L2 (English) spelling is positively correlated with L2 vocabulary size, but not with a measure that putatively taps visual statistical learning

Authors: Matthew H.C. Mak¹, Chen Qiu², Kathy K.M. Shum²

¹ University of Oxford; matthew.mak@psy.ox.ac.uk

² The University of Hong Kong

Abstract

Recent studies showed that a person's visual statistical learning (VSL) ability is positively correlated with L1 and L2 literacy development. That is, people who are more adept at detecting implicit regularities in visual inputs tend to be or to become better readers. Most if not all studies that examined this link looked only into reading ability, oddly neglecting spelling, and importantly, they all relied on relatively small sample sizes. In light of these, we recruited a relatively large sample (64 advanced English learners) to investigate if performance on a commonly adopted VSL task (i.e., embedded triplet task) correlates with L2 (English) spelling proficiency, as indexed by a spelling test involving highly rare regular English words. We also administered a rote memorization task and an L2 vocabulary size test. In line with a growing body of literature, we found no evidence that performance on the VSL task correlated with L2 spelling. We did however find clear evidence that L2 learners with a larger L2 vocabulary size are also better spellers. We argue that while SL is linked to literacy development, traditional VSL measures may not be suitable for the investigation of individual differences owing to poor sensitivity. Our study, alongside Mak (2016), also demonstrates that false positives may arise due to small sample sizes. We, therefore, urge researchers worldwide, especially those interested in individual differences, to work towards reproducibility and better science by for example avoiding unjustifiably small sample sizes.

1. Background

Learning to read can be interpreted as learning the statistical regularities that govern a written language (Nation & Mak, in press). To become literate in an alphabetic language such as English, an individual must possess knowledge of at least two types of regularities: (1) the mapping between graphemes and phonemes and (2) the frequency of permissible letter combinations.

For the first type of regularity, take the letter ‘c’ in English as an example, it is most often mapped to the phoneme /k/ (as in *cat*), but it can be associated with other phonemes depending on its neighbors: it is pronounced as /s/ when followed by an ‘e’ (as in *ceiling*) or an ‘i’ (as in *city*) and is pronounced as /tʃ/ when followed by an ‘h’ (as in *cheat*). To master written English, one must become sensitive to contextual cues such as letter sequences, which provide cues concerning how a grapheme or a set of graphemes is pronounced.

Turning to the second type of regularity, each language permits a different set of letter combinations. For example, ‘izq’ is permitted in Spanish, but not in English. And some letter sequences are more frequent than others, for example, ‘qu’ is markedly more frequent than ‘qi’ in English. Mastering such statistical frequencies is believed to be essential to skilled reading and spelling (Arciuli & Simpson, 2012).

What so special about these statistical regularities is that they are not always transmitted in an explicit manner; instead, they are usually implicitly and incrementally acquired through print exposure (Castles, Rastle & Nation, 2018; Nation & Castles, 2017). In other words, an individual can become a proficient reader even without explicit understanding of such statistical knowledge (Treiman & Kessler, 2005).

In light of these, it has been proposed that individuals who are more adept at detecting and extracting implicit regularities in perceptual inputs are more likely to be better readers than someone who are very insensitive to perceptual regularities (Elleman, Steacy & Compton,

2019). This proposition entails that poor reading and developmental dyslexia may be a consequence of an inability to detect and extract implicit regularities in written languages.

In the psychology literature, this process of detecting and extracting perceptual regularities is usually referred to as *statistical learning* (hereafter SL), which is often seen as a domain-general cognitive construct that is independent of general intelligence, executive functions, and verbal/phonological working memory (Arciuli & Simpson, 2012; Kaufman et al., 2010; Kidd & Arciuli, 2016; Siegelman & Frost, 2015). If this hypothetical SL construct is indeed causally related to reading development, one would expect a positive correlation between reading proficiency and performance on tasks that putatively measure SL ability. Several studies have tested this prediction.

One of the earliest studies is Arciuli and Simpson (2012), in which 38 typically developing children and 37 healthy adults were exposed to a visual stream made up of images that were covertly sequenced (a visual SL task known as the embedded triplet paradigm; Fiser & Aslin, 2002). Participants were then assessed on their awareness of the covert sequences embedded. Importantly, after partialling out the effect of age and attention, correlational analysis revealed that performance on this embedded triplet paradigm was positively correlated with higher print-word decoding ability in both children ($r = .327, p < .05$) and adults ($r = .338, p < .05$). This finding was interpreted as indicating that “a capacity for more effective SL is related to higher (L1) reading ability in the general population” (Arciuli & Simpson, 2012).

Comparable result has also been reported by Frost et al. (2013). In their study, 27 native speakers of English, who were learning Hebrew as an L2, completed an embedded triplet task as well as a number of Hebrew reading tests (e.g., non-word decoding). After a year of learning Hebrew, the participants completed another set of Hebrew reading tests. Those who were more capable of detecting and identifying implicit sequences in the triplet task were more likely to become better Hebrew readers a year later ($r = 0.43-0.57, p < .05$). This study provided direct

evidence for the very first time that L2 literacy acquisition is related to a domain-general ability to detect regularities in visual inputs.

Interestingly, all previous research studies (e.g., Arciuli & Simpson, 2012; Schmalz, Moll, Mulatti & Schulte-Korne, 2019) that examined the link between Visual Statistical Learning (VSL) ability and literacy proficiency looked only into decoding ability, oddly neglecting spelling. Noting the fact that reading and spelling are two core components of literacy and dissociable to a certain extent (Torppa, Georgiou, Niemi, Lerkkanen, & Poikkeus, 2016), Mak (2016) set out to investigate if performance on a VSL task is correlated with spelling proficiency in a second language (L2). In the study, 26 native speakers of Cantonese, who were highly proficient English speakers, completed a modified embedded triplet task and an English spelling test that involved highly rare regular English words (e.g., *abomasum*). Surprisingly, a negative correlation was found ($r = -.537$, $p = .047$; for scatterplot, see Figure 1); that is, the more proficient a speller is, the poorer their performance on the embedded triplet task.

Two explanations to this surprising result were considered. First, the participants, all of whom are native Cantonese speakers, may have been more predisposed towards using rote memorization (aka explicit learning) when it comes to learning English spelling due to the logographic nature of Chinese characters and to the fact that Hong Kong's education system equates explicit learning with academic success (Au & Noel, 1999). Good spellers, relative to less proficient spellers, may have relied more on explicit learning in the process of learning to spell in English. As proposed by Ullman (2004), a trade-off relation may exist between explicit and implicit learning (i.e., a see-saw effect). If such a proposed relation indeed exists, that will explain why the better spellers performed more poorly on the implicit VSL task in Mak (2016).

Second, while Frost et al. (2013) looked into the role of VSL in *early* L2 literacy, Mak (2016) looked into the role of VSL in *advanced* L2 literacy. The fact that the same positive correlation was not observed in Mak (2016) hinted at the possibility that the centrality of VSL in L2 literacy acquisition may diminish as L2 proficiency increases. Mak (2016) then put forth the proposal that VSL is a core component in L2 language aptitude, whose predictive power of individual differences in language outcome diminishes as a function of L2 proficiency

Figure 1. Scatterplot showing the relationship between L2 spelling proficiency and performance on the embedded triplet task in Mak (2016)

2. The Current Study

Using a similar experimental design as Mak (2016), the current study investigated if performance on a non-linguistic Visual Statistical Learning (VSL) task (i.e., embedded triplet task) correlates with individual differences in L2 (English) spelling proficiency.

One of the major weaknesses of Mak (2016) and in fact most previous studies (e.g., Arciuli & Simpson, 2012; Frost et al., 2013) is that they all relied on fairly small sample sizes [Mak (2016): $N = 26$; Frost et al. (2013): $N = 27$]. In light of this, we recruited a much larger

sample size ($N = 64$). This sample size, well-justified by power analysis (documented in Appendix 1), gives us approximately 95% statistical power.

Another major difference between Mak (2016) and the current study was that we included an explicit learning task (a modified version of the Modern Language Aptitude Test-Part V) that involved rote memorization of English-Maya word pairs. The inclusion of such a test was to help us evaluate whether a trade-off relation indeed exists between explicit and implicit learning and whether explicit learning ability correlates with individual differences in L2 spelling proficiency.

3. Methods

3.1. Participants

A total of 64 psychology undergraduates from the University of Hong Kong participated in the study (43 females). Three were excluded from data analysis due to equipment failure; the final sample, therefore, encompassed 61 participants. All of them are highly skilled English learners who have achieved a minimum of 6.5¹ on the International English Language Test System (IELTS) or equivalent. All of them are native speakers of Cantonese and are sequential bilinguals who began to learn English after 3 years of age. They also use English on a daily basis as the medium of instruction at the University of Hong Kong is English. Their mean age was 19.1 (SD: 1.98). All participants have normal or correct-to-normal vision, and have no known history of speech, reading or hearing disability, ASD, or ADHD. The participants received partial course credits for their participation in the study. This study adhered to basic ethical considerations for the protection of human participants in research and was approved by the Human Research Ethics Committee at the University of Hong Kong. Written consent was obtained from the participants before the experiments began.

¹ A 6.5 on IELTS is indicative of a “Competent English User”. The highest possible score on IELTS is 9.0, which is indicative of an “Expert English User”.

3.2. Experimental Tests

3.2.1. L2 (English) Spelling Test

Participants spelt out 40 highly rare English words (e.g., *papaveraceous*) in the spelling test. In one trial, the participants listened to an audio clip containing one English vocabulary item. They could, in theory, replay the audio clip as many times as they wished; nevertheless, only 30 seconds were given within which the participants had to listen to the audio *and* spell out the word. A timer (counting down from 30s) was visible to the participants in each trial. After 30 seconds, next trial began automatically, regardless of whether an answer was given. There was one practice trial prior to the actual test, and the participants were required to wear a headphone during the test. This test was administered using Qualtrics on a desktop, with the presence of a research assistant.

Test items. The selection of the 40 vocabulary items were aided by three native speakers of English (two British and one American) who rated their level of familiarity with 55 rare English vocabulary items (0–*Not At All*; 1–*I feel like I know this word*; 2–*Quite sure I know this word*; 3–*Certain that I know this word*). The 40 items with the lowest level of familiarity were chosen for the spelling test (mean: 0.18; range 0-1).

These lexical items are almost non-existent in typical daily discourse; examples include *abomasum* and *papaveraceous* (for a complete list, see Appendix 2). The purpose of using such rare regular English words was to force the participants to spell using their knowledge of the statistical regularities between phonemes and graphemes as well as of letter permissibility. Such knowledge, as laid out in Introduction, is believed to rely on a domain-general statistical learning mechanism.

Scoring. As a matter of fact, when asked to spell /,di:glu:'tɪfən/, spellers who rely solely on the statistical regularities between phoneme and grapheme may produce multiple

phonologically plausible spellings such as *deglutition* and **deglutician*. In other words, a participant who spelt *deglutition* as **deglutician* was appropriately applying one of the few acceptable phoneme-grapheme translation rules and correctly using their knowledge of letter permissibility. Given this experiment was designed to seek any potential association between spelling using acquired statistical information and VSL, participants should be awarded some points for **deglutician* even though it is not a lawful lexical item.

In light of this, a pilot study was conducted in which five native English speakers (3 British and 2 American) gave all the phonologically plausible spellings of the 40 test items. Afterwards, these responses were passed onto another four native English speakers (all postgraduate students at Cambridge University) who judged the degree of “orthographic appropriateness” of each plausible spelling.

For example, the word *amphibology*, according to the five native English speakers, can be alternatively spelt **amfibology* based on sound-to-letter regularities alone; this variant spelling was then rated by four other native English speakers on the degree to which they think this particular spelling conforms to English orthographies (0–*Not At All*; 1–*50% English-like*; 2–*100% English-like*). The average rating on orthographic appropriateness was then used to compute a score that a plausible spelling is worth (see Table 1). For example, the mean rating of orthographic appropriateness for **amfibology* is 0.25, so if a participant in the actual spelling test gives **amfibology* for *amphibology*, their answer would be translated into 0.5 point. On the other hand, if he or she is able to produce *amphibology*, 2 points (which is the maximum point for one test item) will be awarded.

Table 1. Points to be awarded to an alternative spelling based on orthographic appropriateness

Mean rating of orthographic appropriateness	Points to be awarded
1.67 – 2	2 191
1.33 – 1.66	1.5
1 – 1.32	1 192
0.25 – 0.99	0.5
0 – 0.24	0

Audio clips. Each audio clip contained one vocabulary item, and hence, there were 40 audio clips. Each vocabulary item was repeated once in the clip (i.e., two identical tokens in one clip). Between the two tokens was a 1-second silence. The audio was taken from Oxford Dictionary (<https://en.oxforddictionaries.com/>). Each test item was digitized in an individual sound file at a sampling rate of 44.1kHz. The duration of one clip ranges from 2200ms to 2800ms.

3.2.2. VSL Task

There are two major types of VSL task: Visual *Sequential* and Visual *Spatial* (Yu, 2016). Yu (2016) showed that L2 reading correlated with Visual Sequential task more strongly than Visual Spatial task, so the current study adopted the Sequential version. In addition, the stimuli in a VSL task can be either non-linguistic (e.g., pictures) or linguistic (e.g., words), but most relevant studies (e.g., Arciuli & Simpson, 2012, Frost et al., 2013) employed only non-linguistic stimuli, so the current VSL task followed in their footsteps. The design of the current VSL task was similar to Frost et al. (2013), which comprised 2 phases: an exposure phase and a test phase.

Exposure phase. The participants were instructed at the beginning that all they had to do was to pay full attention to a visual stream. They were told to watch the video without doing anything else. Prior to viewing the actual visual stream, each participant was presented with a 10-second visual stream that resembled the actual one; this was to give the participants a sense of what to expect.

Visual stream. Eighteen arbitrary black shapes were taken and modified from Frost et al. (2013) (see Figure 2). Each shape was shown one at a time, against a white background on a computer screen (see Figure 3). Following Arciuli & Simpson (2012b), each shape was

presented for 400ms, with an interstimulus interval of 200ms. The 18 shapes were organized in 6 triplets (referred to as ABC, DEF, GHI, JKL, MNO, and PQR). Within one triplet, the transitional probabilities (TPs) between each shape were $p = 1.0$, such that, for triplet ABC, A is always followed by B, and B always appears before C. No reference was made to participants about such predictability. The triplets were presented in a pseudorandom order, with three constraints:

1. A particular triplet (e.g., T_1) could not appear twice in immediate succession (i.e., ... $T_1 T_1$... is not allowed).
2. No repeated pairs of triplets were allowed (i.e., ... $T_1 T_2 T_1 T_2$... is not allowed).
3. A triplet and the next identical triplet must not be separated by more than 10 triplets.

Each triplet appeared 17 times in the visual stream, giving a total of 102 triplets.

Figure 2. The 18 shapes (6 triplets) in the VSL task

Figure 3. Sample exposure stream using triplets ABC & GHI.

Cover task in the visual stream. As a cover task, participants were told before the exposure phase that something “out-of-place” was embedded in the visual stream and that they must memorize what they were. A picture of a red flower, yellow bananas, and a dog appeared at 1:00, 2:00, and 3:00 timestamp of the visual stream, respectively, for 200ms.² The three objects stood out from the rest of the stimuli as all other stimuli were black meaningless figures; in other words, the three real-life objects should leave a lasting impression on the participants. Immediately following the exposure phase, the participants were asked to indicate the real-life objects they saw in three multiple-choice questions. This was to ensure that the participants had paid sufficient attention to the visual stream. If the participants failed to give the correct answers, it implied that they had not attended sufficiently to the visual stream, and hence, their data would be discarded. Encouragingly, all participants were able to indicate what the three real-life objects were, suggesting that they had paid at least some attention to the visual stream.

² These real-life objects appeared in between triplets, e.g., ABC [flower] DEF.

Test phase. In Mak (2016) and previous studies (Frost et al., 2013), the exposure phase was followed by a surprise two-alternative forced-choice task as a way to measure participants' implicit awareness of the covert sequences. In this task, participants discriminated between a seen and an unseen sequence. As informed by Schmalz et al. (2019), the two-alternative forced-choice task used in Arciuli & Simpson (2012) or Frost et al. (2013) is very poor at generating sufficient variability, preventing any meaningful correlations from arising. This observation, therefore, prompted us to adopt a slightly different test to tap statistical learning.

In the test phase, participants were presented with only one sequence each time: either one that appeared in the visual stream (namely *possible sequence*) or one that did not (namely *impossible sequence*), and the participants had to judge on a scale whether they thought they had seen it during the viewing phase. Participants were given 25 seconds to choose, using their gut instincts, from one of the following four options upon seeing a sequence:

1. Didn't appear: Very certain
2. Didn't appear: Slightly certain
3. Appeared: Slightly certain
4. Appeared: Very certain

With this task, we generated two VSL scores for each participant: **(a)** one without taking the confidence rating into account (as in previous studies), **(b)** another taking the confidence rating into account.

For **(a)**, in trial where a possible sequence was shown, options 3 and 4 are both worth one point, while the other two options are worth zero point. On the other hand, in trials where an impossible sequence was shown, options 1 and 2 are both worth one point, where the other two options are worth zero point.

As for **(b)**, in trials where a possible sequence was shown, options 3 and 4 are worth one and two points respectively, while the other two options are worth zero point. On the other

hand, in trials where an impossible sequence was shown, options 2 and 1 are worth one and two points respectively, while the other two options are worth zero point. The reason for introducing **(b)** is that we wished to increase the chance of detecting sufficient variability in this task, and no previous studies, to the best of our knowledge, have tested whether taking confidence rating into account improves the predictive power of the current VSL paradigm.

Participants were encouraged to use the entire scale. It is inferred that the higher score a task tester gets in this VSL task, the more efficient their cognitive capacity for learning the implicit transitional probabilities among the visual shapes in the stream (Schlichting et al., 2017).

In order to further increase our chance of detecting meaningful individual differences, a test sequence contained only two, instead of three, shapes. The advantage of using two shapes (i.e., duplets) over triplets in the test phase is that more unique test trials could then be created. Below, we explain the specificities of these duplets.

Within one possible triplet (e.g., ABC), there are two duplets (e.g., AB & BC) within which the TPs were $p=1$. Given six possible triplets, there are 12 possible duplets (AB, BC, DE, EF, GH, HI, JK, KL, MN, NO, PQ, and QR).

Twelve new duplets were also created by using the original black figures (namely, AH, BF, DK, EI, GN, HL, JQ, KO, MB, NR, PE, and QC). These duplets, dubbed as the *impossible duplets*, did not appear in the visual stream, and hence the transitional probabilities within each triplet are zero.

The duplets were presented with the identical presentation time and interstimulus interval used in the exposure phase. Each possible and impossible sequence was shown twice in the test phase, hence there were a total of 48 trials.³ The order of trial presentation was randomized for each participant.

³ Previous studies had few test trials, for example Frost et al. (2013) only had 32 trials.

3.2.3. Explicit learning task – Modern Language Aptitude Test, Part V (MLAT-V)

On the task, participants were given two minutes to memorize the English translation of 24 Maya words. Following the memorization phase was a recognition test made up of 20 multiple-choice questions, which asked participants to match the Maya words to their corresponding English meaning. The participants were given five minutes to complete the task.

3.2.4. Vocabulary size test

To measure whether L2 spelling proficiency is correlated with how many L2 word items a learner knows, we administered the vocabulary size test designed by Nation & Beglar (2007). This paper-and-pencil test involved 50 multiple-choice questions in which participants decided which of the four supplied meanings best suits a vocabulary item. An example is shown below:

puree: This is bright green <puree>.

- a. fruit or vegetables in liquid form
- b. dress worn by women in India
- c. skin of a fruit
- d. very thin material for evening dresses

In the original test, there were 100 vocabulary items, ranging from very basic words (e.g., *see*) to very obscure words (e.g., *casuist*). Since all participants were proficient English speakers, it was expected that they know all the rudimentary items. Therefore, these relatively basic items ($N = 50$) were removed, leaving only the more advanced items ($N = 50$) on the test. One of the major reasons for this abridged version was that we would like to minimize the amount of fatigue experienced by the test takers. This test was untimed; participants could take as long as they wished to complete the task. Participants were encouraged to pick an answer only if they felt confident about the meaning of a word; otherwise, they should avoid making wild guesses. One score is awarded for every correct answer (max: 50). It is inferred that the higher the score, the larger the person's English vocabulary size.

329 **3.3. Procedure**

330 The procedure of the current study is summarized in Figure 4. The whole experiment required
331 approximately one hour.

332
333 **Figure 4.** Experimental procedure
334
335
336
337
338
339
340
341
342
343
344
345

4. Results

Table 2 shows the overall descriptive statistics of the four experimental tasks. We first computed a one-sample *t* test to determine if the participants' performance on the VSL task is greater than chance. Despite the current VSL test being slightly different from previous studies (e.g., Frost et al., 2013), participants' performance well exceeded chance level ($t(60) = 6.70, p < .001$), indicative of VSL being at work.

Table 2. Descriptive statistics of the four experimental tasks

	VSL Task (Max = 48) [^]	VSL Task (Max = 96) [*]	Spelling Test (Max = 80)	Explicit Learning Task (Max = 20)	Vocabulary Size Test (Max = 50)
Chance Level	24	36	NA	5	12.5
Mean	28.1	37.7	34.2	14.2	19.79
Median	28	38	37	15	17
Range	18-40	20-58	8-60.5	2-20	7-42
SD	4.8	8.5	13.4	3.9	8.8
Greater than chance level?	Yes $t(60) = 6.70,$ $p < .001$	No $t(60) = 1.56,$ $p = .12$	NA	Yes $t(60) = 18.13,$ $p < .001$	Yes $t(60) = 6.41,$ $p < .001$

[^] This does not take confidence rating into account. Both very certain and slightly certain response were worth 1 point.

^{*} This takes confidence rating into account. Very certain response was worth 2 points while slightly certain were worth 1 point.

For data analysis, we followed Schmalz et al.'s (2019) procedure by generating (a) a correlation matrix containing Pearson's correlations and (b) Bayes factors for the presence of each correlation. These are shown in Table 3. Scatterplots of each correlation are available in Figure 5.

367

Table 3. Correlation matrix showing the Pearson's correlation and Bayes factors

Measures	1	2	3	4	5
1. VSL (Without Confidence Rating)	1	0.70*** BF > 1000	0.17# BF = 0.64	0.04 BF = 0.30	0.03 BF = 0.30
2. VSL (With Confidence Rating)	-	1	0.19# BF = 0.82	0.17 BF = 0.61	-0.15 BF = 0.51
3. English Spelling Test	-	-	1	0.43*** BF = 76.6	0.13 BF = 0.46
4. Vocabulary Size Test	-	-	-	1	0.14 BF = 0.50
5. Explicit Learning Task (MALT-V)	-	-	-	-	1

*** $p < .001$

Bayes Factors (BFs) that have a value larger than 3 are in bold font and provide evidence for the presence of a correlation (Schmalz et al., 2019). BFs were calculated using the BayesFactor package (Morey & Rouder, 2018) in R Version 3.4.1 (R Core Team, 2017).

Partial correlations between the VSL scores and the spelling task were also computed, with measure 4, L2 vocabulary size, serving as the control. Pearson's r s were .17 (Measures 1 & 3) and .14 (Measures 2 & 3) after partialling out the effect of L2 vocabulary size.

Mirroring Schmalz et al.'s (2019) findings, the current study found no reliable correlations between performance on the VSL task and any of the language/learning measures. Importantly, we failed to replicate the finding of Mak (2016), who reported a negative correlation between performance on a VSL task and L2 spelling proficiency. The result of Mak (2016), therefore, may plausibly have been a false positive owing to the small sample size employed ($N = 26$).

The only two significant correlations in the current are the one between the two VSL scores ($r = .70, p < .001$) and the one between the spelling task and the vocabulary size test ($r = .43, p < .001$).

Figure 5. Scatterplots for each correlation in Table 3

5. Discussion

The current study aimed to ascertain if performance on a Visual Statistical Learning (VSL) task, namely the embedded triplet task, correlates with individual differences in L2 (English) spelling among a relatively large group of highly skilled L2 English speakers ($N = 61$). Contrary to our hypothesis and previous findings, performance on the VSL task did not correlate with the L2 spelling test, or the explicit learning task. Spelling proficiency also did not correlate with the explicit learning task. We did, however, find that L2 spelling and L2 vocabulary size are moderately correlated ($r = .43, p < .001$). In short, we did not find evidence supporting the proposal that a more efficient VSL capacity is related to higher L2 spelling proficiency.

We first of all consider the lack of any meaningful correlation between performance on the spelling test and the VSL task. This study was motivated by Mak (2016) who reported a negative correlation between these two tasks. Despite using essentially the same set of tasks and test items, we failed to replicate the finding of Mak (2016). This discrepancy, we argue, can be attributed to Mak (2016) employing a small sample size, which might have given rise to a false positive. Taken together with Mak (2016), our study illustrated the danger of using small sample sizes to explore individual differences. Researchers interested in individual differences should therefore avoid unjustifiably small sample sizes in order to prevent false positives. We by no means suggest or imply that the positive results from previous studies (e.g., Arciuli & Simpson, 2012; Frost et al., 2013) are false positives, as there are fundamental differences between the current study and those studies, for example they investigated decoding ability while ours investigated spelling ability; these two literacy skills may in fact rely on VSL in different ways or to a different extent. However, we would like to stress that those studies relied on relatively small sample sizes ($N = 37$ adults and $N = 26$ adults, respectively), and to the best of our knowledge, most if not all attempts to replicate those studies were not met with any success (see for example Schmalz et al., 2019; Steacy et al.,

2019; Qi et al., 2019). At the end of this Discussion section, we return to the issues associated with small sample sizes and discuss in detail why they are dangerous to the progression of the scientific community.

Next, we turn to the VSL paradigm employed in this study: the embedded triplet task. This task was adopted by Arciuli & Simpson (2012) and Frost et al. (2013), whose findings have generated considerable excitement among reading researchers. This task has been used in studies concerned with reading proficiency (Qi et al., 2019) and developmental dyslexia (e.g., Hjartarson, 2017; Witteloostuijn, Boersma, Wijnen & Rispens, 2018). Despite its satisfactory test-retest reliability (Siegelman & Frost, 2015; Qi et al., 2019), this task has been criticized: for example, all the test items in the test phase have the same level of difficulty (Siegelman et al., 2017), but in written languages, some statistical regularities (e.g., less frequent ones) are more difficult for acquisition, implying variability in difficulty. So a question then necessarily follows: is it possible for a test that contains absolutely no variability in difficulty to predict reading or spelling acquisition that contains much variability? Our data seem to suggest that the answer is no. Furthermore, in for example Frost et al. (2013), the VSL task contained only 32 test trials; having such a limited number of trials is difficult for sufficient variability to emerge, in turn making meaningful correlations hard to come by (Schmalz et al., 2019). We attempted to resolve this issue in the current study by slightly modifying the test phase by first of all increasing the number of unique ($N = 24$) and total test trials ($N = 48$) and by taking confidence rating into consideration (for details, see *Test phase* in section 3.2.2). Rather disappointingly, these efforts did not seem to bear much fruit: while distribution of the VSL scores is normal, it showed a high degree of clustering around the mean (see Figure 6a). Another related issue is that a large number of participants achieved the same VSL score: for example, 10 of the 61 participants attained the mode VSL score (26, i.e., 54.1%). Such clustering stands in contrast with performance on the spelling task, where the score range is much wider (see Figure 6b) and the mode spelling score (41, i.e., 51.25%) is only shared by

three participants. Therefore, these seem to suggest that the current VSL task is not very sensitive to individual differences.

Secondly, our attempt to take confidence rating into account proves inconclusive: when the confidence rating is taken into account, overall VSL performance did not significantly differ from chance ($t(60) = 1.56, p = .12$). This may be related to the fact that most participants, for unknown reasons, had a strong tendency for avoiding the Very Certain options.

Figure 6a. Data distribution of VSL scores (without confidence rating) (in %)

Figure 6b. Data distribution of spelling scores (in %)

Briefly reiterating here, the mastery of written language can be interpreted as the acquisition of the statistical regularities encoded in written languages (Nation & Mak, in press).

It is therefore a very reasonable claim that reading development, be it in L1 or L2, is causally related to a person's ability to detect and extract statistical regularities. However, a growing body of evidence suggests that traditional SL tasks such as the embedded triplet task, artificial grammar learning (Nieuwenhuis et al., 2013), and serial reaction time task (Kuppuraj & Arpitha, 2018) may not be suitable for discerning such a relation: For one, these implicit learning tasks have been criticized for their low reliability (West, Vadillo, Shanks, Hulme, 2017). For another, inter-task correlations are unsatisfactorily low, which then leads to the question of whether they are measuring the same underlying cognitive mechanism (Schmalz et al., 2019; West et al., 2017). Even though these findings do not necessarily undermine the validity of the tasks in tapping statistical or implicit learning, they cast serious doubt over whether they can predict other cognitive abilities (Siegelman & Frost, 2015). Our study adds to this literature by showing that implicit SL tasks like the one used in this study may not be sensitive enough to individual differences, so when a study is sufficiently powered (like the current study), this lack of sensitivity becomes apparent, potentially explaining why no correlation is found between this VSL task and L2 spelling proficiency. Note, however, a lack of correlation here does not negate the argument of (V)SL being causally related to L2 spelling; it is plausible that the null result here is a consequence of the current VSL task being incapable of revealing any meaningful differences in VSL ability. We are therefore delighted to see that efforts are dedicated to developing new SL measures (Siegelman, Bogaerts & Frost, 2017). Nevertheless, as noted in Nation & Mak (in press), caution should be exercised in the development of new SL tasks, because “the closer an SL task becomes to the task in hand—reading (or spelling)—the harder it is to maintain a distinction between SL as a cause of individual differences in reading vs. a finer-level description of them”.

So far, we attributed the lack of correlation between the VSL task and spelling task to two potential factors: (1) Positive results from previous studies (e.g., Mak, 2016) may well have been a consequence of small sample sizes, (2) the VSL task employed in this study (i.e.,

the embedded triplet task) is not sensitive enough for individual differences. These two explanations are both methodologically-based. Here, we put forth a theoretically-based explanation to the lack of correlation between the two tasks. In the current study, the 61 participants recruited were relatively *advanced* learners of English; they all have received a minimum of 15 years of English education and are using English on a day-to-day basis. This stands in contrast to the participants recruited by Frost et al. (2013), who reported a positive correlation between VSL ability and the amount of improvement in L2 reading proficiency in *beginning* L2 learners. Taken together, these findings highlight the possibility that the importance of VSL in literacy development may actually diminish as L2 proficiency increases. Such a trajectory (i.e., a cognitive capacity becomes decreasingly important as a function of L2 proficiency) is not uncommon: For example, Serafi & Sanz (2016) reported a robust positive correlation between phonological working memory and L2 competence in beginning L2 learners, but not in advanced learners. Comparable findings have also been reported by Hummel (2009) and Winke (2013). In light of such evidence, we also entertain the possibility that the role of (V)SL is more critical and apparent in the early stages of L2 reading development, but its role will gradually diminish as proficiency increases. This will explain why there was no correlation between VSL ability and spelling in the current group of highly skilled English learners.

Next, we turn to the positive correlation between the spelling test and the vocabulary size test ($r = .43, p < .001$). This correlation should be very self-explanatory. Having more L2 vocabulary knowledge suggests that the learner is more fine-tuned to the frequency of permissible letter strings and the mapping between graphemes and phonemes in the second language. Such refined statistical knowledge is perhaps of essential importance to L2 spelling proficiency. In addition, when a person attempts to spell out a never-encountered word (e.g., *ozonosphere*), they may compare its auditory form to a familiar word (e.g., *atmosphere*) and utilize their orthographic knowledge of this familiar word to deduce the spelling of this never-

encountered word. So for someone whose vocabulary bank does not contain any words that end with *-sphere*, they should experience greater difficulty spelling *ozonosphere* than someone whose vocabulary bank contains *atmosphere*. In other words, the more words a person knows, the more templates they could use; this should allow them to spell out never-encountered words with higher accuracy. We would like to point out one interesting aspect of this positive correlation between the spelling test and vocabulary size test. While the former test primarily tapped the mapping between phonology and orthography, the latter primarily tapped the mapping between orthography and semantics. Note that although spelling out presumably unknown words requires absolutely no knowledge of semantics, the current data suggest that knowing more orthography-semantics mappings is associated with better phonology-orthography mappings. A potentially interesting research question is thus: Does semantics play a role in the learning and accumulation of orthographic regularities? If it does, how does it contribute to literacy development (for discussion, see Castle, Rastle & Nation, 2018; Nation, 2009)? These questions, although not within the scope of the present study, are worth investigating in future research.

Moving on to the periphery goal of the present study: whether the explicit learning task (MLAT-V) correlates with the spelling task and the VSL task. Our study did not find any correlation between these tasks.

First, let us focus on the potential relation between the explicit learning task and the spelling task. The current null result can only leave us to speculate: it is possible that the capacity for explicit learning crucially contributes to spelling development in the initial phases of L2 acquisition and less so in the advanced stages of L2 acquisition. If this is the case, we may expect a positive correlation in beginning L2 learners, but not in advanced L2 learners. This suggestion echoes well with Hamrick (2015) who reported that initial L2 syntactic learning was best predicted by explicit learning while longer-term learning was best predicted

by an implicit learning measure. Future studies should consider pinpointing how the role of explicit and implicit learning changes according to L2 proficiency.

Finally, let us turn to the lack of correlation between the explicit learning task (MLAT-V) and the (implicit) VSL task. As proposed by Ullman (2004), there exists a trade-off relation between explicit and implicit learning; that is, good explicit learning may translate into poor implicit learning. Admittedly, this proposal is mainly concerned with language impairment, but Mak's 2016 (2016) finding seemed to suggest that a trade-off relation may also exist in language-typical adults. We found no evidence in support of such between the explicit learning task and the implicit VSL task. However, we would like to stress that we are not aware of any substantial evidence that lends credence to Ullman's (2004) original proposal. While findings from Lukacs, Kemeny, Lum & Ullman (2017) and Hedenius, Ullman, Alm, Jennische & Persson (2013) suggested an enhanced explicit learning mechanism in children with developmental dyslexia and developmental language impairment (DLD), these studies were not intended to test this specific trade-off relation (i.e., whether a negative correlation between explicit and implicit learning exists). We are only aware of one study which did exactly that: Kupperraj, Rao & Bishop (2016) recruited 30 children with DLD and administered an implicit learning measure and two explicit learning measures. They reported no evidence of trade-off. In sum, whether a trade-off between explicit and implicit learning genuinely exists in language impaired or even language-typical individuals awaits further research.

We end our discussion with a note on small sample sizes. Small sample size reduces statistical power, which in turn reduces "the likelihood that a statistically significant result reflects a true effect" (Buttons et al., 2013). For example, with a hypothetical true effect size of 0.35, a sample size of 30, and a Type 1 error rate of 0.05, studies using these parameters will on average obtain an effect size of 0.50, meaning that there is an effect inflation rate of 42.9% ($(0.5-0.35)/0.35$; Andrews, 2017). Since a large number of SL studies used highly similar parameters, we have reasons to suspect that the effect sizes in the VSL-Reading literature are

overestimated. In addition, it is widely known that low power due to small sample sizes raises the possibility of yielding false positives. For example, we estimated the statistical power of Arciuli & Simpson (2012, Experiment 2) using the pwr package (Champely, 2017) in R Version 3.4.1 (R Core Team, 2017), which gave a statistical power estimation of 0.54. With a hypothetical probability of finding a true hypothesis ranging from 0.1 to 0.4, the false positive rate of a study similar to Arciuli & Simpson (2012) is estimated to be 10% to 41% (Andrews, 2017). These statistics are for illustration purpose only, and we are by no means saying the results of Arciuli & Simpson (2012) are false positives. However, our study, along with Mak (2016), provided a clear illustration of the potential peril of using small sample sizes: with 26 participants, Mak (2016) found a significant correlation between VSL and L2 spelling, but in the current study where we used essentially the same methodologies but a much larger sample size ($N = 61$; justified by power analysis), we failed to find support for such a correlation. False positives as a result of small sample sizes are detrimental to the progression of psychology and neuroscience as they may lead us down onto fruitless research avenues. Some might argue that science is self-correcting, so false positives will be disproven and somehow lose their place, but due to for example publication bias and self-censorship which prevent most null results from seeing the light of day, science will not be able to correct itself. High-quality studies with large sample sizes and high transparency (by the means of for example pre-registration) are conducive to the advancement of the fields. In closing, we urge psychologists and neuroscientists worldwide to join forces to achieve better science by for example avoiding unjustifiably small sample sizes.

References

- Arciuli, J., & Simpson, I. C. (2012). Statistical Learning Is Related to Reading Ability in Children and Adults. *Cognitive Science*, 36(2), 286–304. <https://doi.org/10.1111/j.1551-6709.2011.01200.x>
- Andrew, M. (2017). Some demos related to the psychology replication crisis, *GitHub repository*. <https://github.com/lawsofthought/replication-crisis-demos>
- Au, C., & Noel, E. (1999). “Memorisation with understanding” in approaches to studying: cultural variant or response to assessment demands? In *European Association for Research on Learning and Instruction Conference*. Gothenburg, August 1999. Retrieved from <http://www.leeds.ac.uk/educol/documents/00001728.htm>
- Button, K. S., Ioannidis, J. P. a, Mokrysz, C., Nosek, B. a, Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews: Neuroscience*, 14(5), 365–76. <https://doi.org/10.1038/nrn3475>
- Castles, A., Rastle, K., & Nation, K. (2018). Ending the Reading Wars: Reading Acquisition From Novice to Expert. *Psychological Science in the Public Interest*, 19(1), 5–51.
- Champely, S. (2018). pwr: Basic Functions for Power Analysis. R package version 1.2-2. <https://CRAN.R-project.org/package=pwr>
- Elleman, A. M., Steacy, L. M., & Compton, D. L. (2019). The Role of Statistical Learning in Word Reading and Spelling Development: More Questions Than Answers. *Scientific Studies of Reading*, 23(1), 1–7. <https://doi.org/10.1080/10888438.2018.1549045>
- Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 28(3), 458–467. <https://doi.org/10.1037/0278-7393.28.3.458>
- Frost, R., Siegelman, N., Narkiss, A., & Afek, L. (2013). What Predicts Successful Literacy Acquisition in a Second Language? *Psychological Science*, 24(7), 1243–1252.

<https://doi.org/10.1177/0956797612472207>

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

<https://doi.org/10.1016/j.lindif.2015.10.003>

Hedenius, M., Ullman, M. T., Alm, P., Jennische, M., & Persson, J. (2013). Enhanced Recognition Memory after Incidental Encoding in Children with Developmental Dyslexia. *PLoS ONE*, 8(5), 1–7. <https://doi.org/10.1371/journal.pone.0063998>

Hjartarson, K. H. (2017). *High-Level Vision in Dyslexia: The Possibility of Inefficient Visual Expertise as a Contributing Factor to Reading Problems*. (Unpublished master's thesis). Háskóli Íslands, Reykjavík, Iceland.

Hummel, K. M. (2009). Aptitude, phonological memory, and second language proficiency in nonnovice adult learners. *Applied Psycholinguistics*, 30(2), 225.

<https://doi.org/10.1017/S0142716409090109>

Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition*, 116(3), 321–340.

<https://doi.org/10.1016/j.cognition.2010.05.011>

Kidd, E., & Arciuli, J. (2016). Individual Differences in Statistical Learning Predict Children's Comprehension of Syntax. *Child Development*, 87(1), 184–193.

<https://doi.org/10.1111/cdev.12461>

Kuppuraj, S., Rao, P., & Bishop, D. V. (2016). Declarative capacity does not trade-off with procedural capacity in children with specific language impairment. *Autism & Developmental Language Impairments*, 1, 1–17.

<https://doi.org/10.1177/2396941516674416>

Kuppuraj, S. & Arpitha, V. (2018, July 13). Initial and Later Procedural Learning Rates of Implicit Sequences in Good and Poor Readers of English.

<https://doi.org/10.31234/osf.io/w86rx>

644 Lukacs, A., Kemeny, F., Lum, J. A. G., & Ullman, M. T. (2017). Learning and Overnight
 645 Retention in Declarative Memory in Specific Language Impairment. *PLOS ONE*, 12(1),
 646 e0169474. <https://doi.org/10.1371/journal.pone.0169474>
 647 Mak, H.C. (2016). *L2 spelling proficiency is related to the capacity for Visual Statistical*
 648 *Learning (VSL), but negatively*. Paper presented at the Interdisciplinary advances in
 649 statistical learning, Bilbao, 28-30 June (p.75). Bilbao: BCBL. Retrieved from
 650 <https://www.bcbi.eu/events/statistical-learning2017/en/speakers/desde0/ver/2629/>
 651 Morey, R.D. & Rouder, J.N. (2018). BayesFactor: Computation of Bayes Factors for
 652 Common Designs. R package version 0.9.12-4.2. [https://CRAN.R-](https://CRAN.R-project.org/package=BayesFactor)
 653 [project.org/package=BayesFactor](https://CRAN.R-project.org/package=BayesFactor)
 654 Nation, K. (2009). Form-meaning links in the development of visual word recognition.
 655 *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1536),
 656 3665–3674. <https://doi.org/10.1098/rstb.2009.0119>
 657 Nation, K., & Castles, A. (2017). Putting the learning in to orthographic learning. In K. Cain,
 658 D. Compton, & R. Parrila (Eds.), *Theories of reading development* (pp. 147–168).
 659 Amsterdam, The Netherlands: John Benjamins.
 660 Nation, K., & Mak, M. H. C. (in press). Orthographic learning and learning to read:
 661 Implications for developmental dyslexia. In D. L. Compton, J. A. Wahington,
 662 & P. McCardle (Eds.), *Dyslexia 101: Revisiting etiology, diagnosis, treatment, and*
 663 *policy*. Baltimore, MD: Brookes Publishing.
 664 Nation, I.S.P. & Beglar, D. (2007) A vocabulary size test. *The Language Teacher*, 31(7), 9-
 665 13.
 666 Nieuwenhuis, I. L. C., Folia, V., Forkstam, C., Jensen, O., & Petersson, K. M. (2013). Sleep
 667 Promotes the Extraction of Grammatical Rules. *PLoS ONE*, 8(6).
 668 <https://doi.org/10.1371/journal.pone.0065046>
 669 Qi, Z., Sanchez Araujo, Y., Georgan, W. C., Gabrieli, J. D. E., & Arciuli, J. (2019). Hearing

670 Matters More Than Seeing: A Cross-Modality Study of Statistical Learning and Reading
671 Ability. *Scientific Studies of Reading*, 23(1), 101–115.
672 <https://doi.org/10.1080/10888438.2018.1485680>

673 R Core Team (2017). R: A language and environment for statistical computing. R Foundation
674 for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

675 Schmalz, X., Moll, K., Mulatti, C., & Schulte-Körne, G. (2019). Is Statistical Learning
676 Ability Related to Reading Ability, and if so, Why? *Scientific Studies of Reading*, 23(1),
677 64–76. <https://doi.org/10.1080/10888438.2018.1482304>

678 Schlichting, M. L., Guarino, K. F., Schapiro, A. C., Turk-Browne, N. B., & Preston, A. R.
679 (2017). Hippocampal Structure Predicts Statistical Learning and Associative Inference
680 Abilities during Development. *Journal of Cognitive Neuroscience*, 29(1), 37–51.
681 https://doi.org/10.1162/jocn_a_01028

682 Serafi, E. J., & Sanz, C. (2016). Evidence for the decreasing impact of cognitive ability on
683 second language development as proficiency increases. *Studies in Second Language*
684 *Acquisition*, 38, 607–646. <https://doi.org/10.1017/S0272263115000327>

685 Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical
686 perspectives and empirical evidence. *Journal of Memory and Language*, 81(4), 105–
687 120. <https://doi.org/10.1016/j.jml.2015.02.001>

688 Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017). Towards a theory of
689 individual differences in statistical learning. *Philosophical Transactions of the Royal*
690 *Society B*, 372, 20160059. <https://doi.org/10.1098/rstb.2016.0059>

691 Steacy, L. M., Compton, D. L., Petscher, Y., Elliott, J. D., Smith, K., Rueckl, J. G., ... Pugh,
692 K. R. (2019). Development and prediction of context-dependent vowel pronunciation in
693 elementary readers. *Scientific Studies of Reading*, 23(1), 49–1563.
694 <https://doi.org/10.1080/10888438.2018.1466303>

695 Torppa, M., Georgiou, G. K., Niemi, P., Lerkkanen, M. K., & Poikkeus, A. M. (2016). The

- precursors of double dissociation between reading and spelling in a transparent orthography. *Annals of Dyslexia*, 1–21. <https://doi.org/10.1007/s11881-016-0131-5>
- Treiman, R., & Kessler, B. (2005). Writing Systems and Spelling Development. In M. J. Snowling & C. Hulme (Eds.), *The Science of Reading: A Handbook*. Oxford: Blackwell Publishing Limited.
- West, G., Vadillo, M. A., Shanks, D. R., & Hulme, C. (2017). The procedural learning deficit hypothesis of language learning disorders: we see some problems. *Developmental Science*, (January), e12552. <https://doi.org/10.1111/desc.12552>
- Winke, P. (2013). An Investigation Into Second Language Aptitude for Advanced Chinese Language Learning. *The Modern Language Journal*, 97(1), 109–130. <https://doi.org/10.1111/j.1540-4781.2013.01428.x>
- van Witteloostuijn, M., Boersma, P., Wijnen, F., & Rispens, J. (2018). *Relationships between statistical learning and literacy skills in children with and without dyslexia*. Paper presented at the *Twenty-Fifth Annual SSSR Meeting*, Brighton, 18-21 July. Retrieved from <https://www.triplesr.org/understanding-role-statistical-learning-reading-and-spelling-development-across-languages>
- Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, 92(1–2), 231–270. <https://doi.org/10.1016/j.cognition.2003.10.008>
- Yu, A. Y. (2016). *Correlation of statistical learning of different implicit transitional probabilities with Chinese literacy acquisition in non-native adult learners*. (Unpublished doctoral thesis). National Central University, Taiwan.

Disclosure Statement

The authors report no conflict of interest.

Appendix 1: Power analysis

Studies	Correlation between literacy tasks and the VSL task (embedded triplet task)
Arciuli & Simpson (2012; Exp 2)	$r = .338, p < .05$
Frost et al. (2013)	$r = .57, p < .05$ (Pointed nonword decoding) $r = .43, p < .05$ (Unpointed naming accuracy) $r = .57, p < .05$ (Morphological priming effects)

We first averaged the correlation coefficients among the above studies ($r = .47$). We then used the pwr package (Champely, 2017) in R Version 3.4.1 (R Core Team, 2017). By entering $r = 0.47$, sig.level = 0.05, power = 0.95 (= 95%), a sample size estimation equal to 52 participants was obtained. To guard against data loss and equipment failure, we decided to recruit up to 65 participants.

Appendix 2: Vocabulary items used in the spelling test

3 syllables	4 syllables	5 syllables
Latin Origin		
canorous	abomasum	commensalism
fastigate	calcination	entomophagy
inspissate	comminatory	temerarious
patulous	divaricate	papaveraceous
nigrescent	erubescant	
	manubrium	
	mastodontic	
	noctambulist	
	nugacity	
	piscatorial	
	rapscallion	
	rubiginous	
	sternutator	
umbriferous		
French Origin		
ablative	adulation	amphibology
bombazine	bituminous	
divagate	concubinage	
tardigrade	deglutition	
	deracinate	
	fecundity	
	scurrilously	
Unknown/Other Origins		
laccolith	anthophilous	
wainscoting	galactoside	
	neostigmine	