017). Effects of sound, vocabulary, and grammar learning aptitude on adult second language speech attainment in foreign language classrooms. *Language Learning*, *67*, 665–693.

Sasaki, M. (2012). The modern language aptitude test (paper-and-pencil version). *Language Testing*, *29*, 315–321.

Scott, R. B., & Dienes, Z. (2010). Knowledge applied to new domains: The unconscious succeeds where the conscious fails. *Consciousness and Cognition: An International Journal*, *19*, 391–398.

Segalowitz, N., & Frenkiel-Fishman, S. (2005). Attention control and ability level in a complex cognitive skill: Attention shifting and second-language proficiency. *Memory & Cognition*, *33*, 644–653.

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

Skehan, P. (2019). Language aptitude implicates language and cognitive skills. In Z. E. Wen, P. Skehan, A. Biedroń, S. Li, & R. L. Sparks (Eds.), *Language aptitude: Advancing theory, testing, research and practice* (pp. 56–77). Routledge.

Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics*, *38*, 1229–1261.

Suzuki, Y. (2020). Optimizing fluency training for speaking skills transfer: Comparing the effects of blocked and interleaved task repetition. *Language Learning*. Advance online publication. https://doi.org/10.1111/lang.12433.

Suzuki, Y., & DeKeyser, R. M. (2017a). Exploratory research on L2 practice distribution: An aptitude × treatment interaction. *Applied Psycholinguistics*, *38*, 27–56.

Suzuki, Y., & DeKeyser, R. M. (2017b). The interface of explicit and implicit knowledge in a second language: Insights from individual differences in cognitive aptitudes. *Language Learning*, *67*, 747–790.

Wang, W.-C., & Yonelinas, A. P. (2012). Familiarity is related to conceptual implicit memory: An examination of individual differences. *Psychonomic Bulletin & Review*, *19*, 1154–1164.

Wen, Z. E., Skehan, P., Biedroń, A., Li, S., & Sparks, R. L. (2019). *Language aptitude: Advancing theory, testing, research and practice*. Routledge.

Williams, J. N., & Paciorek, A. (2015). Indirect tests of implicit linguistic knowledge. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice* (pp. 37–54). Routledge.

Yalçın, Ş., Çeçen, S., & Erçetin, G. (2016). The relationship between aptitude and working memory: An instructed SLA context. *Language Awareness*, *25*, 144–158.

Yilmaz, Y., & Grañena, G. (2016). The role of cognitive aptitudes for explicit language learning in the relative effects of explicit and implicit feedback. *Bilingualism: Language and Cognition*, *19*, 147–161.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*, 441–517.