Running head: Sleep and language plasticity

Sleep underpins the plasticity of language production
M. Gareth Gaskell[1], Jill Warker[2], Shane Lindsay[1], Rebecca Frost[1], James Guest[1], Reza Snowdon[1], Abigail Stackhouse[1]
[1]University of York, UK
[2]University of Scranton, USA

The constraints that govern acceptable phoneme combinations in speech perception and production have considerable plasticity. We addressed whether sleep influences the acquisition of new constraints and their integration into the speech production system. Participants repeated sequences in which two phonemes were artificially restricted to syllable onset or coda depending on the vowel. After 48 sequences, participants either had a 90-minute nap or remained awake. Participants then repeated 96 sequences to examine implicit constraint learning, and were tested for constraint generalization in a forced-choice task. The nap but not the wake group produced speech errors at test that were consistent with training constraints. Furthermore, only the nap group generalized this knowledge to new materials. Polysomnography data showed that implicit constraint learning was associated with slow-wave sleep. These results show that sleep facilitates the integration of new linguistic knowledge with existing production constraints. We interpret these data in the context of systems consolidation models of sleep.

Recent studies have provided numerous demonstrations of the enduring plasticity underlying language processing. This allows us to learn languages (Mårtensson et al., 2012), new words in our language (Gaskell & Dumay, 2003), and adapt to new accents or environments (Bradlow & Bent, 2008; Norris, McQueen, & Cutler, 2003). Some studies have revealed that plasticity in language may involve memory consolidation during sleep (e.g., Dumay & Gaskell, 2007; Gomez, Bootzin, & Nadel, 2006). These studies fit with a broader literature highlighting a critical role for sleep in the consolidation of many forms of memory, including paired-associate learning (Fenn & Hambrick, 2012; Marshall, Helgadottir, Molle, & Born, 2006), procedural knowledge (Plihal & Born, 1997), and statistical learning (Durrant, Taylor, Cairney, & Lewis, 2011).

Studies of sleep and plasticity in language have focused on two particular areas: word learning (Dumay & Gaskell, 2007; Tamminen, Payne, Stickgold, Wamsley, & Gaskell, 2010) and acquisition of grammar (Gomez et al., 2006; Hupbach, Gomez, Bootzin, & Nadel, 2009). The grammar data suggest that sleep can assist in the abstraction of structure from new instances of sequences. The word learning data show that sleep can facilitate the embedding of novel words in the mental lexicon. Both areas are noteworthy because they provide evidence that sleep does more than merely strengthen new memories (Walker & Stickgold, 2010). Rather, they show that sleep facilitates an integrative process that situates the newly learned information in its appropriate context in long-term memory. This is a key prediction of systems models of sleep and memory consolidation (Diekelmann & Born, 2010; Walker & Stickgold, 2006), which propose that slow-wave sleep (SWS) in particular provides a means of integrating episodic hippocampal memories with long-term neocortical memory (Alvarez & Squire, 1994).

However, the generality of these findings remains in question, both within the language domain and in the memory consolidation literature. Within language, it is important to determine what other aspects of learning and plasticity are related to sleep and consolidation. More broadly, there is a particular need for further paradigms that involve relating newly learned information to existing stable memories, given the value of such data to the predictions of systems consolidation models. In order to further understand the relationship between language plasticity and memory consolidation the experiment presented here employed an implicit learning paradigm whereby recent experiences are incorporated into the language production system and constrain the nature of subsequent speech errors.

All languages have phonotactic constraints, which set restrictions on the acceptable placement of phonemes or combinations of phonemes. First-order constraints concern the syllabic positions of isolated phonemes. For example, in English /ŋ/ is acceptable at syllable coda (e.g., *sing*) but not onset (e.g., *ngis*). Second-order constraints involve phoneme combinations (e.g., /r/ is acceptable in English at syllable onset in combination with /d/ (e.g., *drink*), but not with /s/ (e.g., *srink*); for /l/ the acceptable combinations are reversed (e.g., *dlink* versus *slink*). These constraints affect performance in both perception (Dupoux, Kakehi, Hirose, Pallier, & Mehler, 1999) and production, with speech errors that violate the phonotactic constraints of a speaker's language (e.g., *srink*) being extremely rare (Wells, 1951).

The susceptibility of phonotactic constraints to recent experience was exposed in a seminal study by Dell, Reed, Adams, and Meyer (2000). They tested whether speech errors might be influenced by laboratory-induced first-order constraints, in which certain consonants were artificially restricted to particular syllable positions over several sessions. Speech errors began to mirror these constraints, suggesting that participants had begun to incorporate new phonotactic constraints into their productions as a consequence of their recent experience. These first-order effects, based on simple occurrence frequency of phonemes, emerged quickly, on the first day of testing. This learning appears to be implicit, in that informing participants of the nature of the constraints made no difference to their speech errors, and uninformed participants were unable to describe the constraints when prompted.

Warker and Dell (2006) found that second-order constraints could also influence speech errors, but the effects of these more subtle constraints tended to emerge later in the experiment, on the second day of testing. Warker (2013) looked at whether this delayed emergence was a consequence of extra training or the opportunity to consolidate the initial learning, and found that exposure level had a modest effect but the opportunity for consolidation was particularly important. Although Warker could not implicate sleep over time in any consolidation process that resulted in constraint acquisition, there are clear precedents in the literature favoring such an account, such as studies showing that sleep can enhance statistical learning (Durrant et al., 2011).

In the current work, we used the phonotactic learning paradigm to address both the mechanisms of language plasticity and nature of memory consolidation. The use of this paradigm allows examination of the extent to which plasticity in speech production makes use of sleep to consolidate new phonotactic information. Furthermore, it addresses the extent to which newly learned materials (phonotactic constraints) impact the operation of a well-established neocortical mechanism (errors in the speech production system). The study of speech errors has had a long history of informing our understanding of the normal mechanisms of word production (Levelt, 1999).

We examined the association between sleep and acquisition of second-order phonotactic constraints using a nap design. Following Warker (2013), we asked two groups of participants to repeat a short block of syllable sequences containing second-order constraints where the syllable position of two phonemes (/f/ and /s/) was restricted based on the identity of the syllable's vowel. One group then slept in our lab with polysomnography applied. The other group remained awake for a similar period, and then both groups were tested for the impact of constraints on their speech errors by undergoing two further blocks of syllable repetition. Based on Warker's results, we expected that the wake group would show no evidence of the within-experiment second-order constraints in their speech errors, but our prediction for the sleep group was that sleep would facilitate consolidation of the newly learnt phonotactic knowledge, leading to a change in speech error patterns. We further predicted, on the basis of existing models of declarative memory consolidation (Diekelmann & Born, 2010), that SWS duration would be correlated with the strength of this change. In a post-test, we looked at whether sleep influenced the extent to which participants could generalize their newly acquired phonotactic knowledge in a perceptual task,

given prior evidence for the role of sleep in abstraction of statistical knowledge (Durrant et al., 2011). This generalization task asked participants to select between untrained syllables that either corresponded to or violated the within-experiment constraints.

Method

Participants

Data were collected for thirty-eight adults (20 male, 18 female) with a mean age of 22 (range 18-27), with one further participant excluded due to experimenter error.

Design and Materials

Sleep was manipulated between participants (sleep vs. wake) and all participants received two sessions of syllable repetition (pre- and post-interval). The pre-interval session (one block of 48 sequences) was intended to be short enough to provide sufficient information about the within-experiment constraints, but without inducing an effect on speech errors within that short session. The post-interval session was longer (two 48-sequence blocks), but previous research suggests that the wake group should still show no effect of constraints on speech errors (Warker, 2013).

The sequences comprised 4 CVC syllables presented visually to be read aloud in time with a metronome (e.g., *haf kan sang gam*). The vowel (/æ/ or /ɪ/) remained the same in each sequence but alternated between sequences. Of the 8 consonants, /h/ was always at syllable-onset, /ŋ/ was always at syllable-coda, the positions of /f/ and /s/ depended on the vowel (restricted), and the remaining consonants (/n/, /m/, /g/ and /k/) could appear as either syllable-onset or coda (unrestricted). All consonants were used once per sequence. Two versions of the experiment language counterbalanced the constraints between participants. In one version, /s/ was restricted to onset and /f/ to coda when the vowel was /æ/, whereas /f/ was restricted to onset and /s/ to coda when the vowel was /ɪ/. Version 2 reversed these constraints. A sequence generator applied these rules pseudorandomly to create the syllable sequences (Warker, 2013).

A 2-alternative forced choice (2AFC) recognition task assessed explicit memory of the syllables. Twenty syllables containing restricted consonants were taken from one language version and were randomly paired with 20 equivalent syllables from the second version (e.g. *fim*, *hif*). The two vowels were equally represented. Participants were presented with both syllables and asked to select the syllable from the experiment.

A 2AFC generalization task assessed participants' ability to make judgments about the correspondence between novel visually presented consonant-vowel-consonant sequences (CVCs) and the repetition materials. Forty CVCs were used in which the vowel was either *a* or *i* (20 of each). One consonant came from the restricted set (*s* or *f*; 20 of each) and the other was from a larger set intended to contain only new consonants that were not used in the repetition task. Twenty target CVCs adhered to the second-order constraints used in the repetition task, with each constraint equally represented. The foils followed the opposite constraints. Targets were pseudorandomly paired with foils and participants had to select which CVC best fitted the language they had seen. Due to experimenter error, the set of "new" consonants contained one consonant that was used in the repetition task (*m*) alongside four correctly chosen new

consonants (*t, d, l*, *z*). This actually added an interesting complexity to the task. In the majority of trials (12 or 13 depending on list) the generalization pairs both contained one restricted consonant and one novel consonant (categorized as "both unfamiliar", e.g., *das-fid*). For the remainder one of the CVCs used only consonants from the repetition task whereas the other had a novel consonant (e.g., *mas-fid*). Half these items (3 or 4) used the unfamiliar consonant in the target CVC ("target unfamiliar") and half used it in the foil ("target familiar"). All these items were retained for analysis, but the data were categorized by the type of trial.[1]

Materials relating to assessment of sleepiness, explicit knowledge of the sequences and reading ability are provided in the Supplemental Materials available online.

Procedure.

On arrival, a reading ability test was administered, and participants were pseudorandomly allocated to the sleep or wake condition. Sleep participants were wired up using a 9 EEG montage (see Supplementary Materials). During the first session of syllable repetition (48 sequences), participants sat in individual booths, with testing administered using PsychoPy (Peirce, 2007). Participants were instructed to read each sequence aloud in time to a metronome, once slowly (1 Hz) and 3 times at a faster pace (2.5 Hz). The metronome began 1 second after the sequence appeared on screen and was delivered via headphones. The ISI was 1 second. Spoken responses were recorded using a Sony MP3 recorder.

Participants were told if they would be napping only after training. Sleep participants were given 90 minutes to nap in a bedroom, followed by an interval of at least 20 minutes to reduce sleep inertia. Wake participants stayed in the testing booth and watched a 105-minute video followed by a 5-minute break. The video had little dialogue and was chosen to occupy the participants while minimizing interference with the language tasks. In the test session (25 mins) participants completed two 48-sequence blocks of syllable repetition followed by recognition and generalization tests.

As in Warker (2013), speech errors were coded based on whether a slipping consonant slipped to the same syllable position in a given sequence, creating a "same-position" error, or to a different syllable position, creating a "different-position" error (see Supplementary Materials).

Results

Speech errors

Speech errors for one participant who did not complete the task were excluded (see Table 1 for the remaining data). As in previous studies, the primary measure was the percentage of errors (per participant) involving a particular consonant type that were classed as "same-position" (i.e., that did not change syllable position). To provide the cleanest estimate of production error performance given small numbers of errors, data from the post-interval blocks were pooled for analysis. For unrestricted consonants, the percentage of errors that are same-position provides a baseline measure of the extent to which a participant's speech errors preserve syllable position. For restricted consonants, the key question is whether the measure rises

---

[1] One item in one list used *m* in both alternatives, making it equivalent to a recognition trial. This was excluded from analyses.

significantly above the unrestricted baseline rate, suggesting that newly acquired phonotactic constraints are influencing production errors. We will describe this difference between restricted and unrestricted same position percentages as the *phonotactic learning score*, with positive values suggesting that phonotactic learning has taken place.

Table 1.

Mean (and SD) number of errors of each type per participant during training and testing, with the percentages of these errors that preserved syllable position

	Sleep		Wake	
	Unrestricted	Restricted	Unrestricted	Restricted
Training				
Same position %	85% (11%)	80% (8%)	79% (10%)	90% (14%)
Total errors	45.3 (14.1)	9.5 (4.7)	50.2 (28.5)	8.8 (5.4)
Test Block 1				
Same position %	84% (11%)	92% (9%)	85% (11%)	86% (15%)
Total errors	32.3 (15.4)	4.9 (3.8)	39.8 (17.3)	6.1 (4.7)
Test Block 2				
Same position %	84% (9%)	94% (17%)	84% (8%)	85% (16%)
Total errors	28.3 (25.4)	6.0 (5.4)	29.5 (25.6)	6.8 (4.9)

Because participants and conditions varied substantially in the numbers of speech errors observed and the dependent measure is binary, we used mixed effects logistic regression using lme4 in R on the error data, using the syllabic position of the incorrect consonant (same or different) as the binary dependent variable. Session (pre-interval vs. post-interval), Restriction (restricted vs. unrestricted consonants) and Group (sleep vs. wake) were all effect coded (-.5, .5). Random effects for participants were incorporated with maximal inclusion of random slopes for within-participants variables and their interactions (Barr, Levy, Scheepers, & Tily, 2013), and *p*-values were calculated using Wald-z. The analysis revealed a significant three-way interaction between Session, Restriction and Group (β=1.97, *SE*=.50, *z*=3.31, *p*<.001), as well as less important effects of Session (β=.30, *SE*=.14, *z*=2.14, *p*=.03), Restriction (β=.65, *SE*=.20, *z*=3.34, *p*<.001) and Group by Session (β=.57, *SE*=.28, *z*=2.01, *p*=.044).

Planned comparisons examining the phonotactic learning score showed no effect prior to the interval for the sleep group (β=-.09, *SE*=.35, *z*=-.28, *p*=.78) but a clear effect after sleep (β=1.37, *SE*=.41, *z*=3.38, *p*<.001). When broken down by block, the effect was as strong in the first block after sleep (phonotactic learning score=14%; β=1.40, *SE*=.40, *z*=2.48, *p*=.013) as it was in the second block (phonotactic learning score=13%; β=1.02, *SE*=.43, *z*=2.35, *p*=.019)). Surprisingly, the wake group did show an effect prior to the interval (β=1.11, *SE*=.33, *z*=3.35, *p*<.001), but this did not remain significant when tested after the interval (β=.21, *SE*=.28, *z*=.74, *p*=.46). The effect for the wake group prior to the interval was unexpected for several reasons.

Prior experiments (Warker, 2013; Warker & Dell, 2006) suggest that such effects should not be present after exposure to so few sequences. Furthermore, the sleep group received identical training and showed no such effect at that point. Finally, the effect in the wake group was not observed after the interval, when the increase in the number of sequences provided greater power to observe such an effect. It seems likely, therefore, that it was a Type I error. It is worth noting that the number of restricted different-position errors wake participants made was near constant across the three blocks (training: 0.9; test 1: 0.8; test 2: 1.1), whereas there was a clear drop in this error type between training and test for the sleep participants (training: 1.9; test 1: 0.4; test 2: 0.4).

Figure 1. Same-position error rate across conditions for the subset of participants matched on training performance. Asterisk marks significant difference (p < .05). Error bars indicate standard error.
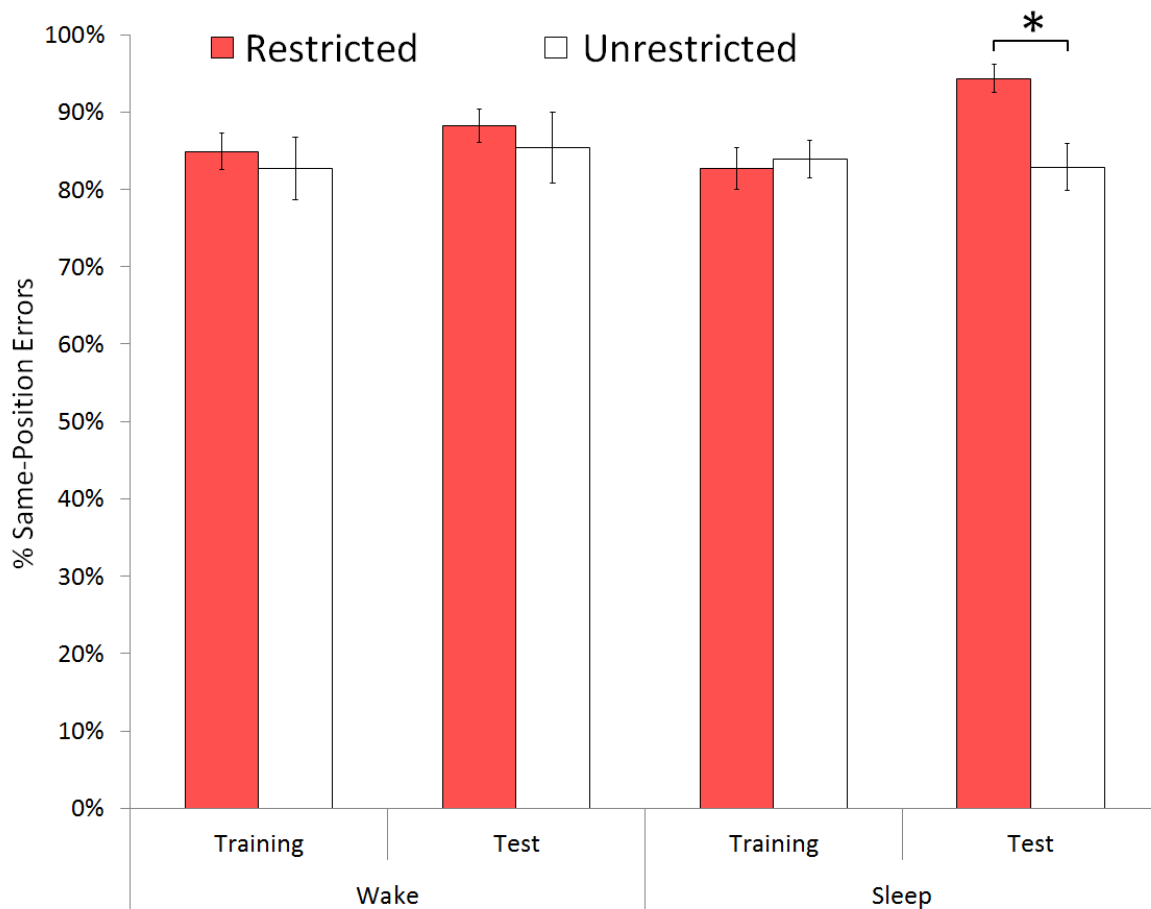
In order to examine whether the groups showed a different effect of the interval when matched on performance prior to the interval, a subset of 28 of the 38 participants was selected by progressively excluding outliers from both groups based on their pre-interval phonotactic learning score (see Figure 1). Unsurprisingly, for this subset there was no difference in the nature of their errors in training (no effect of Group, Restriction or interaction, all $ps>.17$). However,

the Group x Restriction x Session interaction remained significant for this subset (β=1.35, SE=.61, z=2.20, p=.028), with the Restriction x Session interaction found for participants who slept (β=1.15, SE=.50, z=2.31, p=.021) but not those who remained awake (β=-.66, SE=.44, z=-1.50, p=.13). After the interval there was still no effect for the wake group (β=-.19, SE=.24, z=-.81, p=.41), whereas the sleep group showed a clear effect (β=1.04, SE=.44, z=2.38, p=.017).

2AFC Tests

Recognition performance was uniformly good, with no differences between groups (sleep: 88%, wake: 87%) in correct response rates.

For the generalization test, mixed-effects logistic regression was run on the correct response rates, with Group (sleep vs. wake) and Trial Type (target familiar vs. target unfamiliar vs. both unfamiliar) as effect-coded independent variables (see Table 2). This showed that the familiarity of the two response options affected selection in that both target familiar (β=2.95, SE=.52, z=5.70, p<.001) and target unfamiliar (β=-1.94, SE=.31, z=-6.29, p<.001) conditions differed from the overall mean performance level. More importantly, the sleep group (69%) scored more highly than the wake group (56%; β=1.11, SE=.53, z=2.08, p=.038). In fact, only the sleep group showed familiarity with the constraints learned during the experiment and an ability to generalize those constraints, as assessed by comparisons with chance for the both-unfamiliar items (sleep: β=0.43, SE=.17, z=2.59, p=.0096; wake: (β=.09, SE=.15, z=.61, p=.54). Note however, that no participant was able to describe explicitly any of the constraints that they had learned (Supplementary Materials), showing that even the participants who could generalize their knowledge did so in a somewhat implicit way. Interactions between Group and Trial Type were not significant (target familiar: β=.79, SE=1.04, z=0.77, p=.44; target unfamiliar β=-.03, SE=.62, z=-.05, p=.96).

Table 2.
Mean (and SD) percentage of correct trials in the generalization test.

	Sleep	Wake
Target familiar	99% (3%)	92% (3%)
Target unfamiliar	47% (7%)	24% (7%)
Both unfamiliar	60% (4%)	52% (4%)

Correlation Analyses

Our key finding was that sleep participants showed a shift in their phonotactic learning score across the interval, suggesting that their use of phonotactic constraints in speech production had been altered by recent experience. Given the *a priori* evidence for an involvement of SWS in memory consolidation we predicted that sleep participants would show an association between SWS and this behavioral shift. A logistic regression analysis looking at the interaction of Session x Restriction x SWS duration did indeed reveal a significant relationship (β=0.06, SE=.02, z=3.10, p=.002).

Table 3.

Sleep stage data and their association with changes in phonotactic learning score

| Sleep Stage	Duration in mins (SD)	Correlations	
		Logistic regression β *(SE)*	Correlation *r*
Stage 1	8.6 (6.9)	-.06 (.04)	-.21
Stage 2	43.2 (18.9)	-.03 (.02)	-.06
SWS	14.6 (13.5)	.06** (.02)	.47*
REM	4.3 (6.8)	-.07* (.03)	-.49*
Sleep period	82.5 (8.2)	-.06* (.03)	-.22

Note: Beta values are for interactions with Session x Restriction. The *r*-values are by participant correlations with changes in phonotactic learning score across the interval and are given in addition to β for ease of interpretation. *: $p < .05$, **: $p < .01$ without Bonferroni correction.

When logistic regression interactions with other stages were examined *post hoc* unexpected effects in the opposite direction were found for REM sleep duration and overall sleep period were also found, perhaps partly because SWS and REM sleep durations were themselves negatively correlated ($r= -.51$, $p=.026$). However, a regression equation that included all the three-way interactions with sleep parameters left only the beta value for SWS duration as a significant independent predictor ($\beta=0.05$, $SE=.03$, $z=2.15$, $p=.03$).

To provide additional evidence, another measure of slow-wave activity was taken: the average log power in the delta band (0.3-4.0 Hz) across all sleep stages. This measure includes oscillatory activity for epochs that do not qualify as SWS. Delta activity was significantly correlated with time in SWS ($r=.79$, $p<.001$), but not REM sleep ($r= -.35$, $p=.14$) and is independent of sleep duration. This measure also showed a significant interaction with Session x Restriction ($\beta=0.006$, $SE=.003$, $z=2.64$, $p=.008$), supporting the proposition that the incorporation of new phonotactic constraints in the production system is associated with slow-waves (see also Figure 2 for equivalent correlation).

Given the group difference in generalization ability, correlation analyses were run for this measure. There was no association between time in slow-wave sleep duration and generalization ability ($r=.011$, $p=.965$) and *post hoc* analyses suggested that no other stage was associated with this effect.

Figure 2. Correlation between delta power (averaged across central and frontal electrodes, all sleep stages) and change in phonotactic learning score between training and test for the sleep participants.
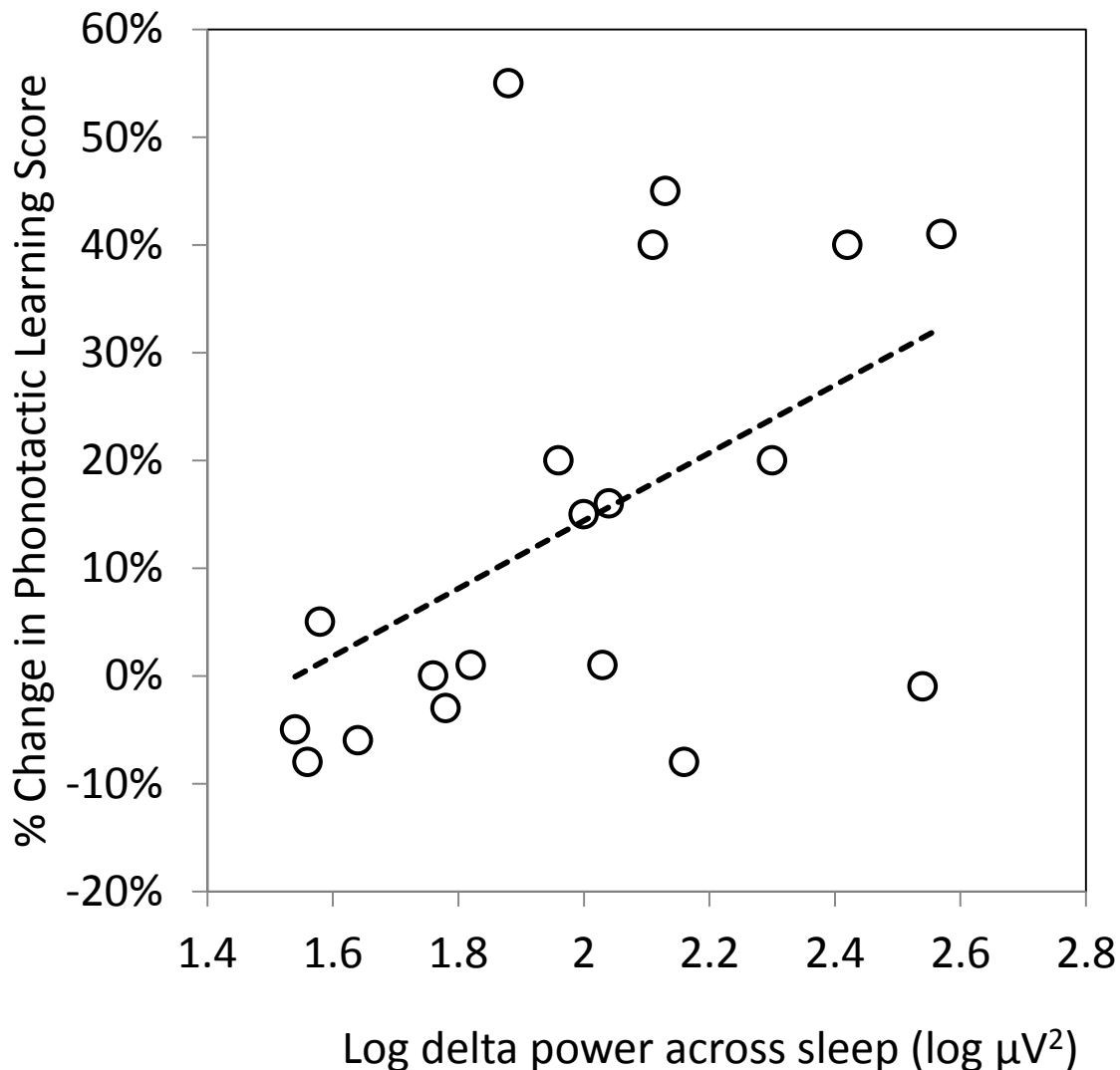
Discussion

Our experiment revealed a clear association between sleep and the engagement of new phonotactic knowledge in the production system. Participants who stayed awake between a short block of syllable repetition and a longer test session showed no influence of the recent experience on their pattern of speech errors during test. On the other hand, participants who slept in the intervening period demonstrated implicit effects of this recent experience in their speech errors, with this effect being present in both blocks of the test session. Furthermore, when confronted with two novel syllables, sleep participants were better able to select items that fit with the constraints from the experiment than wake participants, who remained at chance. Whereas this latter effect was not associated with any particular component of sleep, the implicit

effect on speech errors showed a positive correlation with time in SWS and a negative correlation with time in REM and sleep duration. On balance, the correlation with SWS is more likely to be causally implicated in the consolidation effect, given further correlations with EEG delta power.

The speech error result is compatible with a growing body of evidence that sleep has a particular role to play in the transformation rather than stabilization of recent memories, such that after sleep the new knowledge is in a more useful state for future actions (Stickgold & Walker, 2013). This argument has been made on the basis of several studies in which participants are required to generate inferences based on the integration of a set of new facts (Djonlagic et al., 2009; Ellenbogen, Hu, Payne, Titone, & Walker, 2007). Other studies have found that sleep benefits the integration of new memories with existing long-term memories (Dumay & Gaskell, 2007, 2012; Gaskell & Dumay, 2003).

The current paradigm has some facets of both the above forms of integration. The phonotactic constraints can only be formed by integrating statistical knowledge across many production sequences (cf. Djonlagic et al., 2009; cf. Ellenbogen et al., 2007). Furthermore, the implicit phonotactic effect was observed in normal speech errors, which are clearly also influenced by long-term knowledge of phonotactic constraints in the speaker's native language (Dell et al., 2000). This aspect can therefore be explained as integration of new experiences with long-term memory (cf. Dumay & Gaskell, 2007). More specifically, a participant enters the experiment with existing phonotactic knowledge from years of experience speaking English. During the experiment, the participant experiences a different phonological distribution that is not predicted by the existing phonotactic knowledge (e.g., the participant implicitly learns that /f/ tends to occur at syllable onset only if followed by /ɪ/). This distribution becomes encapsulated with the experimental context, allowing the recently experienced constraints to become integrated with existing phonotactic knowledge in long-term memory without overwriting that knowledge. During sleep, the association between syllable position and vowel identity is consolidated and becomes embedded in the language production system meaning that, when the participant returns to the experimental context, the newly acquired constraints influence production errors in combination with more established language-wide constraints. This ensures that the participant's language model is as up-to-date as possible, optimizing the nature of future linguistic interactions.

It is worth noting that integration as just described may not be the end of the integration process. Potentially, after further weeks or months of consolidation the new knowledge may become decontextualized so that it has a more general (and yet less dramatic) effect on speech production. A further point is that first-order constraints (e.g., /f/ is always a syllable onset) show integration effects without the need for sleep or extended experience. In this case it may be that the simplicity of the constraint (observable in simple positional frequency measures) allows it to become influential prior to sleep. Performance on a transitive inference task has shown a similar dissociation. Ellenbogen et al. (2007) found that simpler inferences (single mediating item) did not benefit from sleep over wake, but there was a sleep-specific boost in performance on more

complex inferences over two mediating items. The relative complexity or context-dependence of the second-order constraints here may be the reason that sleep is facilitates their emergence in production errors.

Participants presumably made use of their updated language model during the forced-choice generalization task. Their ability to apply implicitly acquired phonotactic knowledge to novel syllables in a separate, explicit decision suggests that participants did not merely memorize the syllables they experienced during the production task. Rather, it suggests participants learned a more abstract form of the second-order constraints (but still largely unavailable to verbal description). Given that those who slept performed better at this task, sleep may have a particular role to play in the reorganization and formation of abstract knowledge of information about the experimental constraint, allowing novel sequences to be accepted as matching those constraints (cf. Durrant et al., 2011).

The current data establish that the acquisition of these second-order constraints is associated with sleep rather than simply the passing of time (cf. Warker, 2013). This effect is not predicted by Warker and Dell's (2006) connectionist model of plasticity in speech production, but can be accommodated by a complementary systems variant (Davis & Gaskell, 2009; McClelland, McNaughton, & O'Reilly, 1995) in which the existing network is supplemented by a complementary network (modeled on the hippocampus) that specializes in the temporary acquisition of new mappings that cannot be easily acquired in the existing network (within-experiment second-order phonotactic constraints). Like other studies of this nature, the evidence that slow-wave sleep is related to restructuring is correlatory and does not imply a causal explanation. Nonetheless converging evidence from a wide range of methods (e.g., Ngo, Martinetz, Born, & Mölle, 2013) fits parsimoniously with an account in which sleep, and more particularly slow-wave sleep, provides a mechanism for transfer of knowledge to the main network just as in the systems consolidation models referred to above (e.g., Diekelmann & Born, 2010). McClelland (2013) demonstrated how a complementary systems model explains the incorporation of new memories into existing schema (Tse et al., 2007; van Kesteren et al., 2013). The current paradigm can be described in the same terms. Participants had pre-established long-term knowledge of the likelihood of encountering various phoneme combinations in different syllable positions. The syllable repetition task altered these expectations in specific ways, and sleep provided the means for offline consolidation of this knowledge, updating the phonotactic schema.

In conclusion, we have shown that sleep-associated consolidation is beneficial to the integration of complex phonotactic constraints in speech production. After a small amount of exposure to sequences that subtly conflict with participants' phonotactic knowledge of English syllables, sleep provides a means for the updating of their language models. Participants who slept exhibited changes in their pattern of speech errors, particularly those with high levels of slow-wave sleep, whereas those who stayed awake showed no change. These results fit with a general model of language learning and processing in which complementary systems underpin the substantial plasticity that remains in the system even during adulthood.

Authorship

M.G. Gaskell developed the study concept. All authors were involved in the design of the experiment. R. Frost., J. Guest., R. Snowdon and A. Stackhouse ran the experiment and preliminary analyses under the supervision of M. G. Gaskell and S. Lindsay. M. G. Gaskell ran final analyses. M. G. Gaskell wrote the paper, based on an initial draft by R. Frost. J. Warker and S. Lindsay provided critical feedback. All contactable authors approved the final version of the paper for submission.

Acknowledgements

References

Alvarez, P., & Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: a simple network model. *Proceedings of the National Academy of Sciences, 91*(15), 7041-7045.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255-278. doi: http://dx.doi.org/10.1016/j.jml.2012.11.001

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707-729. doi: http://dx.doi.org/10.1016/j.cognition.2007.04.005

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: neural and behavioural evidence. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences, 364*(1536), 3773-3800. doi: 10.1098/rstb.2009.0111

Dell, G. S., Reed, K. D., Adams, D. R., & Meyer, A. S. (2000). Speech errors, phonotactic constraints, and implicit learning: a study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(6), 1355.

Diekelmann, S., & Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience, 11*(2), 114-126. doi: 10.1038/nrn2762

Djonlagic, I., Rosenfeld, A., Shohamy, D., Myers, C., Gluck, M., & Stickgold, R. (2009). Sleep enhances category learning. *Learning & Memory, 16*(12), 751-755.

Dumay, N., & Gaskell, M. G. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science, 18*(1), 35-39.

Dumay, N., & Gaskell, M. G. (2012). Overnight lexical consolidation revealed by speech segmentation. *Cognition, 123*(1), 119-132. doi: DOI 10.1016/j.cognition.2011.12.009

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance, 25*(6), 1568.

Durrant, S. J., Taylor, C., Cairney, S., & Lewis, P. A. (2011). Sleep-dependent consolidation of statistical learning. *Neuropsychologia, 49*(5), 1322-1331.

Ellenbogen, J. M., Hu, P. T., Payne, J. D., Titone, D., & Walker, M. P. (2007). Human relational memory requires time and sleep. *Proceedings of the National Academy of Sciences of the United States of America, 104*(18), 7723.

Fenn, K. M., & Hambrick, D. Z. (2012). Individual differences in working memory capacity predict sleep-dependent memory consolidation. *Journal of Experimental Psychology: General, 141*(3), 404.

Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition, 89*(2), 105-132.

Gomez, R. L., Bootzin, R. R., & Nadel, L. (2006). Naps promote abstraction in language-learning infants. *Psychological Science, 17*(8), 670-674.

Hupbach, A., Gomez, R. L., Bootzin, R. R., & Nadel, L. (2009). Nap-dependent learning in infants. *Developmental Science, 12*(6), 1007-1012. doi: 10.1111/j.1467-7687.2009.00837.x

Levelt, W. J. M. (1999). Models of word production. *Trends in Cognitive Sciences, 3*(6), 223-232. doi: http://dx.doi.org/10.1016/S1364-6613(99)01319-4

Marshall, L., Helgadottir, H., Molle, M., & Born, J. (2006). Boosting slow oscillations during sleep potentiates memory. *Nature, 444*(7119), 610-613.

Mårtensson, J., Eriksson, J., Bodammer, N. C., Lindgren, M., Johansson, M., Nyberg, L., & Lövdén, M. (2012). Growth of language-related brain areas after foreign language learning. *NeuroImage, 63*(1), 240-244. doi: http://dx.doi.org/10.1016/j.neuroimage.2012.06.043

McClelland, J. L. (2013). Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *Journal of Experimental Psychology: General, 142*(4), 1190-1210.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning-systems in the hippocampus and neocortex - insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*(3), 419-457.

Ngo, H.-Viet V., Martinetz, T., Born, J., & Mölle, M. (2013). Auditory Closed-Loop Stimulation of the Sleep Slow Oscillation Enhances Memory. *Neuron, 78*(3), 545-553. doi: http://dx.doi.org/10.1016/j.neuron.2013.03.006

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*(2), 204-238.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1), 8-13.

Plihal, W., & Born, J. (1997). Effects of early and late nocturnal sleep on declarative and procedural memory. *Journal of Cognitive Neuroscience, 9*(4), 534-547.

Stickgold, R., & Walker, M. P. (2013). Sleep-dependent memory triage: evolving generalization through selective processing. *Nature Neuroscience, 16*(2), 139-145.

Tamminen, J., Payne, J. D., Stickgold, R., Wamsley, E. J., & Gaskell, M. G. (2010). Sleep spindle activity is associated with the integration of new memories and existing knowledge. *Journal of Neuroscience, 30*(43), 14356-14360. doi: 10.1523/jneurosci.3028-10.2010

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., . . . Morris, R. G. M. (2007). Schemas and memory consolidation. *Science, 316*(5821), 76-82.

van Kesteren, M. T., Beul, S. F., Takashima, A., Henson, R. N., Ruiter, D. J., & Fernández, G. (2013). Differential roles for medial temporal and medial prefrontal cortices in schema-dependent encoding: from congruent to incongruent. *Neuropsychologia*.

Walker, M. P., & Stickgold, R. (2006). Sleep, memory, and plasticity. *Annual Review of Psychology, 57*, 139-166.

Walker, M. P., & Stickgold, R. (2010). Overnight alchemy: sleep-dependent memory evolution. *Nature Reviews Neuroscience, 11*(3). doi: 10.1038/nrn2762-c1

Warker, J. A. (2013). Investigating the retention and time course of phonotactic constraint learning from production experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(1), 96.

Warker, J. A., & Dell, G. S. (2006). Speech errors reflect newly learned phonotactic constraints. *Journal of Experimental Psychology-Learning Memory and Cognition, 32*(2), 387-398. doi: 10.1037/0278-7393.32.2.387

Wells, R. (1951). Predicting slips of the tongue. Reprinted from The Yale Scientific Magazine XXVI, 9-30. *Speech Errors as Linguistic Evidence. The Hague: Mouton*.

Supplemental Online Material

Supplementary Methods

Participants (supplementary)

Recognition and generalization data for one wake participant were lost due to equipment failure. All were right-handed monolingual native-English-speakers. On the day of the experiment, participants were asked to wake before 8 am and refrain from napping or consuming any caffeinated products. Participants were pseudo-randomly allocated to sleep and wake groups, and received £10 for participation. The desired number of participants was determined (20 per group) with reference to previous studies that had used the speech error paradigm and that had shown correlations with sleep measures. Two no-shows left us with 19 participants per group. The research was approved by the Department of Psychology Ethics Committee at the University of York.

Design and Materials (supplementary)

The Test of Word Reading Efficiency (TOWRE) was administered to provide an indicator of participants' reading ability (Torgesen, Wagner, & Rashotte, 1999). Upon completion of the experiment, participants were given a short post-test questionnaire on their sleep habits over the previous 4 days in order to determine whether any participants had unusual sleep habits (none had). A subset of participants (N = 24) completed the Stanford Sleepiness Scale (Hoddes, Dement, & Zarcone, 1972) at four points during the experiment (before and after both training and testing) to provide a measure of any major changes in sleepiness and motivation that may have affected performance. These participants were also given a post-test to determine whether they had noticed any rules in the items they were presented.

Error Coding

Production errors were coded as in previous studies of this kind (e.g., Warker, 2013) using two error types (same-position and different-position) applied to the restricted (/f/, /s/) and unrestricted (/k/, /g/, /m/, /n/) consonants. An error in which a slipping consonant retained the same syllable position in a given sequence (e.g., mistakenly saying *sang sam* instead of *sang gam*) was described as a same-position error whereas an error that involved a change of syllable position (e.g., mistakenly saying *sang gas* instead of *sang gam*) was classified as a different-position error. Most speech errors tend to preserve the syllable position of the slipping consonant, although of course individuals vary in their adherence to this constraint. For unrestricted consonants, the percentage of all errors (same-position + different-position) that were same-position errors establishes a baseline for this tendency in the absence of any new constraints. For example, a typical participant in our experiment (see Table 1 of the main article) might make 45 speech errors involving unrestricted consonants in the course of the training session. If 7 of those errors involved a change of position, with the remaining 38 being same-position errors, then the same-position percentage for that participant would be 38/45 = 84%. For restricted consonants, a value for this percentage that is higher than the baseline level for unrestricted consonants would indicate that the participant's speech errors were adhering to the within-experiment constraints (given that when errors occurred, the uttered syllable always had

the same vowel as the intended syllable). For example, if the same participant in training made 85% same-position errors on the restricted consonants during training then their phonotactic learning score at that point in the experiment would be 85%-84% = 1%. Coding was carried out by one of the authors, with cross-checking of a subset of error recordings by two other coders showing high concordance rates. When the primary coder found no error, the agreement rate was 97%, whereas when the primary coder found an error the agreement rate was between 71% and 78%, similar to other studies using this methodology.

Sleep Recording

Sleep participants were wired up using 9 EEG montage (F3-A2, F4-A1, C3-A2, C4-A1, O1-A2, O2-A1) using Ag-AgCl cup electrodes, which were applied according to the international 10-20 system. Frontal and central electrodes were selected in order to observe slow-wave and spindle activity, and delta power analyses were carried out across all four electrodes. Occipital electrodes were used to observe alpha activity for visual scoring of sleep, but were not further analysed. Two electro-oculogram electrodes and two electromyogram electrodes were also applied. Impedance levels, sampling rates and filter settings were set according to the American Academy of Sleep Science Manual (Iber, 2007).

Sleep Scoring

Sleep data were recorded digitally at a sample rate of 200 Hz using an Embla N7000 system with RemLogic 3.0. Sleep data for each participant were manually scored in RemLogic by two independent coders. Data were scored in 30-second epochs using standard criteria (Iber, 2007), with close inter-scorer agreement. Sleep spindle data were analyzed, but did not show a significant correlation with the key behavioral results.

Supplementary Results

Potential confounds: TOWRE

Performance on the TOWRE did not correlate with any sleep stage or measure of power, but given modest non-significant correlations between this variable and measures of slow-wave sleep (SWS minutes, $r = .30$; delta power, $r = .37$; SWA, $r = .37$) it is worth noting that the correlations between the slow-wave sleep measures and the change in speech error patterns all remained significant when TOWRE performance was partialed out.

Potential confounds: Sleepiness

For the subgroup who provided sleepiness scale data, the scores were analyzed using a three-way ANOVA with a between-participants variable, Group (sleep vs. wake), and two within-participants variables: Session (training vs. test) and Order (before vs. after the tests). This analysis revealed a marginal main effect of Session [$F(1, 19) = 3.9$, $p = .064$, $\eta_p^2 = .17$] and two significant interactions. All other effects and interactions did not approach significance. The interaction between Order and Session [$F(1, 19) = 4.5$, $p = .048$, $\eta_p^2 = .19$] was not of theoretical value, but the Session x Group interaction [$F(1, 19) = 15.3$, $p < .001$, $\eta_p^2 = .45$] was more relevant to the key behavioral effects. Whereas wake participants showed a non-significant increase in their rated sleepiness (from 2.8 to 3.1) between the two sessions, the sleep group showed a significant reduction in their ratings (from 3.2 to 2.3). Given that the second session

was after the sleep group had rested, this change was not surprising, but it could perhaps be argued that the change in the type of speech errors for this group was a consequence not of memory consolidation during sleep but somehow of the greater alertness that this group had after sleep. We addressed this possibility in several ways. First, we calculated the change in sleepiness across sessions for both wake and sleep participants and correlated this with the two key behavioral effects that we found. Neither correlation was significant (phonotactic measure: $r =$ .343, $p = .13$; generalization: $r = .19$, $p = .41$). Although these non-significant effects are partly reassuring, it is hard to rule out a potential confound on the basis of failure to reject the null hypothesis. Consequently a second form of analysis examined whether, for the sleep group, the correlations between the key sleep variables and performance changes held when changes in sleepiness were partialed out. The significant correlations between slow-wave sleep measures and the phonotactic effect in fact got numerically stronger rather than weaker when controlling for changes in sleepiness (SWS Duration: $r = .69$, $p = .029$; Delta Power: $r = .66$, $p = .038$). A similar result was obtained using logistic regression. We examined the model containing the 2-way interaction between Session and Restriction, the 3-way interaction between Session, Restriction and SWS Duration, the 3-way interaction between Session, Restriction and Sleepiness Change, and the 4-way interaction between all these variables, along with an intercept. If sleepiness was underlying the change in performance then we would expect interactions involving this variable to be significant and not interactions without this term. In fact only the Session x Restriction x SWS Duration interaction showed an independent effect on speech errors (β=0.07, SE=.04, z=2.03, p=.04). Equivalent analyses replacing SWS Duration with Delta Power produced the same pattern of results (Session x Restriction x Delta Power: β=0.01, SE=.005, z=2.08, p=.04). In sum, there was good evidence that the change in performance for the sleep group was associated with structural properties of the nap rather than changes in sleepiness.

<u>Explicit knowledge</u>

The same subgroup were asked to write down any rules that they could determine about the items in the experiment. Some participants noticed that the vowels alternated between *a* and *i* and some noticed that the items began and ended with a consonant. One participant incorrectly thought that the *i* sequences had more *f* and *s* consonants. However, most participants did not volunteer any observations, and none of them noticed any association between consonants and vowels. Hence, as in previous studies, the newly extracted knowledge was implicit in that it was not describable.

References

Hoddes, E., Dement, W., & Zarcone, V. (1972). The development and use of the Stanford Sleepiness Scale (SSS). *Psychophysiology, 9*(150), 431-436.

Iber, C. (2007). *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*: American Academy of Sleep Medicine.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). Test of word reading efficiency: Austin, TX: Pro-Ed.

Warker, J. A. (2013). Investigating the retention and time course of phonotactic constraint learning from production experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(1), 96.