Studies in Second Language Acquisition (2024), **46**, 28–50
doi:10.1017/S0272263122000559

RESEARCH ARTICLE

Scrutinizing LLAMA D as a measure of implicit learning aptitude

Takehiro Iizuka* and Robert DeKeyser

University of Maryland, College Park, Maryland, USA
*Corresponding author. E-mail: tiizuka@terpmail.umd.edu

(Received 26 May 2022; Revised 09 December 2022; Accepted 13 December 2022)

Abstract

Since Gisela Granena's influential work, LLAMA D v2, a sound recognition subtest of LLAMA aptitude tests, has been used as a measure of implicit learning aptitude in second language acquisition research. The validity of this test, however, is little known and the results of studies with this instrument have been somewhat inconsistent. In this study, we tested the hypothesis that researchers' variable test instructions are the source of the inconsistent results. One hundred fourteen English monolinguals were randomly assigned to take LLAMA D v2 under one of three test instruction conditions. They also completed two implicit aptitude tests, three explicit aptitude tests, and a sound discrimination test. The results showed that, regardless of the type of test instructions, LLAMA D scores did not align with implicit aptitude test scores, indicating no clear evidence of the test being implicit. On the contrary, LLAMA D scores were negatively associated with scores on one implicit aptitude test, the Serial Reaction Time (SRT) task, but only in the condition where the instructions drew participants' focal attention to the stimuli. This negative association was interpreted as focal attention working against learning in the SRT task. Implicit learning aptitude may be the degree to which one is able to process input without focal attention.

Introduction

Cognitive psychologist Arthur Reber, who coined the term "implicit learning," viewed implicit learning mechanisms as a fundamental aspect of human cognition that varies minimally across individuals (A. Reber, 1967; A. Reber et al., 1991). This view has been embraced by some Second Language Acquisition (SLA) researchers as well (e.g., Krashen, 1981), leading to the implicit assumption that cognitive individual differences, if examined in SLA, are concerned with explicit learning abilities (Wen et al., 2017). Although explicit learning abilities do indeed have predictive power for successful adult SLA (see, e.g., Abrahamsson & Hyltenstam, 2008; DeKeyser, 2000), it might also be true that the complex process of SLA involves implicit as well as explicit learning, and without empirical examination we might not want to dismiss the possibility of individual differences in implicit learning. Calls for this line of research (e.g., Kaufman et al., 2010; Woltz, 2003) were answered by preliminary findings that there may be individual differences in implicit learning abilities influencing adult SLA (e.g., Granena, 2013b; Suzuki & DeKeyser, 2015).

This line of inquiry, however, is still in its infancy: We have little evidence that there is a reliable unitary construct of implicit learning aptitude (see P. Reber, 2013 for a review of relevant neuroimaging studies), and even if such a construct exists, we know little about how to measure it. The present study examined LLAMA D, a subtest of the LLAMA Language Aptitude Tests (Meara, 2005), which Granena (2013a, 2019) has proposed taps implicit language aptitude. Specifically, this study explored whether LLAMA D scores depend upon test administration procedures (i.e., test instructions) and, if they do, with which variant of test instructions the test best aligns with other measures of implicit learning abilities and is dissociated from measures of explicit learning abilities. A secondary purpose of the study was to see if currently available measures of implicit learning abilities (sequence learning and priming) are associated with each other, capturing the same underlying construct of implicit learning aptitude. Note that this study was conducted with LLAMA D version 2, and the findings reported here might not apply to version 3, a beta version of which is now available.[1]

Literature review

Defining implicit learning aptitude

As "implicit learning" is not always used consistently in the field, we will be explicit about what is meant by the term. Following the research traditions of implicit/explicit learning in cognitive psychology and psycholinguistics (e.g., DeKeyser, 2003; Jiménez, 2002; Kaufman et al., 2010; P. Reber, 2013; Rebuschat, 2013; Shanks, 2005; Williams, 2009), implicit learning was defined in this study as learning under conditions in which all the following criteria were met: (a) no intention to learn the object of learning, (b) no awareness of what is being learned (i.e., process) or the product of learning, at least at the time of learning, and (c) no focal attention to the object of learning through one's use of central executive attentional resources. Accordingly, we view "implicit learning aptitude" as the ability to learn something unintentionally, without awareness, irrespective of focal attention.

Also, we see adult second language acquisition as a process largely driven by domain-general cognition (DeKeyser, 2003), while acknowledging some language-specific aspects (see, e.g., Skehan, 2016 for the discussion of domain generality and specificity). In this article, "implicit learning aptitude" (or, more simply, "implicit aptitude") will be used for domain-general aptitude, and "implicit language (learning) aptitude" for language-specific aptitude.

Measures of implicit learning aptitude

Perhaps the most widely used measure of implicit learning aptitude to date is the implicit sequence learning paradigm, the Serial Reaction Time (SRT) task in particular.[2] In the SRT task, participants are instructed to respond as quickly and as accurately

[1] The latest version of the LLAMA tests can be accessed from Meara's website: https://www.lognostics.co.uk/tools/

[2] Some of the instruments used in the research of explicit/implicit learning are used in the research of declarative/procedural memory as well. The current study was framed by the paradigm of explicit/implicit learning (for a somewhat similar study with the framework of declarative/procedural memory, see, e.g., Buffington et al., 2021).

https://doi.org/10.1017/S0272263122000559 Published online by Cambridge University Press

as possible to the location of a stimulus that appears on a computer screen. Unbeknownst to the participants, a series of trials follows a certain regularity as to where the stimulus appears, and thus reaction time to the regular sequence decreases over the trials. As long as the task is probabilistic (as opposed to deterministic), intermixing the regular sequence with random sequences, participants are usually not aware of the regularity, and the learning is implicit (Janacsek & Nemeth, 2013; Jiménez, 2002). In fact, this type of sequence learning has been shown not to be significantly influenced by intention to learn (Jiménez et al., 1996), awareness of regularity (Cleeremans & Jiménez, 1998), or the amount of attention (Jiménez & Méndez, 1999), thereby meeting the criteria for implicit learning. Several studies have demonstrated that individual differences measured by this sequence learning paradigm predict success in adult SLA, particularly the likelihood of reaching high proficiency (Linck et al., 2013) and of developing grammatical sensitivity to subtle second language (L2) features (Granena, 2013b; Suzuki & DeKeyser, 2015), both of which presumably require a certain degree of implicit learning.

Priming is another paradigm that has been proposed to tap implicit learning aptitude (Woltz, 2003). The basic idea of priming is that performance is facilitated by a past experience, which, in an experimental situation, is usually operationalized by a reaction time difference between a primed trial (i.e., preceded by a related trial) and a control trial. Priming can largely be divided into two kinds—perceptual and conceptual priming, each of which is caused by a preceding stimulus similar either in form or meaning, respectively (see Tulving & Schacter, 1990; Woltz, 2003). This paradigm is also considered implicit on the grounds that participants have another task goal (e.g., deciding whether a stimulus is a word or nonword), priming being merely a by-product of task completion. Research indeed shows that increased memory load with a concurrent task does not affect priming effects, suggesting that priming is independent of explicit, attention-controlled recourses (e.g., Woltz & Was, 2006). Several studies have shown that individual differences measured by priming can predict success in cognitive skill acquisition in general and language acquisition in particular. Woltz (1988, 1999), for instance, reported that repetition (i.e., perceptual and conceptual) priming predicted a later stage of cognitive skill acquisition across verbal, numeric, and spatial content domains. Larkin, Woltz, Reynolds, and Clark (1996) also observed that conceptual priming was significantly associated with reading ability for sixth graders. Similarly, Was and Woltz (2007) found that the capacity for conceptual priming has a significant impact on first language (L1) listening ability in adults. Conceptual priming has been shown to be related to fluency of L2 speech production too (Granena, 2019).

Although both paradigms more or less appear to succeed in measuring individual differences relevant to some kind of implicit learning, it is too early to say that they can be a yardstick of across-the-board implicit learning aptitude. In addition to the behavioral evidence that implicit learning measures often do not correlate with one another (e.g., Buffington et al., 2021; Gebauer & Mackintosh, 2007; Godfroid & Kim, 2021; Suzuki & DeKeyser, 2017), neuroimaging studies have shown that the locus of brain activation is not consistent across paradigms for implicit learning, which is in contrast to the case of explicit learning, where the learning process relies on the medial temporal lobe memory system (P. Reber, 2013). One of the implications is that we might not want to hastily assume that implicit learning outcomes of interest can be predicted by any one implicit aptitude measure, but instead might want to target the specific implicit process of interest. To measure implicit language learning aptitude then, a language-based paradigm may be preferable. The sequence learning paradigm

https://doi.org/10.1017/S0272263122000559 Published online by Cambridge University Press

we just reviewed is not language based. The priming paradigm, particularly in the case of conceptual or semantic priming, involves language, but it only concerns the process of implicit activation of already acquired knowledge, excluding the encoding process of new knowledge. Hence, LLAMA D, which we will introduce in the next section, might hold promise for those who want to measure implicit language learning aptitude because it is a language-based test and involves both encoding and retrieval processes. Note again that, as mentioned in the introduction, the current study dealt with LLAMA D version 2 and that some of the features described in the following text may not apply to the newer version (more on this in the “Discussion” section).

LLAMA D as a measure of implicit learning aptitude

The LLAMA Language Aptitude Tests were developed by Paul Meara (2005). The tests are language-independent (see Granena, 2013a; Rogers et al., 2017 for negligible effects of test takers’ L1), and available to everyone for free, computer-delivered with automatic score calculation, making them easily accessible to a wide range of researchers. The test battery consists of four subtests, LLAMA B, D, E, and F, measuring vocabulary learning, sound recognition, sound-symbol association, and grammatical inferencing, respectively. Granena’s (2013a) exploratory validation study, using various cognitive ability tests along with LLAMA, demonstrated that there may be two different aptitude dimensions the LLAMA tests tap into, namely, *explicit* and *implicit* language learning aptitude. More specifically, her factor analysis showed that LLAMA B, E, and F loaded onto one factor, whereas LLAMA D loaded onto another factor. An additional factor analysis further revealed that LLAMA B, E, and F constituted a factor with intelligence (measured by an IQ test), while LLAMA D constituted a separate factor with implicit learning skills (measured by SRT), which was taken as evidence that LLAMA D is a measure of implicit language learning aptitude (see also Granena, 2019, for a similar finding, where LLAMA D aligned with conceptual priming). After this finding by Granena, researchers have started to use LLAMA D as a test of implicit learning aptitude, and the number of such studies has been increasing (e.g., Artieda & Muñoz, 2016; Forsberg Lundell & Sandgren, 2013; Granena, 2013b, 2016, 2019; Granena & Long, 2013; Lee, 2018; Li & Qian, 2021; Ma et al., 2018; Martens et al., 2016; Montero et al., 2018; Moorman, 2017; Mueller, 2017; Rodríguez Silva, 2017; Saito, 2017, 2019; Saito et al., 2019; Suzuki, 2021; Yalçın et al., 2016; Yalçın & Spada, 2016; Yi, 2018). The areas of these studies range from phonology (Saito et al., 2019) to collocations (Yi, 2018) to grammar (Yalçın & Spada, 2016), covering beginners (Artieda & Muñoz, 2016) as well as advanced learners (Forsberg Lundell & Sandgren, 2013). The stakes, therefore, are getting higher and higher.

However, the creator of the LLAMA tests noted that the tests have not been properly validated, and thus that “they should NOT be used in high-stakes situations” (Meara, 2005, p. 21). More than a decade later this still appears to be true (although some validation work is under way, e.g., Bokander & Bylund, 2020; Rogers et al., 2023). On top of that, the suggestion that LLAMA B, E, and F and LLAMA D tap explicit and implicit language learning aptitude, respectively, is essentially based solely on Granena’s (2013a, 2019) work, and such use of the tests is beyond the original intention of the test developer. The situation, therefore, may warrant a careful examination.

When looking into empirical studies with LLAMA D, we see rather mixed results (see Appendix S1 in Supplementary Materials for the summary). The outcomes are in the expected direction: LLAMA D scores are associated with the attainment of, for

https://doi.org/10.1017/S0272263122000559 Published online by Cambridge University Press

example, L2 collocations (Forsberg Lundell & Sandgren, 2013; Granena & Long, 2013), agreement structures (Granena, 2013b), and sound-symbolic intuitions (Mueller, 2017), all of which are assumed to be mainly the products of data-driven, implicit learning. LLAMA D scores have also been shown to predict long-term development of L2 speech production (Saito et al., 2019), which also aligns with the idea of LLAMA D as a measure of implicit learning aptitude, if we think implicit learning becomes increasingly important in later stages of SLA. Yet the outcomes are not always easy to interpret: LLAMA D scores, in some cases, predict accuracy in L2 writing (Lee, 2018) and early stages of adult foreign language learning (Artieda & Muñoz, 2016), in which explicit learning presumably should play a more important role. Also noteworthy is that, when both LLAMA tests and the SRT task are used in the same study, the correlation coefficient between LLAMA D and SRT (implicit aptitude) is smaller than the one between LLAMA D and LLAMA B, E, or F (explicit aptitude) (Granena, 2016; Yi, 2018). Furthermore, the relationships between LLAMA D and the other subtests are not consistent: There are sometimes moderate positive correlations (e.g., LLAMA D–B: $r = .35$, $p < .01$ in Yalçın & Spada, 2016; $r = .34$, $p < .05$ in Yi, 2018), while at other times no correlations were found (e.g., LLAMA D–B: $r = .05$, $p > .05$ in Saito, 2019; $r = .03$, $p > .05$ in Yalçın et al., 2016).

To make sense of these apparent inconsistencies and use LLAMA D with more confidence, we might want to explore what is taking place in the test. Overall, the test goes like this: Test-takers listen to 10 unfamiliar sound strings—computer-generated words based on a Native American language from British Columbia. After that, they move on to the test stage, where they listen to 30 sound strings one by one and indicate whether they have heard the strings or not. Unlike the other LLAMA subtests, which include a study phase (2 minutes for LLAMA B and E; 5 minutes for LLAMA F), LLAMA D only exposes test-takers to the stimuli once, all in a row, before the test, arguably placing itself at the more implicit end of the spectrum for an aptitude test (Granena, 2013a). A thing to note, though, is that, partly as a result of its language-independent nature, the test does not have any standardized test instructions, and thus how to administer it is up to individual researchers. This potential inconsistency in test administration can be critical, particularly if we intend to use the test as a measure of implicit learning aptitude, because implicitness is, as mentioned in the earlier section, very sensitive to intention, awareness, and attention. In the next section, therefore, we will explore possible variations of LLAMA D test instructions and their relevance to implicit learning.

Test instructions for LLAMA D

While the LLAMA manual (Meara, 2005) encourages researchers to create their own test instructions, it provides some ideas of how to administer the test, which are considered default instructions and which many researchers presumably use. One publication by Meara and his colleagues demonstrates this default type of instructions: "You must listen to the sound recording and it will play with [*sic*] a number of made up words. Your task is to learn and memorise as many of these words as possible" (Rogers et al., 2016, p. 209). Now, reflecting on these instructions, we notice differences with other measures of implicit learning, namely, the LLAMA D instructions entail test-takers' (a) intention to learn the stimuli, (b) awareness of what is being learned, and (c) focal attention to the stimuli, which all go against the criteria of implicit learning. Incidentally, the previously mentioned studies (Artieda & Muñoz, 2016; Lee, 2018),

https://doi.org/10.1017/S0272263122000559 Published online by Cambridge University Press

which offered alleged counterevidence to LLAMA D as a measure of implicit learning aptitude, used this type of instructions.

Granena (2013a, 2013b, 2016, 2019; Granena & Long, 2013) used a slightly different set of instructions: "You will hear a set of words in a language you are not familiar with. Your task is simply to listen carefully" (G. Granena, personal communication, April 13, 2018). As with the default instructions, focal attention is drawn to the stimuli by this type of instructions. The existence of intention and awareness, however, is less clear. Not being informed of the test phase in advance, test-takers would most likely have less intention and awareness of learning, compared with the default instructions, yet at least some participants would anticipate some kind of test, partly because, with the order of the LLAMA subtests randomized (Granena, 2013a), some of them have already taken other subtests before LLAMA D, becoming familiar with the sequence from the study/exposure phase to the test phase. In any case, many individuals are expected to try to discover patterns, if not try to memorize them, under this condition.

Yet another set of instructions was adopted by Saito (2017, 2019; Saito et al., 2019). To prevent test-takers from learning the stimuli intentionally, the researcher pretends that the exposure phase is just a sound check: They are only told to check if they can hear sound without any difficulty, which is followed by a surprise test. Their lack of intention to learn is confirmed by interview shortly after the test. Also, the test is always administered first among the entire test battery, making the "sound check" session reasonable and minimizing their anticipation of the test. Thus, the intention and awareness of learning are considered absent in this instruction condition. The existence of focal attention, however, is less clear, as some participants might pay attention to the sound strings, while others might just process them without any particular focus, only making sure the volume is okay. Table 1 summarizes the characteristics of the three types of test instructions, labeled as "memorization," "just listen," and "sound check" conditions, respectively. As seen in the table, the three types of instructions can be construed as varying from most explicit ("memorization") to least explicit ("sound check"). In fact, previous studies suggest that the correlation coefficient between LLAMA D and the clearly explicit subtests is larger when the test instructions are more explicit (e.g., LLAMA D–B: $r = .34$, $p < .05$ in Yi [2018] with "memorization" instructions; $r = .29$, $p < .05$ in Granena [2016] with "just listen" instructions) than when less explicit (e.g., LLAMA D–B: $r = .05$, $p > .05$ in Saito [2019]; $r = .05$, $p > .05$ in Suzuki [2021] both with "sound check" instructions), giving us the impression that less explicit instructions are preferable if we want to measure implicit learning aptitude. Yet, the dissociation from explicit tests alone, of course, does not guarantee that the test is tapping the construct of implicit learning. Therefore, a systematic comparison of different types of instructions coupled with various cognitive aptitude tests appears to be justified.

So far we have discussed the test instructions before the exposure phase. The instructions before the test phase may be equally as important. It is important to remember that LLAMA D is a sound recognition test, where participants decide

Table 1. Characteristics of three types of test instructions for LLAMA D

Instruction type	Memorization	Just listen	Sound check
Intention	Yes	Maybe	No
Awareness	Yes	Maybe	No
Focal attention	Yes	Yes	Maybe
Representative study	Rogers et al. (2016)	Granena (2013a)	Saito et al. (2019)

https://doi.org/10.1017/S0272263122000559 Published online by Cambridge University Press

whether each sound string was in the exposure phase or not. A recognition test in general is considered a test of explicit memory because it is accompanied by conscious memory retrieval. However, it may also be a test of implicit memory when (a) the test is forced choice, (b) the stimuli bear high similarity to one another, (c) there are no conceptual or contextual cues, and (d) the instructions discourage analytic retrieval strategies (Voss & Paller, 2009). Arguably, LLAMA D has the first three features. The fourth point is a procedural feature that researchers need to consider. Research has shown that when test instructions encourage participants to recall specific details of the learning objects, they tend to access their explicit memory, which is dependent on the amount of attention during encoding, whereas when test instructions encourage them to use a vague feeling of familiarity to approach the recognition, they tend to tap into their implicit memory, which is minimally affected by attentional resources during encoding (Mulligan, 1998; Whittlesea & Price, 2001). It has also been shown that familiarity-based recognition judgments (but not conscious recollection) are associated with conceptual implicit memory (Wang & Yonelinas, 2012). Taken together, if we were to measure implicit learning aptitude with LLAMA D, familiarity-based recognition judgments might need to be encouraged through test instructions before the test phase.

Present study

Overall research design

The literature review has made it clear that scrutiny of LLAMA D test instructions is necessary for future use of the test as a measure of implicit learning aptitude. The present study attempted to uncover the impact of test instructions by empirically comparing three instruction conditions: (a) "memorization," (b) "just listen," and (c) "sound check." Participants were randomly assigned to one of the three instruction conditions in which they took LLAMA D. The participants also completed two other relatively more established measures of implicit learning aptitude: (a) probabilistic SRT task and (b) Available Long-Term Memory (ALTM) task (i.e., conceptual priming). The participants further completed three measures of explicit learning aptitude: (a) paired associates task, (b) digit span task, and (c) Stroop task. These measures of rote learning ability and working memory were selected because of their hypothesized relevance to LLAMA D (see the next section). Phonological short-term memory and the central executive have long been held to be two components of working memory relevant to the storage and processing of verbal information (Baddeley & Hitch, 1974), which were measured by the digit span task and the Stroop task, respectively, in this study.[3] The Stroop task, a test of inhibitory control ability, was chosen because inhibitory control ability appears to be implicated in various tasks of executive functions and is viewed as the core component of executive functions (Miyake & Friedman, 2012). Additionally, the participants completed a sound discrimination task as a measure of phonetic sensitivity because previous studies suggested the potential role of sound-related ability in LLAMA D; blind people performed significantly better on LLAMA D than sighted people (Smeds, 2015); LLAMA D scores were associated with musical aptitude (Martens et al., 2016) and phonetic acuity (Drozdova et al., n.d.).

[3] The updated version of Baddeley's model of working memory also includes the episodic buffer, which is a temporary storage system that is capable of integrating information from a variety of sources (Baddeley, 2000). It is, however, yet unclear how to measure this component or how it relates to language learning.

https://doi.org/10.1017/S0272263122000559 Published online by Cambridge University Press

We included this last instrument so that the role of implicit versus explicit learning aptitude in LLAMA D could be discussed without confound with sound-related ability. In sum, LLAMA D scores obtained with three types of test instructions were examined in light of two measures of implicit learning aptitude and three measures of explicit learning aptitude, holding specific sound-related aptitude constant.

Research questions and hypotheses

Before delving into the main question about LLAMA D test instructions, this study addressed the issue of whether currently available measures of implicit learning aptitude tap the same underlying construct. Thus, the first research question was formulated as follows:

> RQ1: Are two measures of implicit learning aptitude (SRT and ALTM) substantially associated with each other?

Previous studies (e.g., Buffington et al., 2021; Gebauer & Mackintosh, 2007; Godfroid & Kim, 2021), though with different sets of instruments, have demonstrated that measures of implicit learning often do not correlate substantially with one another. Thus, we proposed the following hypothesis:

Hypothesis 1: Two measures of implicit learning aptitude (SRT and ALTM) will show no more than small correlation ($r < .30$).

The main research question concerned the impact of test instructions on LLAMA D scores in light of implicit and explicit learning aptitude:

> RQ2: Which of the three types of test instructions ("memorization," "just listen," or "sound check") best yields LLAMA D scores that align with scores on implicit learning aptitude measures (SRT and ALTM) without association with scores on explicit learning aptitude measures (paired associates, digit span, and Stroop) when controlling for phonetic sensitivity (sound discrimination)?

We predicted that the "memorization" instructions, stimulating participants' intention, awareness, and focal attention, would call upon their explicit learning aptitude, rather than implicit learning aptitude, and thus the scores would be affected by rote learning ability and working memory:

Hypothesis 2a: Under the "memorization" instruction condition, LLAMA D scores will be significantly predicted by scores for rote learning ability (paired associates) and working memory (digit span and Stroop), when controlling for phonetic sensitivity (sound discrimination).

The "just listen" instructions, encouraging participants' focal attention to the stimuli, might make working memory play a role in the test:

Hypothesis 2b: Under the "just listen" instruction condition, LLAMA D scores will be significantly predicted by scores for working memory (digit span and Stroop), when controlling for phonetic sensitivity (sound discrimination).

The "sound check" instructions minimize participants' intention and awareness of learning. Arguably, such less explicit learning condition might be conducive to

https://doi.org/10.1017/S0272263122000559 Published online by Cambridge University Press

data-driven, implicit learning (see, e.g., DeKeyser, 1995; Granena & Yilmaz, 2019; Toomer & Elgort, 2019):

Hypothesis 2c: Under the "sound check" instruction condition, LLAMA D scores will be significantly predicted by scores on implicit learning aptitude measures (SRT and ALTM), when controlling for phonetic sensitivity (sound discrimination).

Methodology

Participants

One hundred fourteen monolingual native English speakers (77 females; 18–38 years of age, $M = 20.06$, $SD = 2.69$) participated in this study for course credit or financial compensation at the University of Maryland, College Park. They were randomly assigned to one of three LLAMA D test instruction conditions. All the participants took an identical test battery, the only between-group difference being LLAMA D test instructions. Five participants were excluded because they did not follow instructions and/or were later found out to be ineligible for the study (see Appendix S2 for the eligibility criteria). The final sample size was $N = 109$ ("memorization" group, $n = 37$; "just listen" group, $n = 36$; "sound check" group, $n = 36$).

Instruments

LLAMA D

LLAMA D is a sound recognition test. The participants listened to 10 unfamiliar sound strings once, all in a row (exposure phase). Then they listened to 30 sound strings one by one, some of which were in the exposure phase while others were not, and made "old" or "new" decisions by clicking the corresponding icon on the computer screen (test phase). The stimuli were computer-synthesized sound strings based on words in a Native American language from British Columbia. The test phase was self-paced. The software calculated a score with a maximum of 75. Depending on assigned conditions, one of three test instructions was given as follows.[4]

Memorization condition. As with the original test instructions by Meara and colleagues (e.g., Rogers et al., 2016), in this condition the participants were instructed to memorize the stimuli. They were also informed that there would be a test and what the test would be like. This set of instructions, therefore, activated the participants' intention, awareness, and focal attention, along with conscious recollection. See Appendix S3 for the exact wording of this type of instructions as well as of the other two types.

Just-listen condition. Following Granena's (2013a, 2013b, 2016, 2019; Granena & Long, 2013) test instructions, in this condition the participants were instructed simply to listen to the stimuli carefully. The subsequent test was not mentioned. After the exposure phase, the participants were encouraged to adopt familiarity-based judgments for the recognition test.

[4] It should be noted that, although previous studies inspired our test instruction types, our test instructions were not identical to those in previous studies. Therefore, the results are not directly comparable. The present study neither validates nor invalidates the findings of the previous studies cited in this article.

https://doi.org/10.1017/S0272263122000559 Published online by Cambridge University Press

Sound-check condition. Following Saito's (2017, 2019; Saito et al., 2019) test instructions, the participants in this condition were (falsely) informed that the exposure phase would be just a sound check. They were only told to check if they could hear sound without any difficulty. After the exposure phase, they were encouraged to adopt familiarity-based judgments for the recognition test. Shortly after the test, the researcher confirmed through a brief interview that the participants in fact did not try to learn the stimuli during the "sound check" phase.

Probabilistic serial reaction time task

The probabilistic SRT task (Kaufman et al., 2010) was used as one of implicit learning aptitude tests. In each trial, the participants saw a stimulus appear at one of four locations on the computer screen. Their task was to press the corresponding key as quickly and as accurately as possible. Unknown to them, the sequence of stimulus appearance followed a certain pattern (1–2–1–4–3–2–4–1–3–4–2–3) 85% of the time, which was intermixed with another pattern (3–2–3–4–1–2–4–3–1–4–2–1) 15% of the time. This version of SRT task was particularly hard for the participants to learn explicitly because the probability of stimulus appearance was governed not by first-order information but by second-order information; that is, the preceding trial alone did not provide any useful information for prediction, but the most recent two trials offered such information (e.g., after 1–2, one occurred 85% of the time while four occurred 15% of the time). After an initial practice block where the two patterns occurred with equal likelihood, the participants completed eight main blocks (120 trials each) in which the sequence followed the previously mentioned differential probability. Following Granena's (2013a, 2013b, 2016) studies, learning was quantified by calculating the average reaction time difference between probable and improbable trials. Reaction time to the probable trials but not to the improbable trials was expected to decrease, and so the greater the reaction time difference (i.e., improbable – probable), the more learning.[5]

Available long-term memory task

The ALTM category task (Was et al., 2012; Was & Woltz, 2007; Woltz & Was, 2006, 2007) was used as another implicit learning aptitude test. This test probed individual differences in conceptual priming. The test consisted of two tasks: a priming task and a comparison task. In the priming task, the participants saw five words, one at a time for 2 seconds each, presented on the computer screen. Three of the words were exemplars of one category and two were exemplars of another category (e.g., *apple*, *dagger*, *banana*, *pear*, and *bomb*). The participants were then asked to indicate which of two categories had more exemplars in the word list (e.g., *Were there more weapons or more fruits?*). Following this priming task was the comparison task, where the participants saw pairs of words, a pair at a time, on the computer screen and indicated (for each pair) whether the two words were from the same or different categories as quickly and as accurately as possible by pressing one of two keys. To measure priming effects, there were two conditions for the comparison task, one in which one or both words were exemplars of one of the two categories from the preceding priming task (primed

[5] Following a reviewer's request, the reaction time difference between probable and improbable trials was examined; there was a significant difference, $t(108) = 10.35$, $p < .01$, suggesting that learning occurred in this task.

https://doi.org/10.1017/S0272263122000559 Published online by Cambridge University Press

condition), and another in which neither word was an exemplar of the categories from the preceding priming task (unprimed condition). The participants completed 18 rounds of the priming and comparison tasks after two rounds of practice. In half the rounds (nine rounds) the comparison task was with the primed condition and, in the rest of the rounds (nine rounds), with the unprimed condition. Each round included eight comparison task trials after four warm-up (unrelated) trials. The difference in the number of correct responses per minute between the primed and unprimed conditions (i.e., primed − unprimed) was used as an index of priming effects (Woltz & Was, 2006, 2007).

Paired associates task

The verbal paired associates task (Wechsler, 2009) was used as a test of rote learning ability. The participants saw a list of 14 word pairs (e.g., *way* and *body*), a pair at a time, on the computer screen. Their task was to memorize these word pairs. For each word pair, the participants saw the first word of the pair on the left side of the screen for one second, and then the second word of the pair on the right side of the screen for one second. All the 14 pairs were presented one after another with 2-second intervals between pairs. Immediately after this presentation, the participants were presented with the first word of each pair as a prompt and typed in the missing second word. Feedback (either *correct* or *incorrect*) was provided for each response. This procedure (presentation of word pairs followed by a recall test) was repeated four times. The number of correctly recalled items (max. 56) served as a score.

Digit span task

The forward digit span task (Woods et al., 2011) was used as a test of the phonological short-term memory component of working memory. In each trial, the participants saw a sequence of digits, a digit at a time for one second each, presented on the computer screen. Their task was to recall them in the order presented. One second after the presentation of the last digit, the participants entered digits by clicking on-screen buttons. This was an adaptive test, where the number of digits for the next trial increased by one when the participants answered correctly on a given trial and the number of digits decreased by one after consecutive unsuccessful trials at the same level. The task began with three digits and continued for 14 trials. The Mean Span, the number of digits for which a given participant had a 50% chance of successful recall (Woods et al., 2011), served as a score.

Stroop task

The color-word Stroop task (Stroop, 1935) was used as a test of the central executive component of working memory. In each trial, the participants saw a color word (*red*, *green*, *blue*, or *black*) or rectangle shape, presented in red, green, blue, or black—the physical color may or may not match the meaning of the word—on the computer screen. Their task was to respond to the physical color of the word (or rectangle shape) as quickly and as accurately as possible by pressing the corresponding key, ignoring the meaning of the word. In an incongruent trial where a color word was presented in another color, the participants needed to inhibit their prepotent response (i.e., responding to the meaning of the word). The task consisted of 84 trials, in which each of the four colors appeared seven times in each of the three conditions (congruent word, incongruent word, and control rectangle shape). The average reaction time

https://doi.org/10.1017/S0272263122000559 Published online by Cambridge University Press

difference between incongruent and control trials (i.e., control − incongruent) was used as an index of the participants' ability of inhibitory control.

Sound discrimination task

A sound discrimination task was used as a test of phonetic sensitivity. In each trial, the participants listened to a pair of words from languages unfamiliar to them (Russian, Chinese, and Japanese). Their task was to indicate whether the pair was the same word in the given language. The stimuli were minimal pairs, but the critical elements were not phonemic in the participants' native language, namely Russian hard/soft consonants, Chinese tones, and Japanese short/long vowels. The task consisted of 72 target trials (24 trials with each of the languages) and 24 (easier) filler trials, with equal numbers of positive (same) and negative (different) response trials. The two words to be compared in each trial were spoken by different speakers of the same gender. The Russian materials were adopted from the study by Chrabaszcz and Gor (2014), and the Chinese and Japanese materials were recorded for this study. Feedback on accuracy was provided for six trials of practice, but not for the main trials. Percent accuracy was used as a score.

Procedure

The participants completed the test battery individually with the researcher in a research lab (for the order of the tests, see Table 2).[6] All the participants took LLAMA D as their first test, regardless of their assigned test instruction conditions. This procedural decision was made, following Saito's (2017, 2019; Saito et al., 2019) studies, most importantly to make the "sound check" condition reasonable, but also to preempt presumptions potentially brought in by other tests. After LLAMA D, the participants took the other six tests. To allow for more consistent individual difference measures, the

Table 2. Order of tests

Test	Minutes
Consent form and background questionnaire	5
LLAMA D	5
Digit span	5
Sound discrimination	10
LLAMA B	5
Break	5
Available long-term memory	30
LLAMA E	5
Break	5
Paired associates	10
LLAMA F	10
Probabilistic serial reaction time	10
Stroop	5

Note: The test battery included LLAMA B, LLAMA E, and LLAMA F, the results of which are not reported here because they are not within the scope of the present article.

[6] Aside from the seven tests, the participants also completed LLAMA B, LLAMA E, and LLAMA F, the results of which are not reported in this article because they are not within the scope of the present study.

https://doi.org/10.1017/S0272263122000559 Published online by Cambridge University Press

participants completed these tests in the same order (see, e.g., Gebauer & Mackintosh, 2007; Miyake et al., 2000; Was & Woltz, 2007 for a similar practice). The entire session took about 2 hours.

Analysis

Descriptive statistics were calculated for all the tests. The test scores were coded such that greater values always indicated higher levels of the attributes. Test reliability was calculated using Cronbach's alpha. To check if randomization was successful, a series of one-way ANOVAs were conducted. Pearson's correlations among the variables were also computed. Following Plonsky and Oswald's (2014) field-specific guideline, correlation coefficients were considered small when r was around .25, medium when r was around .40, and large when r was around .60. To answer the main research question, multiple linear regression was used, where LLAMA D scores of each group were regressed on the six predictor variables. The assumptions of linearity, homoscedasticity, and normality were examined through residual plots and Q-Q plots. When assumption violations were identified, data were transformed to overcome the problem. Multicollinearity was inspected with tolerance values. These analyses were conducted using R version 4.0.3.

Results

Preliminary analysis

Descriptive statistics for all the measures are summarized in Table 3. A series of one-way ANOVAs indicated that there was no significant difference in scores across the groups for the six tests that all participants took under the same condition, meaning that random assignment successfully resulted in roughly equivalent groups. Of those measures, three (ALTM, paired associates, and Stroop) had good reliability ($\alpha \geq .74$), whereas two (SRT and sound discrimination) had low reliability ($\alpha = .41–.46$). A one-way ANOVA indicated that the effect of group was not significant for LLAMA D, $F(2, 106) = 2.64$, $p = .08$, suggesting that different test instructions did not have an impact on the level of LLAMA D scores. The reliability of LLAMA D also was low regardless of test instructions ($\alpha = .47–.58$).

Table 3. Descriptive statistics for the measures used in the study (N = 109)

Measure	Mean	SD	Min.	Max.	Reliability
LLAMA D					
Memorization condition (*n* = 37)	31.76	15.60	0	65	.47
Just listen condition (*n* = 36)	28.61	14.37	0	55	.49
Sound check condition (*n* = 36)	23.61	15.75	0	50	.58
SRT	23.31	23.51	−48.71	96.26	.41
ALTM	9.21	4.76	−1.84	21.46	.74
Paired associates	27.38	12.27	0	50	.95
Digit span	6.82	0.99	4.83	9.75	N/A[a]
Stroop	−221.58	149.84	−656.97	108.27	.95
Sound discrimination	63.43	7.06	43.10	80.60	.46

Note: SRT = probabilistic serial reaction time task; ALTM = available long-term memory task.
[a] Reliability could not be calculated for this task because it was an adaptive test.

Table 4. Correlations between LLAMA D and other measures with different LLAMA D test instructions

LLAMA D instruction type	SRT	ALTM	Paired associates	Digit span	Stroop	Sound discrim.
Memorization ($n = 37$)	.16	−.09	.32†	.09	.18	.17
Just listen ($n = 36$)	−.51*	−.03	.12	.04	−.11	−.11
Sound check ($n = 36$)	.10	.03	.04	−.26	.02	.22

Note: SRT = probabilistic serial reaction time task; ALTM = available long-term memory task; discrim. = discrimination. An asterisk (*) indicates statistical significance with the Bonferroni correction (i.e., $p < .0028$). A dagger (†) indicates marginal significance (i.e., not significant with the Bonferroni correction, but significant without the correction, $p < .05$).

Correlational analysis

Correlations among the six tests that all participants took under the same condition are summarized in Appendix S4. Regarding the first research question, two measures of implicit learning aptitude, SRT and ALTM, were not significantly correlated, $r(107) = -.09, p = .33$ (Hypothesis 1 was confirmed). The ALTM task instead showed marginally significant positive correlations with the paired associates task, $r(107) = .19$, $p = .04$, and with the Stroop task, $r(107) = .21, p = .03$.

Correlations between LLAMA D and the other measures with each variant of LLAMA D test instructions are summarized in Table 4. With “memorization” instructions, there was a small to medium marginally significant positive correlation between LLAMA D and the paired associates task, $r(35) = .32$, $p = .05$. With “just listen” instructions, there was a medium to large significant negative correlation between LLAMA D and the SRT task, $r(34) = -.51$, $p < .01$. With “sound check” instructions, LLAMA D was not correlated with any measures significantly.

Regression analysis

Based on residual plots, the assumptions of linearity and homoscedasticity did not appear to be violated. Q-Q plots, however, indicated violations of the normality assumption for the models of the two groups (“memorization” and “just listen” groups). Therefore, the data was transformed to normalize it.[7] Based on tolerance values, no multicollinearity issue was found.

Regarding the main research question, each group’s LLAMA D scores were regressed on the six predictor variables. These regression models are summarized in Table 5. In the “memorization” model, the only significant predictor was paired associates (Hypothesis 2a was partially confirmed). Its squared semipartial correlation with LLAMA D scores was .11, meaning that paired associates scores uniquely explained about 11% of the variance in LLAMA D scores. It should be noted, however, that the overall model was not significant, $R^2 = .21$, $F(6, 30) = 1.33$, $p = .27$, which suggests that this set of predictors did not account for a significant proportion of variance in LLAMA D. In the “just listen” model, the only significant predictor was SRT (Hypothesis 2b was not confirmed). Its squared semipartial correlation with LLAMA D scores was .27, meaning that SRT scores uniquely explained about 27% of the variance in LLAMA D scores. Note also that the association between these variables was in the negative direction. The overall model was significant, $R^2 = .35, F(6, 29) = 2.57, p = .04$.

[7] The data was transformed by $\sqrt{K - X}$, where X was each score and K was the largest score of $X + 1$ (Tabachnick & Fidell, 2013).

https://doi.org/10.1017/S0272263122000559 Published online by Cambridge University Press

Table 5. Summary of LLAMA D regression models

	Memorization instructions (n = 37)		Just listen instructions (n = 36)		Sound check instructions (n = 36)	
Variable	*B*	*SE*	*B*	*SE*	*B*	*SE*
SRT	−0.01	0.01	0.04*	0.01	0.13	0.13
ALTM	0.05	0.06	0.03	0.05	0.01	0.65
Paired associates	−0.05*	0.03	−0.03	0.02	0.06	0.22
Digit span	−0.12	0.32	0.12	0.21	−4.76	3.04
Stroop	0.00	0.00	0.00	0.00	0.02	0.02
Sound discrimination	0.02	0.06	0.03	0.03	0.66	0.41
R^2		.21		.35		.17
F		1.33		2.57*		0.99

Note: For normalization, the data for the "memorization" and "just listen" models were transformed by $\sqrt{K-X}$, where X was each score and K was the largest score of $X+1$. SRT = probabilistic serial reaction time task; ALTM = available long-term memory task. *p < .05.

In the "sound check" model, none of the predictors was significant (Hypothesis 2c was not confirmed), and the overall model was also not significant, $R^2 = .17$, $F(6, 29) = 0.99$, $p = .45$.

Discussion

This study was conducted to examine the validity of LLAMA D v2 (Meara, 2005) as a measure of implicit learning aptitude. To see whether and how LLAMA D scores could be influenced by test instructions, participants were randomly assigned to take the test under one of three test instruction conditions. In the text that follows we will discuss the results bearing in mind our guiding research questions and hypotheses. We will then discuss a rather unexpected finding and the future directions for the use of LLAMA D and research on implicit learning aptitude in SLA.

For the first research question, we asked whether two measures of implicit learning aptitude, SRT and ALTM, would be substantially associated with each other. As was hypothesized, and in line with other recent work (e.g., Buffington et al., 2021; Godfroid & Kim, 2021), they were not correlated substantially, $r = -.09$, $p = .33$. This lack of convergence in this study and elsewhere suggests that we have not been successful in capturing a latent construct of implicit learning aptitude. Also, the ALTM task was found to be somewhat associated with rote learning ability (paired associates) and the processing component of working memory (Stroop). Given that other studies also reported some association between this task and explicit learning aptitude measures such as the Antisaccade task (Linck et al., 2013) and the letter span task (Granena, 2019), we need to reevaluate whether the ALTM task is truly a measure of implicit learning aptitude.

For the second, and main research question, we explored the impact of test instructions on LLAMA D scores. We asked which of the three types, "memorization," "just listen," or "sound check" instructions, would be the best for the test as a measure of implicit learning aptitude. Although there was no significant group difference in the level of LLAMA D scores, the relationship between LLAMA D scores and cognitive ability test scores was found to be different across the different test instruction groups, suggesting that different cognitive abilities came into play when performing LLAMA D depending on the instructions given.

https://doi.org/10.1017/S0272263122000559 Published online by Cambridge University Press

We hypothesized that under the "memorization" instruction condition, rote learning ability and working memory would play a role in LLAMA D. This hypothesis was partially confirmed; in the regression analysis, paired associates was a significant predictor, indicating some involvement of rote learning, whereas digit span and Stroop (indices of working memory) were not significant predictors. However, the regression model was not significant and so the set of predictors did not sufficiently explain the overall variance of LLAMA D scores. Nevertheless, rote learning ability explained more than 10% of the variance; therefore, this test instruction type does not seem to be appropriate for LLAMA D as a measure of implicit learning aptitude.

Our hypothesis for the "just listen" instruction condition was that working memory would play a role in LLAMA D. This hypothesis was not confirmed; neither of the indices of working memory, digit span or Stroop, was significant in the regression model. The only significant predictor was SRT, a measure of implicit learning aptitude. The direction of this association, however, was negative (we will revisit this point later in this section). Although working memory was not involved in the test performance, LLAMA D scores were *negatively* predicted by one of the implicit learning aptitude measures; therefore, this test instruction type does not seem to be appropriate either.

We hypothesized that under the "sound check" instruction condition, LLAMA D scores would align with scores on implicit learning aptitude measures. This hypothesis was not confirmed; none of the predictors was significant in the regression model. It was not clear what kind of ability was involved in LLAMA D under this test instruction condition.

In sum, regardless of test instruction types, LLAMA D scores did not align with scores on the two implicit learning aptitude measures. The two implicit measures were not associated with each other either. Therefore, no clear evidence of a unitary construct of implicit learning aptitude was found. Despite these somewhat disappointing results, one noticeable finding is that LLAMA D scores were negatively associated with SRT scores under the "just listen" instruction condition. The SRT task explained about a quarter of variance in LLAMA D in the regression model. This strong association is intriguing given their methodological differences—verbal, auditory, accuracy-based LLAMA D, on the one hand, and nonverbal, visual, reaction-time-based SRT, on the other. In other words, this association cannot be explained by method effects but rather was driven by a certain cognitive process involved in these tasks. What set the "just listen" instructions apart from the other instruction types was that the participants' focal attention was drawn to the material without their strong intention or awareness of learning. Therefore, it would not be unreasonable to think that participants with good focusing ability (henceforth "good focusers") did well on LLAMA D under the test instruction condition, and in turn these good focusers did not do well on the SRT task. It makes sense that focal attention worked against learning in the SRT task because the most adjacent trials did not provide useful information (see the "Methodology" section) and the participants needed to process the material at a macro level to succeed in the task. Potentially, this aspect—whether one is a good focuser or not—could be a key criterion for implicit learning aptitude in individual difference research. As seen in this study and similar work by others (e.g., Gebauer & Mackintosh, 2007; Godfroid & Kim, 2021), the attempt to construe implicit learning aptitude as something that accelerates learning across domains has not been successful, and so a better way to look at implicit learning aptitude may be to see it as an ability to let go of well-developed cognitive functions, perhaps focal attention in particular, and process input as it is. In other words, implicit learning aptitude may be better seen as lack of interference rather than something that individuals have more or less of. Robinson (2005) reported that implicit learning of an artificial grammar was hindered by higher IQ, and other studies also showed that implicit learning was impeded by partial input enhancement (Toomer & Elgort, 2019) and explicit instructions (e.g., Granena & Yilmaz, 2019). All these could be interpreted as the results of focal attention being drawn to part of the learning material, thereby biasing the process of input unfavorably. This interpretation, however, is post hoc, and therefore follow-up studies are needed to (in)validate the finding.[8]

On the practical side, we now have serious reservations about the use of LLAMA D v2 as a measure of implicit learning aptitude (although this is not a criticism of the LLAMA D test and we are only commenting on this version of LLAMA D, and only on its use for assessing implicit language learning aptitude). This standpoint is based on the results of this study and a recent one by Suzuki. Suzuki (2021) modified the LLAMA D test and collected data on reaction time and confidence (i.e., how confident participants were in their judgments) as well as accuracy scores under the "sound check" instruction condition. The results showed that the participants responded faster and more accurately when they were confident, suggesting that they were using conscious knowledge on the test. In the current study we made the whole set of test instructions even more implicit by encouraging the participants to use familiarity-based judgments instead of conscious recollection (see the last paragraph of the literature review). Even with this effort, LLAMA D did not work well as an implicit test. Our recommendation, therefore, is to stop using LLAMA D as a measure of implicit aptitude and use the test as was originally intended, that is, as a test of sound recognition/listening ability (Meara, 2005; Rogers et al., 2023). This is the direction the LLAMA developer team is heading; the newer version of LLAMA D is accompanied by an explicit type of test instructions (Rogers et al., 2023). Additionally, the number of test items increased from 30 to 40 in the newer version, which should mitigate the issue of low reliability (e.g., α = .47–.58 in our study; .54 in Bokander & Bylund, 2020; .20 in Suzuki, 2021). For other changes from Version 2 to 3, see Rogers et al. (2023).

From one point of view, the lack of convergence of the implicit aptitude measures in this study supports the idea that implicit learning aptitude is multidimensional (Li & DeKeyser, 2021). As a reviewer also suggested, this could mean that, when measuring implicit language aptitude, we should use a language-based test and perhaps further narrow it down to the specific domain of interest (e.g., grammar, pronunciation)—a case in point is a study by Saito, Sun, and Tierney (2019), where implicit pronunciation-specific language aptitude was measured through assessing participants' neural encoding of speech. Examining a specific domain in this way is important. At the same time, though, if we want to discuss a cognitive *construct* of implicit learning aptitude, we eventually need to find out at least some commonalities among implicit aptitude measures. In the current study we might have found a

[8] Another way to look at this is to see this tendency to focus on patterns as a style rather than an aptitude. The ability to switch styles depending on context may be more advantageous than scoring very high on one aptitude or the other (or both!) without being able to switch (in this case between implicit and explicit learning). This is reminiscent of the literature on cognitive styles, in particular field dependence/independence, where field independence was seen as important for puzzle solving and various perceptual and motor skills, but field dependence more beneficial for smooth social interaction. See, e.g., Price (2004) and Witkin and Goodenough (1981). For examples of field (in)dependence in the SLA literature, see, e.g., DeKeyser (1984) and Johnson et al. (2000).

https://doi.org/10.1017/S0272263122000559 Published online by Cambridge University Press

starting point for exploration of such commonalities; the negative association between LLAMA D and SRT led us to speculate about interference from focal attention in the SRT task (see the preceding discussion). This is all the more interesting because the SRT task is the most promising measure of implicit aptitude (Li & DeKeyser, 2021) with predictive power already documented in SLA (e.g., Godfroid & Kim, 2021; Granena, 2013b; Linck et al., 2013; Suzuki & DeKeyser, 2015). The next step then may be to examine the SRT task further, using, for example, an eye tracker to uncover the underlying mechanisms behind the task performance and subsequently develop more instruments that require similar cognitive operations.

Lastly, the present study has raised important issues of reliability and validity. It appears that even the instruments that have been frequently used in published research are not necessarily reliable and they might not be measuring what they have been claimed to measure. A great deal more work needs to be done to validate research instruments. It is also important for researchers not to buy too hastily into what a single study has suggested (including our own).

Limitations

A couple of limitations are important to mention, one of which concerns the reliability of instruments. The reliability of the SRT task was particularly low (.41). This level of reliability is certainly not ideal, but a similar level of reliability was reported in previous studies with this instrument and it is considered standard for a measure of implicit learning (see Granena, 2016; Kaufman et al., 2010; Suzuki & DeKeyser, 2015). Because lower reliability results in attenuation of correlation, it is interesting that a strong negative correlation was found between SRT and LLAMA D in this study, despite the low reliability of these instruments. Nonetheless, replications with more reliable instruments are needed to confirm the findings of this study. We may, for example, be able to improve the reliability of the SRT task by increasing the number of blocks of trials. Also, fatigue could increase error variance, so keeping an experiment short may be another way to improve reliability (although the last two suggestions are somewhat conflicting and we need to find a good balance).

Generalizability is another thing to note. As with many other studies in the field, participants were recruited at a university; that is, the sample was drawn from well-educated individuals. Follow-up studies with different kinds of people are needed to see if the findings of this study can be generalized to the population at large (see Andringa & Godfroid, 2019 for a recent call for more diverse sampling).

Conclusion

Before summarizing the present study, a few caveats should be noted. First, the study was conducted with LLAMA version 2 (Meara, 2005) and the results should not be generalized to other versions. Second, the reliability of some instruments (SRT, sound discrimination, and LLAMA D) was low, and therefore the results should be interpreted with caution. Despite these limitations, the current study contributed to the field in some important ways, which are summarized as follows.

In this study, we examined whether test instruction types had an impact on LLAMA D as a measure of implicit learning aptitude. Although instruction types did change the relationship between LLAMA D and other cognitive test scores, regardless of the type of

https://doi.org/10.1017/S0272263122000559 Published online by Cambridge University Press

instructions, LLAMA D scores never aligned with scores on implicit learning aptitude measures, showing no evidence of the test being implicit. However, LLAMA D scores were negatively associated with scores on the SRT task, an implicit learning aptitude measure, under the test instruction condition where participants' focal attention was drawn to the learning material. We interpreted this negative association as a result of focal attention working for (in the case of LLAMA D) versus against (in the case of SRT) learning and proposed the idea that implicit learning aptitude is the degree to which one is able to let go of the tendency to look for patterns and process input without focal attention.

Acknowledgments. This article is based on the first author's qualifying paper under the supervision of the second author for the PhD program in Second Language Acquisition at the University of Maryland, College Park. We thank the committee, Kira Gor, Steven Ross, and Mike Long, for their constructive feedback. This article also benefited from comments by the Handling Editor Kazuya Saito and four anonymous reviewers. We are also grateful to Scott Barry Kaufman and Anna Chrabaszcz for sharing their research materials. Finally, many thanks to Joella (Mei) Huynh, Qi Zheng, Kaitlyn Dorman, Yanlin Peng, Kenta Kurosawa, Jason Struck, and many others, who helped us at various stages of this project.

Supplementary Materials. To view supplementary material for this article, please visit http://doi.org/10.1017/S0272263122000559.

Competing interests. The authors declare none.

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.

Andringa, S., & Godfroid, A. (2019). Call for participation: SLA for all? Reproducing second language acquisition research in non-academic samples. *Language Learning*, *69*, 5–10.

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.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*, 417–423.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. A. Bower (Ed.), *Recent advances in learning and motivation* (Vol. *8*, pp. 47–89). Academic Press.

Bokander, L., & Bylund, E. (2020). Probing the internal validity of the LLAMA language aptitude tests. *Language Learning*, *70*, 11–47.

Buffington, J., Demos, A. P., & Morgan-Short, K. (2021). The reliability and validity of procedural memory assessments used in second language acquisition research. *Studies in Second Language Acquisition*, *43*, 635–662.

Chrabaszcz, A., & Gor, K. (2014). Context effects in the processing of phonolexical ambiguity in L2. *Language Learning*, *64*, 415–455.

Cleeremans, A., & Jiménez, L. (1998). Implicit sequence learning: The truth is in the details. In M. A. Stadler & P. A. Frensch (Eds.), *Handbook of implicit learning* (pp. 323–364). Sage.

DeKeyser, R. M. (1984). The role of field independence in foreign language instruction. *ITL Review of Applied Linguistics*, *63*, 1–21.

DeKeyser, R. M. (1995). Learning second language grammar rules: An experiment with a miniature linguistic system. *Studies in Second Language Acquisition*, *17*, 379–410.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, *22*, 499–533.

DeKeyser, R. M. (2003). Implicit and explicit learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 313–348). Blackwell.

Drozdova, P., van Hout, R., & Scharenborg, O. (n.d.). *Do noise and linguistic skills influence lexically-guided perceptual learning?* http://www.inspire-itn.eu/files/spire2016/02-Drozdova.pdf

https://doi.org/10.1017/S0272263122000559 Published online by Cambridge University Press

Forsberg Lundell, F., & Sandgren, M. (2013). High-level proficiency in late L2 acquisition: Relationships between collocational production, language aptitude and personality. In G. Granena & M. H. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 231–256). John Benjamins.

Gebauer, G. F., & Mackintosh, N. J. (2007). Psychometric intelligence dissociates implicit and explicit learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 34–54.

Godfroid, A., & Kim, K. M. (2021). The contributions of implicit-statistical learning aptitude to implicit second-language knowledge. *Studies in Second Language Acquisition*, *43*, 606–634.

Granena, G. (2013a). Cognitive aptitudes for L2 learning and the LLAMA Language Aptitude Test. In G. Granena & M. H. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 105–129). 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. (2016). Cognitive aptitudes for implicit and explicit learning and information-processing styles: An individual differences study. *Applied Psycholinguistics*, *37*, 577–600.

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., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research*, *29*, 311–343.

Granena, G., & Yilmaz, Y. (2019). Corrective feedback and the role of implicit sequence-learning ability in L2 online performance. *Language Learning*, *69*, 127–156.

Janacsek, K., & Nemeth, D. (2013). Implicit sequence learning and working memory: Correlated or complicated? *Cortex*, *49*, 2001–2006.

Jiménez, L. (2002). Intention, attention, and consciousness in probabilistic sequence learning. In L. Jiménez (Ed.), *Attention and implicit learning* (pp. 43–68). John Benjamins.

Jiménez, L., & Méndez, C. (1999). Which attention is needed for implicit sequence learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 236–259.

Jiménez, L., Méndez, C., & Cleeremans, A. (1996). Comparing direct and indirect measures of sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 948–969.

Johnson, J., Prior, S., & Artuso, M. (2000). Field dependence as a factor in second language communicative production. *Language Learning*, *50*, 529–567.

Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition*, *116*, 321–340.

Krashen, S. (1981). Aptitude and attitude in relation to second language acquisition and learning. In K. C. Diller (Ed.), *Individual differences and universals in language learning aptitude* (pp. 155–175). Newbury House.

Larkin, A. A., Woltz, D. J., Reynolds, R. E., & Clark, E. (1996). Conceptual priming differences and reading ability. *Contemporary Educational Psychology*, *21*, 279–303.

Lee, J. (2018). *The interactive effects of task complexity, task condition, and cognitive individual differences on L2 writing* (Doctoral dissertation). University of Maryland, College Park, MD. https://drum.lib.umd.edu/handle/1903/21755

Li, S., & DeKeyser, R. (2021). Implicit language aptitude: Conceptualizing the construct, validating the measures, and examining the evidence. *Studies in Second Language Acquisition*, *43*, 473–497.

Li, S., & Qian, J. (2021). Exploring syntactic priming as a measure of implicit language aptitude. *Studies in Second Language Acquisition*, *43*, 574–605.

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

Ma, D., Yao, T., & Zhang, H. (2018). The effect of third language learning on language aptitude among English-major students in China. *Journal of Multilingual and Multicultural Development*, *39*, 590–601.

Martens, P., Nakatsukasa, K., & Percival, H. (2016). Music training correlates with visual but not phonological foreign language learning skills. *Proceedings of 14th International Conference on Music Perception and Cognition*, 352–354.

Meara, P. M. (2005). *Llama Language Aptitude Tests: The manual.* Lognostics.

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, *21*, 8–14.

https://doi.org/10.1017/S0272263122000559 Published online by Cambridge University Press

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49–100.

Montero, F., Donate, A., Dixon, D., & Long, M. H. (2018, February). *Language aptitudes and L2 proficiency in Spanish noun gender assignment*. Paper presented at Evolving Perspectives on Advancedness: A Symposium on Second Language Spanish, University of Minnesota-Twin Cities, Minneapolis, MN.

Moorman, C. M. (2017). *Individual differences and linguistic factors in the development of mid vowels in L2 Spanish learners: A longitudinal study* (Doctoral dissertation). Georgetown University, Washington, DC. https://repository.library.georgetown.edu/handle/10822/1047824

Mueller, J. (2017). *An examination of the influence of age on L2 acquisition of English sound-symbolic patterns* (Doctoral dissertation). University of Maryland, College Park, MD. https://drum.lib.umd.edu/handle/1903/20315

Mulligan, N. W. (1998). The role of attention during encoding in implicit and explicit memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 27–47.

Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.

Price, L. (2004). Individual differences in learning: Cognitive control, cognitive style, and learning style. *Journal of Educational Psychology*, *24*, 681–698.

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

Reber, A. S., Walkenfeld, F., & Hernstadt, R. (1991). Implicit and explicit learning: Individual differences and IQ. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 888–896.

Reber, P. J. (2013). The neural basis of implicit learning and memory: A review of neuropsychological and neuroimaging research. *Neuropsychologia*, *51*, 2026–2042.

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

Robinson, P. (2005). Cognitive abilities, chunk-strength, and frequency effects in implicit artificial grammar and incidental L2 learning: Replications of Reber, Walkenfeld, and Hernstadt (1991) and Knowlton and Squire (1996) and their relevance for SLA. *Studies in Second Language Acquisition*, *27*, 235–268.

Rodríguez Silva, L. H. (2017). *The role of cognitive individual differences and learning difficulty in instructed adults’ explicit and implicit knowledge of selected L2 grammar points: A study with Mexican learners of English* (Doctoral dissertation). University of Essex, Essex, UK. http://repository.essex.ac.uk/id/eprint/20626

Rogers, V., Meara, P., Aspinall, R., Fallon, L., Goss, T., Keey, E., & Thomas, R. (2016). Testing aptitude: Investigating Meara’s (2005) LLAMA tests. In S. A. Liszka, P. Leclercq, M. Tellier, & G. D. Véronique (Eds.), *EUROSLA Yearbook 16* (pp. 179–210). John Benjamins.

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

Rogers, V., Meara, P., & Rogers, B. (2023). Testing language aptitude: LLAMA evolution and refinement. In Z. E. Wen, P. Skehan, & R. L. Sparks (Eds.), *Language aptitude theory and practice*. Cambridge University Press.

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.

Saito, K. (2019). The role of aptitude in second language segmental learning: The case of Japanese learners’ English /ɹ/ pronunciation attainment in classroom settings. *Applied Psycholinguistics*, *40*, 183–204.

Saito, K., Sun, H., & Tierney, A. (2019). Explicit and implicit aptitude effects on second language speech learning: Scrutinizing segmental and suprasegmental sensitivity and performance via behavioural and neurophysiological measures. *Bilingualism: Language and Cognition*, *22*, 1123–1140.

Saito, K., Suzukida, Y., & Sun, H. (2019). Aptitude, experience, and second language pronunciation proficiency development in classroom settings: A longitudinal study. *Studies in Second Language Acquisition*, *41*, 201–225.

Shanks, D. R. (2005). Implicit learning. In K. Lamberts & R. Goldstone (Eds.), *Handbook of cognition* (pp. 202–220). Sage.

Skehan, P. (2016). Foreign language aptitude, acquisitional sequences, and psycholinguistic processes. In G. Granena, D. O. Jackson, & Y. Yilmaz (Eds.), *Cognitive individual differences in second language processing and acquisition* (pp. 17–40). John Benjamins.

Smeds, H. (2015). Blindness and second language acquisition: Studies of cognitive advantages in blind L1 and L2 speakers *(Doctoral dissertation)*. Stockholm University, Stockholm, Sweden. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A790294&dswid=6124

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.

Suzuki, Y. (2021). Probing the construct validity of LLAMA_D as a measure of implicit learning aptitude: Incidental instructions, confidence ratings, and reaction time. *Studies in Second Language Acquisition*, *43*, 663–676.

Suzuki, Y., & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language Learning*, *65*, 860–895.

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

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

Toomer, M., & Elgort, I. (2019). The development of implicit and explicit knowledge of collocations: A conceptual replication and extension of Sonbul and Schmitt (2013). *Language Learning*, *69*, 405–439.

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science*, *247*, 301–306.

Voss, J. L., & Paller, K. A. (2009). An electrophysiological signature of unconscious recognition memory. *Nature Neuroscience*, *12*, 349–355.

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.

Was, C. A., Dunlosky, J., Bailey, H., & Rawson, K. A. (2012). The unique contributions of the facilitation of procedural memory and working memory to individual differences in intelligence. *Acta Psychologica*, *139*, 425–433.

Was, C. A., & Woltz, D. J. (2007). Reexamining the relationship between working memory and comprehension: The role of available long-term memory. *Journal of Memory and Language*, *56*, 86–102.

Wechsler, D. (2009). *Wechsler Memory Scale-Fourth Edition (WMS-IV): Technical and interpretive manual*. Pearson.

Wen, Z. E., Biedroń, A., & Skehan, P. (2017). Foreign language aptitude theory: Yesterday, today and tomorrow. *Language Teaching*, *50*, 1–31.

Whittlesea, B. W., & Price, J. R. (2001). Implicit/explicit memory versus analytic/nonanalytic processing: Rethinking the mere exposure effect. *Memory & Cognition*, *29*, 234–246.

Williams, J. N. (2009). Implicit learning in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *The new handbook of second language acquisition* (pp. 319–353). Emerald Group Publishing.

Witkin, H. A., & Goodenough, D. R. (1981). *Cognitive styles: Essence and origins. Field dependence and field independence*. International Universities Press.

Woltz, D. J. (1988). An investigation of the role of working memory in procedural skill acquisition. *Journal of Experimental Psychology: General*, *117*, 319–331.

Woltz, D. J. (1999). Individual differences in priming: The roles of implicit facilitation from prior processing. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants* (pp. 135–156). American Psychological Association.

Woltz, D. J. (2003). Implicit cognitive processes as aptitudes for learning. *Educational Psychologist*, *38*, 95–104.

Woltz, D. J., & Was, C. A. (2006). Availability of related long-term memory during and after attention focus in working memory. *Memory & Cognition*, *34*, 668–684.

Woltz, D. J., & Was, C. A. (2007). Available but unattended conceptual information in working memory: Temporarily active semantic content or persistent memory for prior operations? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 155–168.

Woods, D. L., Kishiyama, M. M., Yund, E. W., Herron, T. J., Edwards, B., Poliva, O., Hink, R. F., & Reed, B. (2011). Improving digit span assessment of short-term verbal memory. *Journal of Clinical and Experimental Neuropsychology*, *33*, 101–111.

https://doi.org/10.1017/S0272263122000559 Published online by Cambridge University Press

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

Yalçın, Ş., & Spada, N. (2016). Language aptitude and grammatical difficulty: An EFL classroom-based study. *Studies in Second Language Acquisition*, *38*, 239–263.

Yi, W. (2018). Statistical sensitivity, cognitive aptitudes, and processing of collocations. *Studies in Second Language Acquisition*, *40*, 831–856.

Cite this article: Iizuka, T. and DeKeyser, R. (2024). Scrutinizing LLAMA D as a measure of implicit learning aptitude. *Studies in Second Language Acquisition,* 46: 28–50. https://doi.org/10.1017/S0272263122000559

https://doi.org/10.1017/S0272263122000559 Published online by Cambridge University Press