Is statistical learning ability related to reading ability, and if so, why?

Xenia Schmalz[*1], *Kristina Moll*[1], *Claudio Mulatti*[2], & *Gerd Schulte-Körne*[1]

Accepted manuscript to appear in the Journal of the Scientific Studies of Reading

* Corresponding author. Email: xenia.schmalz@gmail.com, phone: +4917625464379.

[1] Department of Child and Adolescent Psychiatry, Psychosomatics, and Psychotherapy, Ludwig-Maximilians-University, Munich, Germany.

[2] Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Università degli Studi di Padova, Italy.

Abstract

Previous studies found a relationship between performance on statistical learning (SL) tasks and reading ability and developmental dyslexia. Thus, it has been suggested that the ability to implicitly learn patterns may be important for reading acquisition. Causal mechanisms behind this relationship are unclear: though orthographic sensitivity to letter bigrams may emerge through SL and facilitate reading, there is no empirical support for this link. We test 84 adults on two SL tasks, reading tests, and a bigram sensitivity task. We test for correlations using Bayes Factors. This serves to test the prediction that SL and reading ability are correlated, and to explore sensitivity to bigram legality as a potential mediator. We find no correlations between SL tasks and reading ability, SL and bigram sensitivity, or between the SL tasks. We conclude that correlating SL with reading ability may not yield replicable results, partly due to low correlations between SL tasks.

Keywords: Artificial Grammar Learning; Serial Reaction Time Task; letter bigram legality; word reading; nonword reading

Is statistical learning related to reading ability, and if so, why?

Learning to read involves developing sensitivity to orthographic regularities. For example, during reading acquisition, children have been shown to learn which letter combinations occur frequently, and which do not occur in their orthography (Cassar & Treiman, 1997; Pacton, Perruchet, Fayol, & Cleeremans, 2001; Rothe, Cornell, Ise, & Schulte-Körne, 2015; Rothe, Schulte-Körne, & Ise, 2014). Readers also learn context-dependent pronunciations of letters and graphemes: For a nonword such as "wamp", English-speaking readers pronounce the grapheme *a* as in "wasp" because it is preceded by a *w*, even though the rule */w/a* → /ɔ/ is not taught at schools (Schmalz et al., 2014; Treiman, Kessler, & Bick, 2003; Treiman, Kessler, Zevin, Bick, & Davis, 2006).

Children's sensitivity to such orthographic regularities has led to the proposition that reading acquisition, in part, relies on a domain-general statistical learning ability (hereafter: SL; Arciuli & Simpson, 2012). SL refers to the ability to detect and learn regularities in the environment. It is generally measured by non-linguistic tasks, which contain hidden regularities: for example, in a Serial Reaction Time Task, participants respond to a series of simple stimuli which follow a pre-determined sequence. Without becoming explicitly aware of this sequence, participants' performance improves with practice, and drops when the stimuli are presented in a different sequence.

If SL is important for learning to read, one might predict a correlation between reading ability and performance on SL tasks. Two studies to date have tested this prediction. Using a sample of non-selected participants, Arciuli and Simpson (2012) showed a significant partial correlation (after taking into account age) between performance on a SL task and a reading task, both in English-speaking children and

adults. These results were taken to suggest an important role of SL ability during reading acquisition. Furthermore, the correlation of SL with reading ability in adults suggests that the effect of SL explains variance in reading ability even when the participants have achieved a high level of competence. A second study, which examined the correlation between reading ability and SL (Frost, Siegelman, Narkiss, & Afek, 2013), tested English native speakers who were learning Hebrew as a second language. Here, performance on a visual SL task, akin to the task used by Arciuli and Simpson (2012), was correlated with reading ability: participants who performed well on this task also made faster progress in both unpointed nonword and pointed word reading ability. However, as this study tested adult second-language learners of a new script, it is not clear whether it reflects a similar relationship as the correlation between SL and learning to read one's first language as a child.

While only two studies assessed the relationship between SL and reading ability in an unselected sample, more than a dozen studies compared the performance on SL tasks in a group of participants with developmental dyslexia (hereafter: dyslexia) to a control group. If SL is correlated with reading ability, one should predict such studies to find group differences, as choosing a set of participants with dyslexia increases the range of reading ability among the participants. The results of the studies on SL and dyslexia are mixed: a recent review (Schmalz, Altoè, & Mulatti, 2017) and a meta-analysis (van Witteloostuijn, Boersma, Wijnen, & Rispens, 2017) suggest that there is publication bias. Publication bias refers to the preferential publication of positive results, which consequently become over-represented in the published literature, leading to an increased Type-I error rate and inflated effect sizes (Rosenthal, 1979; Van Elk et al., 2015). With the presence of publication bias, it becomes difficult to determine whether an effect is different from zero, as different

statistical correction methods for meta-analyses often yield conflicting results (Van Elk et al., 2015). Therefore, we do not discuss whether or not there is sufficient evidence for a group difference in SL (but for a critical review, see Schmalz et al., 2017). For our purposes, it is worth describing the tasks which were used by these previous studies. The two studies on SL in an unselected population used a visual SL task: Here, participants see a series of shapes, presented one at a time (Arciuli & Simpson, 2012; Frost et al., 2013). This sequence of shapes includes embedded triplets: Three of the stimuli always follow one another. This means that the first and the second of the three stimuli can be used to predict the next stimuli, once the participant has (implicitly) learnt the sequence. In a subsequent recognition test, participants perceive stimulus pairs which occurred together within this triplet as more familiar than stimulus pairs that did not frequently occur together, suggesting that they learned these transitional statistics.

In contrast, the studies on SL and dyslexia used either Artificial Grammar learning (AGL) (Reber, 1967) or Serial Reaction Time Tasks (SRTT) (Nissen & Bullemer, 1987). In the AGL task, participants first see a set of symbol sequences in a learning phase. These are created according to a set of rules (see Figure 1). The rules specify the positions and sequences in which the symbols can occur. In a subsequent test phase, participants are presented with symbol strings that did not occur during the learning phase, and need to guess, for each string, whether it corresponds to the grammar that constrained the learning strings.

FIGURE 1 ABOUT HERE

In the SRTT, participants see a stimulus which can occur in different positions on the screen (e.g., top, bottom, left, right). The task of the participants is to press a key corresponding to the stimulus' location. Unknown to the participants, the

sequence of the locations repeats. With increased exposure to the repeated sequence, the participants' performance improves across blocks. Critically, towards the end, a block is inserted where the location sequence is randomised. If participants implicitly learned the sequence, their performance on this random block drops, compared to the preceding block.

A commonality between the three tasks is that the participants need to learn to use the available input to predict a future event. In the triplet task and SRTT, these are the identity and location, respectively, of an upcoming stimulus. In the AGL task, participants appear to learn symbol chunks (Pothos, 2007): in their decision about whether a given string is grammatical, participants rely on their knowledge of whether a given symbol can occur next to the other within a letter string. This also involves a prediction based on conditional probabilities: Given the first symbol of the sequence, what is the probability of the observed second symbol?

If SL, as a hypothetical single construct which is measured by all three tasks, is related to reading ability, it is likely that it does so through their shared component: observing regularities in the environment, and using this knowledge to predict an upcoming event. An alternative explanation for any correlations between SL tasks and reading ability is that they reflect participant-level confounds, such as the general level of attention or motivation (Staels & Van den Broeck, 2017; Waber et al., 2003).

Assuming that SL is correlated with reading ability, after partialling out the shared variance with a control task to account for differences in attention and motivation, the next question is about the causal pathways that lead from SL performance to reading ability. Orthographies contain regularities on many levels, and studies have shown that readers develop sensitivities to them. Children learn very quickly which letter sequences do or do not occur in their orthography (Cassar &

Treiman, 1997; Pacton, Fayol, & Perruchet, 2005; Pacton et al., 2001; but see also Deacon, Benere, & Castles, 2012; Rothe et al., 2014, for a failure to find evidence for a causal link between orthographic sensitivity and reading ability, using longitudinal designs). Bigram sensitivity could affect reading ability through a mediating link, namely spelling ability: knowing frequent letter patterns in one's orthography constrains possible spelling patterns of a word, which would improve a child's spelling ability. When writing the word "quick", a child who does not know how to spell it may rely on their knowledge of legal letter patterns of the English orthography to decide against spelling it as *ckwik*, although this spelling is phonologically plausible (for a review, see Chetail, 2015).

A second possible causal pathway between SL and reading ability could be via learning of complex grapheme-phoneme correspondences (GPCs). In alphabetic orthographies, graphemes sometimes have multiple sound associations, which depend on their context (Schmalz, Marinus, Coltheart, & Castles, 2015). In English, the grapheme *a* is generally pronounced as in "cat", but its pronunciation changes when it is preceded by a *w* or *qu*, as in "wasp" (Venezky, 1970). In German, vowel length can often be predicted based on the number of subsequent consonants (Perry, Ziegler, Braun, & Zorzi, 2010). These rules are not taught explicitly at school. However, with reading experience, German speakers become sensitive to such context-dependent regularities, as the number of consonants after a vowel affects the probability of participants reading a vowel as long or short: the nonword *BLAF*, with only one consonant in the coda, is more likely to be pronounced with a long vowel than the nonword *BAMT*, where the vowel is followed by two consonants (Schmalz et al., 2014). SL may be important for learning these complex GPCs through exposure to real words in one's orthography (Apfelbaum, Hazeltine, & McMurray, 2013). This

would enhance nonword reading skills, as readers of even a shallow orthography such as German would compute the correct pronunciation more quickly. The ability to decode unfamiliar words is, in turn, a well-established predictor of orthographic learning and reading ability (Share, 1995, 2008).

Other possible links between SL and reading ability include learning to use probabilistic cues to assign lexical stress in languages without a regular stress pattern (Arciuli, Monaghan, & Seva, 2010; Jouravlev & Lupker, 2015; Mousikou, Sadat, Lucas, & Rastle, 2017; Seva, Monaghan, & Arciuli, 2009; Sulpizio & Colombo, 2013), facilitating written word learning by learning fully-specified links between phonology, orthography, and semantics (Steacy, Elleman, & Compton, 2017), or indirectly, via oral language skills (Saffran, Newport, & Aslin, 1996; Seidenberg & Gonnerman, 2000; Spencer, Kaschak, Jones, & Lonigan, 2015). It is worth noting that, while there are abundant theories on the relation between SL and reading ability, there is less empirical work which would establish (1) the link between the performance on non-linguistic SL tasks and the learning of orthography-specific regularities, and (2) the link between sensitivity to a given orthography-specific regularity and reading ability.

Here, our aim is two-fold. First, we test the proposal that SL is important for reading. In line with previous findings of Arciuli and Simpson (2012), we expect to find a correlation between non-linguistic SL tasks and reading ability. Second, we test the plausibility of two possible mediators in the relationship: sensitivity to bigram legality and nonword reading ability. We use Bayesian correlation analyses, which allows us to draw conclusions about the absence of a correlation rather than only about significant correlations (Dienes, 2014; Rouder, Speckman, Sun, Morey, & Iverson, 2009).

As Arciuli and Simpson (2012), we tested a sample of unselected adults on two reading tests (word and nonword reading fluency) and two SL tasks (serial reaction time and artificial grammar learning). As these two SL tasks have been frequently used in the literature on SL and dyslexia, it is worth establishing whether they correlate with each other and thus measure the same SL construct. In addition, we use a correlational approach to test whether orthographic sensitivity may mediate the relationship between SL and reading. Participants performed an orthographic choice task which measured bigram sensitivity, and a task to control for individual differences in attention or motivation (choice reaction time).

Methods

Participants

Participants were 84 adult German native speakers, recruited at two universities and a research institute in southern Germany. Participant characteristics are summarised in Table 1. Reading percentiles show a wide range of reading ability compared to a normative sample of university students, apprentices, and high school graduates (Moll & Landerl, 2010). Participants were tested individually in sessions lasting about 30 minutes.

TABLE 1 ABOUT HERE

Tasks

Reading tasks

We used a standardised reading task to assess word and nonword reading fluency (Salzburger Lese- und Rechtschreibtest II; Moll & Landerl, 2010). The tests consist of lists of words and nonwords, respectively, arranged in columns and

increasing in difficulty. Participants are instructed to read as many items as possible within 60 seconds.

Dependent variables for the reading tests are the number of words or nonwords, respectively, read correctly within 60 seconds. Though performance on these two reading subtests is correlated, they reflect different cognitive processes, which are dissociated in some readers (Castles & Coltheart, 1993). If SL specifically affects the learning of GPCs, we might expect that readers with poor SL skills may show relatively poor nonword reading skills compared to word reading skills. To test this possibility, we calculated a difference score, by subtracting each participant's *z*-score of their nonword reading performance (compared to the rest of the sample) from the *z*-score of their word reading performance. Negative numbers reflect relatively good nonword reading skills compared to participants' word reading skills; positive numbers reflect relatively good word reading skills.

SL tasks

Serial reaction time task (SRTT). We implemented the SRTT in OpenSesame (Mathôt, Schreij, & Theeuwes, 2012). The stimulus, which occurred sequentially in one of four positions on the screen, was a cartoon-like drawing of a cow. Participants were instructed to indicate the cow's position on the numerical keyboard (8 for up, 4 for left, 6 for right, and 2 for down). The instructions were to respond to each stimulus as fast as possible, but to avoid making too many mistakes. Each trial was presented for two seconds or until a button press occurred. The location sequence repeated after each sixteen trials. There were twelve blocks of sixteen trials each. The eleventh block consisted of a different, pseudo-randomised sequence of sixteen trials.

There are numerous ways to calculate an outcome variable for this task, including improvement across repeated blocks, difference between the random block

and the preceding repeated block, difference between the random block and the succeeding repeated block, or difference between the random block and an average of the preceding and succeeding repeated blocks. Such flexibility is problematic, because multiple comparisons associated with different variables increase the Type I error rate (Elson, 2016). We therefore decided on the outcome variable a priori. For improvement across repeated blocks, it is unclear whether it reflects a practice effect or implicit learning. For the repeated block which succeeds the random block, implicit knowledge is likely to be already diluted by the random block. We therefore calculated the difference between the random block and the preceding repeated block. As accuracy rates are close to ceiling (in our task, average = 98.0%), RTs are better suited to assess individual differences. With RT measures, one needs to be weary of over-additivity: when relying on raw RTs, differences between conditions are numerically larger for participants with longer overall RTs (Faust, Balota, Spieler, & Ferraro, 1999). Thus, we *z*-transformed RTs for each participant. For the analysis, we excluded incorrect trials (2% of the data), and item points which deviated more than 3 SDs from each participant's mean (a further 1.5% of the data). The outcome variable was the *z*-score difference between the random block (Block 11) and the preceding repeated block (Block 10), where larger positive values reflect stronger implicit sequence learning.

Artificial grammar learning. This task consists of a learning phase and a test phase. In the learning phase, participants were exposed to symbol strings, which followed the set of rules summarised in Figure 1. As a cover task for the first phase, we presented participants with two grammatical symbol strings on the screen simultaneously, separated by 25 blank spaces. In half of the trials, the two symbol strings were identical, and in the other half of the trials, they were different. The

number of symbols contained in each string was identical for each pair. Participants were instructed to decide whether the strings of each pair were identical or different, by pressing the right or left shift key. Each trial stayed on the screen until a response occurred. There were 86 trials in total. Throughout the task, participants saw four repetitions of 43 legal symbol strings. The participants completed the cover task from the first phase with very high accuracy (mean: 97.8%, by-participant SD: 2.2%, minimum accuracy: 90.7%), which shows that they attended to the exposure strings. This data is not analysed further.

After the first phase, participants were told that the strings they just saw followed a set of complex rules. For the second phase, participants were presented with symbol strings which had not occurred in the learning phase. They were told that half of the symbol strings were created using the same rules as the strings in the previous part, and that they would need to guess whether each new string was created by the same rules. If the string seemed familiar to them, they were instructed to press the right shift key, and if the string looked less familiar, they were asked to press the left shift key. There were 44 trials altogether. Each symbol string stayed on the screen until a response occurred.

We programmed both phases in DMDX (Forster & Forster, 2003). We used the grammar in Figure 1, based on Knowlton and Squire (1994). In order to remove the potential confound of automatized letter knowledge, we used non-alphanumeric symbols instead of letters (Ziegler, Pech‐Georgel, Dufau, & Grainger, 2010).

Typically, for the critical second phase of the task, accuracy is too low for an RT analysis, therefore only accuracy rates are analysed. We calculated overall accuracy and the sensitivity index (*d'*). The latter measure is a *z*-score difference between the hit rate and the false alarm rate, and accounts for participants' response

bias (Stanislaw & Todorov, 1999). To calculate the *d'* score, hit or false alarm rates of 0 or 1 were changed to 0.00001 and 0.99999, respectively, as 0 or 1 yield *z*-scores of ±∞. Higher values of *d'* indicate better learning, and *d'* = 0 indicates chance performance.[1]

Sensitivity to frequent letter patterns

Participants were presented with nonwords either containing a letter bigram which never occurs in the German orthography, or nonwords with legal letter bigrams only. The task was to decide, for each item, if the nonwords follow the orthographic principles of German. Participants were instructed to respond as quickly as possible, and to guess if they were unsure. The items were presented in random order, for 5 seconds or until a response occurred, with DMDX (Forster & Forster, 2003). The items were 80 legal and 80 illegal nonwords, taken from Bakos, Landerl, Bartling, Schulte-Körne, and Moll (2018). All items were pronounceable in German. Illegal letter clusters were illegal letter doublets (e.g., *ovv, Tüü*) or consonant clusters (e.g., *Lutd* and *Alßt*, where the bigrams *td* and *ßt* do not occur in German). The nonwords were matched across the two conditions on length and syllabic structure. The test was preceded by 10 practice trials.

[1] Originally, we had also included a visual SL (triplet) task, akin to Arciuli and Simpson (2012) and Frost et al. (2013), with cartoon-like pictures of animals instead of aliens or shapes. However, after the first 30 participants completed this task, it became clear that there was not sufficient variability in the learning performance to yield meaningful correlations. This is in line with recent observations about the rather poor psychometric properties of this task (Siegelman, Bogaerts, & Frost, 2016). To save time, we therefore decided to discontinue using this task. The mean accuracy on the test phase was 54.4% (chance level: 50%), SD = 9.4, minimum = 40.0%, maximum = 75.0%. Pearson's correlations with this variable were $r(28) = 0.06$ for word reading, $r(28) = -0.14$ for nonword reading, and $r(28) = -0.07$ for SRTT learning, all BF < 1/3.

We calculated the overall accuracy and sensitivity index (*d'*) for this task. As accuracy was relatively high, we also calculated the overall RTs, after excluding incorrect trials (10.5% of the data).

Control task

To control for overall differences in processing speed which may reflect attention or motivation, participants saw different cartoon animals on the screen, presented in random order, and were instructed to press a key on the right-hand side of the keyboard if the animal was a cat, and a key on the left-hand side if the animal was not a cat. The instructions were to respond as fast as possible, but without making too many mistakes. The task was programmed with OpenSesame (Mathôt et al., 2012). For each trial, the stimulus was presented for 1500 ms or until a response occurred. The stimuli were three different-coloured pictures each of cats, cows, rabbits, and sheep. There were 120 trials, 30 of which required a "yes"-response. Here, we calculated both the accuracy and the average RTs for each subject.

Results

Table 1 shows the overall descriptive statistics. For the analysis, we generated a correlation matrix containing Pearson's correlations, and Bayes Factors (BFs) for the presence of each correlation (Table 2).[2] The scatterplots showing the relationship between the SL tasks and reading, between the two SL tasks, and between bigram sensitivity and reading and SL, are shown in Figure 2. Figure 3 shows the average performance of participants across blocks. A figure with all scatterplots, as well as the data used for the analyses, can be accessed at osf.io/fqdnh. We did not exclude any

[2] We do not report *p*-values, because the multiple comparisons would yield them uninterpretable, but for the readers' reference, given our sample size of 84, correlations exceeding $r \approx \pm 0.22$ reach the traditional significance threshold of $p = 0.05$.

outliers. The figures with scatterplots show that, while there were some outlier points for several tasks, these do not seem to distort any meaningful patterns.

TABLE 2 ABOUT HERE

FIGURES 2 AND 3 ABOUT HERE

The correlation coefficients and BFs were calculated with JASP (Love et al., 2015). BFs compare the extent to which the data is compatible with a pre-specified alternative hypothesis over a null hypothesis ($r = 0$). In JASP, the pre-specified alternative is a beta-distribution centred around $r = 0$. The width of the distribution determines the probability density, under the alternative model, of the occurrence of different correlation coefficients. The default parameter assumes a flat prior distribution, such that the probability density of r values between -1 and 1 is evenly distributed. As we expect the correlation between reading and SL to be small (Schmalz et al., 2017), we changed the default parameter to 0.5, which changes the distribution to one where extreme values become less likely.

BFs > 1 provide relative support for the alternative hypothesis, and BFs < 1 provide relative support for the null hypothesis. In line with guidelines summarised by Rouder et al. (2009), we interpret values between 1/3 and 3 as inconclusive evidence, and values <1/3 or >3 as evidence for the null and alternative hypotheses, respectively.

Critically, neither of the SL tasks show any strong correlations, neither with the reading tasks (all BFs < 0.7), nor did the SRTT outcome variable correlate with either of the AGL measures (BF < 1/3).

Discussion

In a sample of 84 adult readers, we found no correlation between any of SL tasks and reading ability, and no correlation between the two SL tasks. Thus, we did not find support for the proposal that SL ability has an effect on reading ability.

The current findings are not in line with previous results of Arciuli and Simpson (2012), who found a correlation between visual SL and reading ability in children and adults. As our study was not a close replication of the original experiment, it is possible that the methodological differences across the studies are responsible for the different outcomes. Thus, a future, pre-registered study is needed in order to determine whether the presence of a correlation can be confirmed, when the protocol closely follows the methods of Arciuli and Simpson (2012). If this is the case, future empirical work is needed to isolate moderating factors which could have led to the outcomes of the current study.

There are several methodological differences which could explain the different outcomes. In the final sample, we used different tasks compared to Arciuli and Simpson (2012) and Frost et al. (2013). Thus, there may be task-specific processes associated with the visual SL task which are correlated with reading ability. Furthermore, as the visual SL task requires participants to focus on a stimulus sequence, which lasts for several minutes, attention may be a confounding factor (Staels & Van den Broeck, 2017; Waber et al., 2003). However, we did not find a correlation between the visual SL task, either with reading ability or with statistical learning in the SRTT, in a subset of our sample which had a comparable size as the study of Arciuli and Simpson (2012) (see Footnote 1).

The two tasks which we used in the final analyses (AGL and SRTT) have been used by numerous studies on SL and dyslexia. As previous studies have linked

our SL tasks to dyslexia, the current results are interpretable within the literature on SL and reading ability. The lack of a correlation between the two tasks raises issues about their psychometric properties. It is possible that the tasks show insufficient variability to allow us to study individual differences (Hedge, Powell, & Sumner, 2017; Siegelman & Frost, 2015): In the AGL task, average performance was significantly above chance-level. On the individual level, however, all but 6 participants were numerically above chance (i.e., at >50% accuracy), but most of these were only slightly above chance level, such that their accuracy level was not significantly better than chance at the 5%-level (see Table 1; see Siegelman et al., 2016, for a discussion of this problem in SL tasks). This methodological issue prevents us from interpreting the absence of a correlation as evidence against the view that SL is important for reading. However, given the popularity of these tasks, our finding of no correlation is still important for future research.

We also did not find a correlation between SL and bigram sensitivity, or between bigram sensitivity and reading ability. Previous studies have been unable to find evidence for a causal relationship between bigram sensitivity and reading ability (Deacon et al., 2012; Rothe et al., 2014). Furthermore, the role of bigram frequency during reading processes is unclear (Schmalz & Mulatti, 2017). Our results are in line with these studies and may suggest that sensitivity to letter bigrams does not act as a mediating link between SL and reading ability. However, our adult participants were clearly already very sensitive to bigram legality, as shown by the high accuracy on the bigram sensitivity task. Bigram sensitivity may be related to reading ability during the early stages of reading acquisition; for adults, this influence may be masked by other variables which influence reading performance.

Finally, it is worth pointing out that our study was conducted with German speakers, while the participants of Arciuli and Simpson (2012) were English speakers. German and English are different in terms of orthographic depth: the English orthography contains more complex (multi-letter and context-sensitive) GPCs than German, as well as more words where the pronunciation is unpredictable based on print-speech regularities (Schmalz et al., 2015). It is possible that SL is more important for learning to read in English, which could be necessary for extracting the orthographic regularities relating to complex rules. However, we consider this an unlikely explanation for our results: Frost et al. (2013) reported that, in learners of Hebrew, SL predicted both the learning efficiency of pointed nonword reading ability (a very shallow script) and unpointed word reading ability (a very deep script). Thus, the influence of SL ability on reading acquisition does not seem to be moderated by orthographic depth.

In summary, we found that SL tasks which are typically used in the literature on SL and dyslexia do not correlate with reading. We cannot distinguish between the possibility that there is no link between SL and reading from the possibility that the tasks that are generally used are inadequate to show it. Future research may want to address issues of publication bias and poor psychometric properties of SL tasks. Researchers will need design a child-friendly SL task with good psychometric properties (for a SL task for adults, designed to have good psychometric properties, see Siegelman et al., 2016), and test a large sample of children to establish the presence or absence of a correlation between reading and SL.

References

Apfelbaum, K. S., Hazeltine, E., & McMurray, B. (2013). Statistical learning in reading: Variability in irrelevant letters helps children learn phonics skills. *Developmental Psychology, 49*(7), 1348.

Arciuli, J., Monaghan, P., & Seva, N. (2010). Learning to assign lexical stress during reading aloud: Corpus, behavioral, and computational investigations. *Journal of Memory and Language, 63*(2), 180-196. doi:10.1016/J.Jml.2010.03.005

Arciuli, J., & Simpson, I. C. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science, 36*(2), 286-304.

Bakos, S., Landerl, K., Bartling, J., Schulte-Körne, G., & Moll, K. (2018). Neurphysiological correlates of word processing deficits associated with reading versus spelling problems. *Clinical Neurophysiology, 129*(3), 526-540.

Cassar, M., & Treiman, R. (1997). The beginnings of orthographic knowledge: Children's knowledge of double letters in words. *Journal of Educational Psychology, 89*(4), 631.

Castles, A., & Coltheart, M. (1993). Varieties of developmental dyslexia. *Cognition, 47*, 149-180.

Chetail, F. (2015). Reconsidering the role of orthographic redundancy in visual word recognition. *Frontiers in Psychology, 6*, 645.

Deacon, S. H., Benere, J., & Castles, A. (2012). Chicken or egg? Untangling the relationship between orthographic processing skill and reading accuracy. *Cognition, 122*(1), 110-117.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*.

Elson, M. (2016). Flexibility in Methods & Measres of Social Science. Retrieved from flexiblemeasures.com

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin, 125*(6), 777-799. doi:10.1037//0033-2909.125.6.777

Forster, K. I., & Forster, J. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instuments & Computers, 35*, 116-124.

Frost, R., Siegelman, N., Narkiss, A., & Afek, L. (2013). What predicts successful literacy acquisition in a second language? *Psychological Science, 24*(7), 1243-1252.

Hedge, C., Powell, G., & Sumner, P. (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, 1-21.

Jouravlev, O., & Lupker, S. J. (2015). Lexical stress assignment as a problem of probabilistic inference. *Psychonomic Bulletin & Review, 22*(5), 1174-1192.

Knowlton, B. J., & Squire, L. R. (1994). The information acquired during artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(1), 79.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A., & Wagenmakers, E. (2015). JASP (Version 0.7)[computer software]. *Amsterdam, the netherlands: Jasp project*.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44*(2), 314-324.

Moll, K., & Landerl, K. (2010). SLRT-II: Lese- und Rechtschreibtest; Weiterentwicklyng des Salzburger Lese- und Rechtschreibtests (SLRT): Huber.

Mousikou, P., Sadat, J., Lucas, R., & Rastle, K. (2017). Moving beyond the monosyllable in models of skilled reading: Mega-study of disyllabic nonword reading. *Journal of Memory and Language, 93*, 169-192.

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

Pacton, S., Fayol, M., & Perruchet, P. (2005). Children's implicit learning of graphotactic and morphological regularities. *Child Development, 76*(2), 324-339.

Pacton, S., Perruchet, P., Fayol, M., & Cleeremans, A. (2001). Implicit learning out of the lab: the case of orthographic regularities. *Journal of Experimental Psychology: General, 130*(3), 401.

Perry, C., Ziegler, J., Braun, M., & Zorzi, M. (2010). Rules versus statistics in reading aloud: New evidence on an old debate. *European Journal of Cognitive Psychology, 22*(5), 798-812.

Pothos, E. M. (2007). Theories of artificial grammar learning. *Psychological Bulletin, 133*(2), 227.

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

Rosenthal, R. (1979). The "File Drawer Problem" and Tolerance for Null Results. *Psychological Bulletin, 86*(3), 638-641.

Rothe, J., Cornell, S., Ise, E., & Schulte-Körne, G. (2015). A comparison of orthographic processing in children with and without reading and spelling disorder in a regular orthography. *Reading and Writing, 28*(9), 1307-1332.

Rothe, J., Schulte-Körne, G., & Ise, E. (2014). Does sensitivity to orthographic regularities influence reading and spelling acquisition? A 1-year prospective study. *Reading and Writing, 27*(7), 1141-1161.

Rouder, J. N., Speckman, P. L., Sun, D. C., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225-237. doi:10.3758/Pbr.16.2.225

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35*(4), 606-621.

Schmalz, X., Altoè, G., & Mulatti, C. (2017). Statistical learning and dyslexia: a systematic review. *Annals of Dyslexia, 67*(2), 147-162. doi:10.1007/s11881-016-0136-0

Schmalz, X., Marinus, E., Coltheart, M., & Castles, A. (2015). Getting to the bottom of orthographic depth. *Psychonomic Bulletin and Review, 22*(6), 1614-1629. doi:10.3758/s13423-015-0835-2

Schmalz, X., Marinus, E., Robidoux, S., Palethorpe, S., Castles, A., & Coltheart, M. (2014). Quantifying the reliance on different sublexical correspondences in German and English. *Journal of Cognitive Psychology, 26*(8), 831-852. doi:10.1080/20445911.2014.968161

Schmalz, X., & Mulatti, C. (2017). Busting a myth with the Bayes Factor: Effects of letter bigram frequency in visual lexical decision do not reflect reading processes. *The Mental Lexicon, 12*(2), 263–282.

Seidenberg, M., & Gonnerman, L. M. (2000). Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences, 4*(9), 353-361.

Seva, N., Monaghan, P., & Arciuli, J. (2009). Stressing what is important: Orthographic cues and lexical stress assignment. *Journal of Neurolinguistics, 22*(3), 237-249. doi:10.1016/J.Jneuroling.2008.09.002

Share, D. (1995). Phonological recoding and self-teaching: *sine qua non* of reading acquisition. *Cognition, 55*, 151-218.

Share, D. (2008). Orthographic Learning, Phonological Recoding, and Self-Teaching. In R. V. Kail (Ed.), *Advances in Child Development and Behavior*: Elsevier.

Siegelman, N., Bogaerts, L., & Frost, R. (2016). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods*, 1-15.

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language, 81*, 105-120.

Spencer, M., Kaschak, M. P., Jones, J. L., & Lonigan, C. J. (2015). Statistical learning is related to early literacy-related skills. *Reading and Writing, 28*(4), 467-490.

Staels, E., & Van den Broeck, W. (2017). A specific implicit sequence learning deficit as an underlying cause of dyslexia? Investigating the role of attention in implicit learning tasks. *Neuropsychology, 31*(4), 371.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137-149.

Steacy, L. M., Elleman, A. M., & Compton, D. L. (2017). Opening the "black box" of learning to read. In K. Cain, D. L. Compton, & R. Parrila (Eds.), *Theories of Reading Development* (pp. 99-121). Amsterdam: John Benjamins Publishing Company.

Sulpizio, S., & Colombo, L. (2013). Lexical stress, frequency, and stress neighbourhood effects in the early stages of Italian reading development. *The Quarterly Journal of Experimental Psychology, 66*(10), 2073-2084.

Treiman, R., Kessler, B., & Bick, S. (2003). Influence of consonantal context on the pronunciation of vowels: A comparison of human readers and computational models. *Cognition, 88*(1), 49-78. doi:10.1016/s0010-0277(03)00003-9

Treiman, R., Kessler, B., Zevin, J. D., Bick, S., & Davis, M. (2006). Influence of consonantal context on the reading of vowels: Evidence from children. *Journal of Experimental Child Psychology, 93*(1), 1-24. doi:10.1016/J.Jecp.2005.06.008

Van Elk, M., Matzke, D., Gronau, Q. F., Guan, M., Vandekerckhove, J., & Wagenmakers, E.-J. (2015). Meta-analyses are no substitute for registered replications: a skeptical perspective on religious priming. *Frontiers in Psychology, 6*.

van Witteloostuijn, M., Boersma, P., Wijnen, F., & Rispens, J. (2017). Visual artificial grammar learning in dyslexia: A meta-analysis. *Research in Developmental Disabilities, 70*, 126-137.

Venezky, R. L. (1970). *The structure of English orthography* (Vol. 82): Walter de Gruyter.

Waber, D. P., Marcus, D. J., Forbes, P. W., Bellinger, D. C., Weiler, M. D., Sorensen, L. G., & Curran, T. (2003). Motor sequence learning and reading ability: Is poor reading associated with sequencing deficits? *Journal of Experimental Child Psychology, 84*(4), 338-354.

Ziegler, J., Pech‐Georgel, C., Dufau, S., & Grainger, J. (2010). Rapid processing of letters, digits and symbols: what purely visual‐attentional deficit in developmental dyslexia? *Developmental Science, 13*(4).

Figure 1

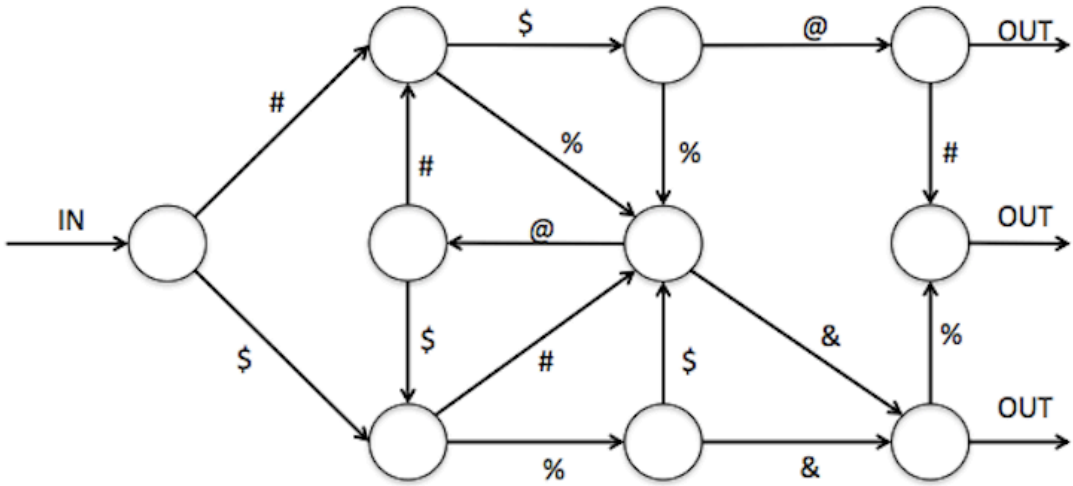
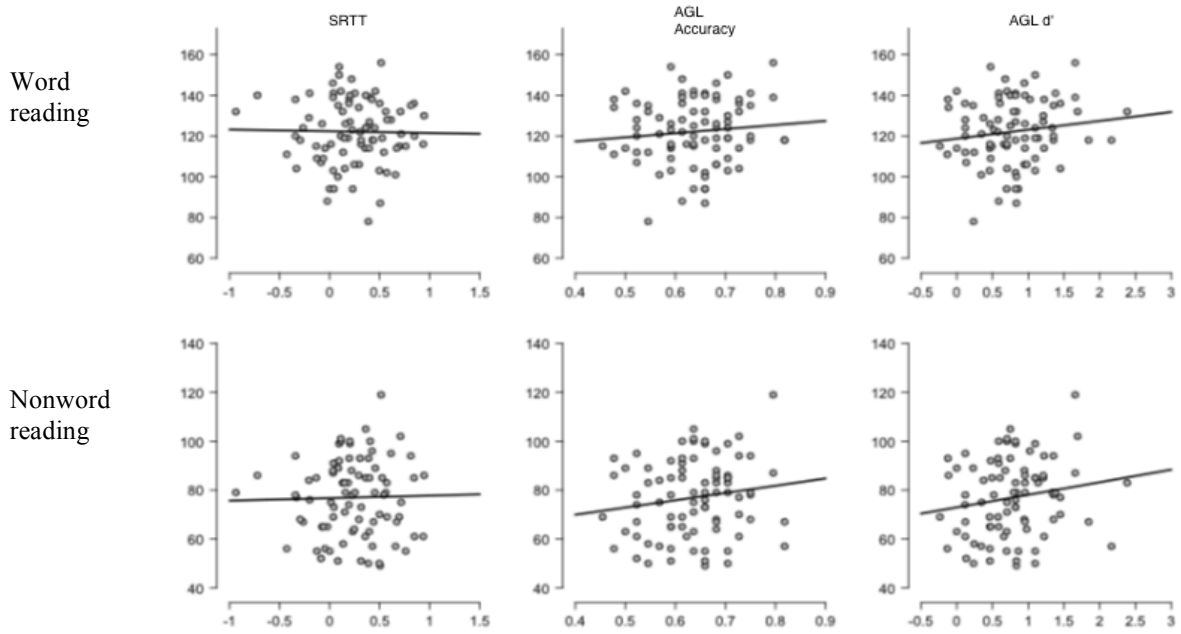
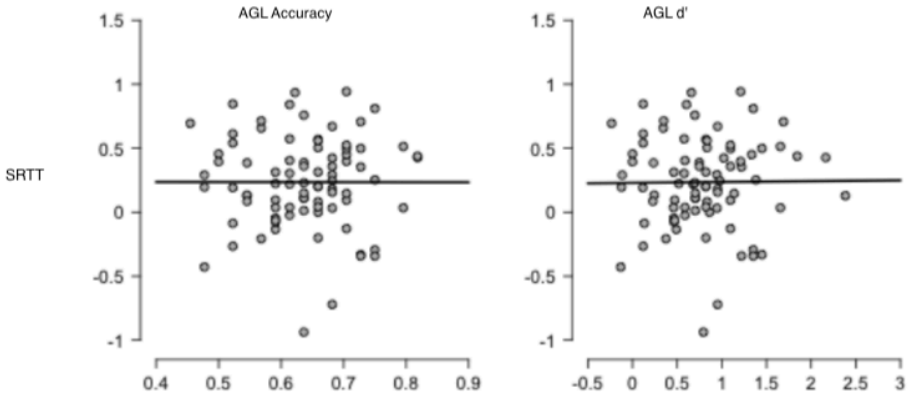


Figure 1: Artificial grammar used in the current task (taken from Knowlton & Squire, 1994).

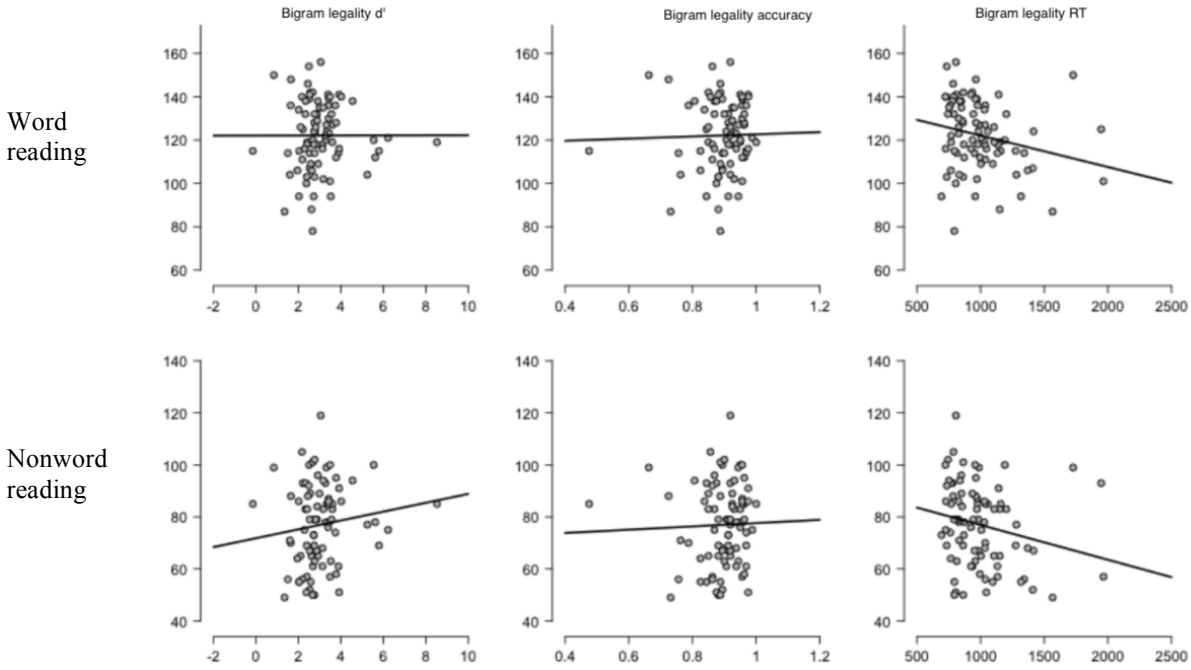
Figure 2


SRTT
AGL Accuracy
AGL d'
Word reading
Nonword reading

A


AGL Accuracy
AGL d'
SRTT

B



C

Figure 2: Scatterplots showing the relationships between the critical variables: (A) Between reading ability and statistical learning, (B) between the two statistical learning tasks, and (C) between bigram sensitivity and reading ability. For SRTT, values >0 reflect that learning occurred, for AGL and bigram legality accuracy, 0.5 reflects chance level, and for AGL and bigram legality d', 0 reflects chance level. For word and nonword reading, the axis reflects the number of words read correctly within 1 minute, and for bigram legality, the average number of *ms* before response.

Figure 3

Response times across blocks

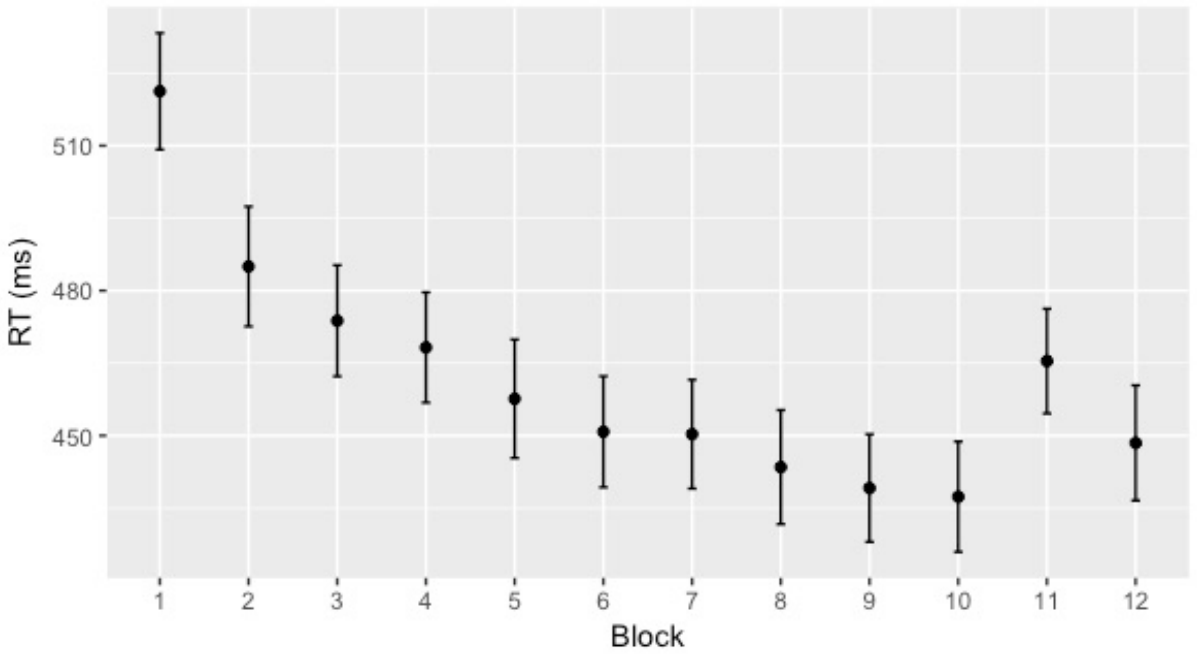


Figure 3: Participants' average performance (trimmed RT) on the SRTT across blocks. Error bars = SEM. The eleventh block contained the random sequence.

Table 1

Table 1: Descriptive Statistics for the obtained variables

	Age	Word reading (SLRT percentile)	Nonword reading (SLRT percentile)	Word reading (correct items per minute)	Nonword reading (correct items per minute)	Reading difference (z-score difference)	SRTT difference (z-score difference)	AGL accuracy (proportion correct)	AGL d'	Bigram legality accuracy (proportion correct)	Bigram legality d'	Bigram legality RT (ms)	Choice Reaction Time Task accuracy (percentage correct)	Choice Reaction Time Task RT (ms)
Mean	27.7	52.4	52.7	122.1	76.9	-0.03	0.23	0.64	0.78	0.89	3.02	996.6	94.2	439.3
Std. Deviation	8.7	28.3	27.9	16.1	15.7	0.71	0.36	0.08	0.51	0.08	1.18	253.5	4.5	62.16
Minimum	19	2	8	78	49	-1.68	-0.94	0.45	0.23	0.48	-0.15	692.3	70.0	319.0
Maximum	62	99	98	156	119	1.27	0.94	0.81	2.38	1.00	8.53	1965	100.0	640.0
Chance level							0	0.5		0.5				
Average deviation from chance level							$t(83) = 6.0$, $p < 0.0001$	$t(83) = 15.3$, $p < 0.0001$		$t(83) = 46.4$, $p < 0.0001$				
Number of participants (out of 84) significantly above chance ($p < 0.05$)							65	40		83				

Note: For age, there are 6 missing values. Average deviation from chance level was calculated with a one-sample *t*-test. For scripts to calculate the number of participants significantly above chance, see osf.io/fqdnh

Table 2: Correlation coefficients with Bayes Factors

		2	3	4	5	6	7	8	9	10	11	12
1. Word reading	Coefficient *r*	**0.75**	**0.4**	*-0.02*	*0.1*	0.14	*0.001*	*0.02*	-0.23	-0.25	**-0.31**	*-0.003*
	Bayes Factor	**>1,000**	**200**	*0.21*	*0.31*	0.43	*0.2*	*0.21*	1.64	2.33	**10.14**	*0.2*
2. Pseudoword reading	Coefficient *r*	—	**-0.312**	*0.02*	0.16	0.17	0.13	*0.03*	-0.22	**-0.29**	**-0.26**	0.19
	Bayes Factor	—	**11**	*0.21*	0.53	0.63	0.39	*0.21*	1.32	**5.23**	**3.22**	0.86
3. Reading difference	Coefficient *r*		—	*-0.06*	*-0.07*	*-0.03*	-0.17	*-0.009*	*-0.03*	*0.03*	*-0.08*	**-0.27**
	Bayes Factor		—	*0.23*	*0.24*	*0.21*	0.68	*0.2*	*0.21*	*0.22*	*0.26*	**3.51**
4. SRTT difference	Coefficient *r*			—	*-0.001*	*0.01*	*0.1*	0.12	*-0.02*	-0.17	-0.15	*0.02*
	Bayes Factor			—	*0.2*	*0.2*	*0.3*	0.35	*0.21*	0.59	0.5	*0.21*
5. AGL accuracy	Coefficient *r*				—	**0.88**	*0.003*	*0.01*	*-0.07*	*-0.008*	*-0.002*	0.18
	Bayes Factor				—	**>1,000**	*0.2*	*0.2*	*0.25*	*0.21*	*0.2*	0.78
6. AGL *d'*	Coefficient *r*					—	*-0.03*	*-0.008*	*-0.02*	*0.02*	*-0.01*	0.19
	Bayes Factor					—	*0.21*	*0.2*	*0.21*	*0.21*	*0.2*	0.8
7. Bigram legality *d'*	Coefficient *r*						—	**0.73**	-0.13	*-0.09*	-0.13	*0.1*
	Bayes Factor						—	**>1000**	0.41	*0.28*	0.38	*0.3*
8. Bigram legality accuracy	Coefficient *r*							—	-0.24	-0.24	**-0.27**	*0.07*
	Bayes Factor							—	2	1.8	**4**	*0.24*
9. Bigram legality RT	Coefficient *r*								—	**0.35**	0.23	*-0.02*
	Bayes Factor								—	**22.1**	1.54	*0.21*
10. Age	Coefficient *r*									—	**0.52**	*0.02*
	Bayes Factor									—	**>1,000**	*0.21*
11. Choice Reaction Time Task RT	Coefficient *r*										—	**0.31**
	Bayes Factor										—	**11**
12. Choice reaction time task accuracy	Coefficient *r*											—
	Bayes Factor											—

Note: The Coefficient *r* is the standard zero-order correlation coefficient. BF values > 3 are marked in bold green font and provide evidence for the presence of a correlation. BF values < 1/3 are marked in italic red font and provide evidence against the presence of a correlation. Values in-between provide insufficient evidence for a conclusion. Word and nonword reading scores are the raw values rather than *z*-scores.