Mattys Sven (Orcid ID: 0000-0001-6542-585X)

Janssen Steve (Orcid ID: 0000-0002-3100-128X)

WORKING MEMORY AND SECOND-LANGUAGE ACCENT ACQUISITION

Sven L. Mattys

and

Alan Baddeley

Department of Psychology

University of York, UK

Corresponding authors: Sven Mattys, sven.mattys@york.ac.uk and Alan Baddeley, alan.baddeley@york.ac.uk

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/acp.3554

Abstract

We explored the proposal that overt repetition of verbal information improves the acquisition of a native accent in a second language. Mandarin-speaking Chinese learners of English were recorded while repeating and reading out English sentences before and after one of three treatments: (1) Repeating native English sentences subvocally, "covert repetition," (2) Repeating sentences out loud, "overt repetition," and (3) Unfilled time of comparable duration. The sentences were rated by English speakers for their nativeness, fluency, and intelligibility. Overt repetition improved accent rating for read-out sentences. Covert repetition did not. Neither condition improved accent rating for repeated sentences, suggesting that immediate repetition depends on temporary rather than long-term representations. Our results provide some support for the use of overt repetition in accent learning. From a theoretical perspective, an interpretation is proposed in terms of a separation between phonological and articulatory coding within the phonological loop component of working memory.

Keywords: accent learning; covert repetition; overt repetition; shadowing; L2 pronunciation; working memory; phonological loop

Introduction

In a recent review, Munro and Derwing (2015) suggested that the study of pronunciation was neglected both in second-language teaching and in applied research. In the case of teaching, the field has seen a surge of interest in recent years, resulting in surveys of current practice in Canada (Foote et al., 2011), in Europe (Henderson, Frost, et al., 2012), and in Brazil (Buss, 2016). However, there remains a comparative lack of studies that attempt to link theory and practice by evaluating the effectiveness of a given method within the context of a theoretical approach to the underlying mechanisms behind pronunciation and accent learning. We report an initial attempt to do so within a broad multi-component working memory framework. Here, we refer to accent as the manner of speaking typically associated with a group or region, but we do not attempt to separate aspects such as "nativeness" and comprehensibility (Levis, 2015).

We are mainly concerned with the role of working memory in accent learning. Working memory is a system responsible for the temporary maintenance of information while performing complex tasks (Baddeley & Hitch, 1974; Baddeley & Hitch, 2018). There is considerable evidence for the involvement in language acquisition of one component of working memory, the phonological loop (Baddeley, Gathercole & Papagno, 1998). The phonological loop holds acoustically-based information that decays over a period of seconds unless reactivated by subvocal rehearsal. The precise nature of the code or codes involved remains unclear although there is general agreement that the principal mode of rehearsal is articulatory and can be separated from acoustic input code by articulatory suppression, a procedure whereby participants overtly repeat an irrelevant utterance such as the word "the" (Baddeley, Lewis, & Vallar, 1984; Murray, 1968). Our study stems from the hypothesis that the phonological loop may play an important role in accent learning, together with the suggestion that overt repetition might be an effective way of teaching accent. This is an

approach that is often advocated in second-language teaching but with little evidence of its effectiveness (Munro & Derwing, 2015). Establishing a method of teaching anything as complex as language and pronunciation is likely to require a series of studies culminating in one or more well designed trials with a large participant base extending over a substantial training period and followed by later retesting to establish durability. Such an enterprise demands extensive pilot testing beginning with studies such as our own in which we combine a limited theoretical question with a limited initial step towards potential application.

A challenge when studying accent learning is whether to focus on quantifiable but narrow problems, such as the difficulty with the /r/-/l/ distinction in Japanese speakers (e.g., Goto, 1971) or to opt for the pronunciation of larger segments such as words and sentences. Early efforts to train phonemic contrasts had some success, but transfer to un-trained items was poor (Logan, Lively, & Pisoni, 1991). However, later work showed that intensive training across speakers and tokens could lead to genuine and long-lasting perceptual improvement (Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999). There is also some evidence that perceptual training may enhance pronunciation and accent. For example, Lambacher, Martens, Kakehi, Marasinghe, and Molholt (2005) reported that Japanese participants trained to recognize English vowels over six weeks improved in both their perception and production. A recent meta-analysis of 25 years of research on training the perception of second-language accent by Sakai and Moorman (2017) reported a medium-sized improvement in the perception of accented speech accompanied by a small improvement in production.

One limitation of much of this work from the language teaching viewpoint is that it has tended to focus on the segmental level (focusing on the /r/-/l/ distinction or on vowels, Iverson and Evans, 2007; Lambacher et al., 2005). Derwing and colleagues (Derwing, Munro, & Wiebe, 1998; Derwing & Rossiter, 2003) compared training on segments with training on

whole words, finding that improvement in pronunciation occurred but was limited to individual sounds rather than whole words. However, a later review of the literature by Lee, Jang, and Plonsky (2014) reports that evidence for training effectiveness tends to be clearer when using larger segments of language. Furthermore, real-life cases show that native accents can be learned in a holistic way and without explicit training, as suggested by the literature on second-language pronunciation by immigrants as a function of age of immigration and length of exposure (Piske, Flege, MacKay, & Meador, 2002; Piske, MacKay, & Flege, 2001).

Assessing the gradual change in accent in students from the North of England who had joined a university in the South, Evans and Iverson (2007) found a change in production, but no significant difference in their capacity to perceive words or passages in noise. Relatedly, Huensch and Tremblay (2015), who studied the effect of training on Korean speakers, found improvement in both perception and production, but no one-to-one link between the perception and production of specific phonological features. They interpreted their results as supporting separate but linked speech perception and production systems. Such a distinction is consistent with classic language models in cognitive psychology. For instance, in Morton's (1979) logogen model, lexical representations (logogens) are divided between a perceptual input logogen system and a production output logogen system. Levelt's (1989) model of speech production assumes a shared mental lexicon but separate sub-lexical representations for perception and production tasks, though, again, links between the two systems are evident.

Our own approach to bridging the gap between theory and practice in the field of accent acquisition differs from the earlier work in two ways. First, it starts with a method of language teaching, overt repetition, that is widespread but relatively untested empirically. Second, it approaches the teaching of accent from a theoretical framework concerned with memory and learning rather than linguistics, an approach we regard as complementary to

earlier studies based on phonology. Overt repetition is the requirement to listen to and repeat a spoken stimulus out loud. The paradigm originated in experimental psychology under the term "shadowing" (e.g., Broadbent , 1958; Cherry, 1953), but it is now widely used as a teaching tool, including training programs for simultaneous interpreters (Bovee & Stewart, 2009; Foote & McDonough, 2017; Mori, 2011). Generally, there is a growing market for on-line or app-based methods to teach second languages using variants of the overt-repetition task, although with little sign of concrete empirical evidence for its efficacy—see Hopman and MacDonald (2018) for an exception. The situation appears to be changing, however. Foote and McDonough (2017) trained non-native speakers using a repetition task in an eight-week program. Learners were then tested on tasks requiring repetition and spontaneous speech. Native speakers acting as judges reported improvement on a range of measures, but not on accentedness. However, this was a preliminary study with a number of limitations, as the authors point out. These include the absence of a control group and minimal control over the amount of time participants engaged with the repetition method.

As introduced earlier, our theoretical framework to test the efficacy of overt-repetition is based on the phonological loop component of the working memory model proposed by Baddeley and Hitch (1974, see also Baddeley, 2000, 2012). Subvocal rehearsal within the phonological loop is critical for maintenance of the information in short-term memory, as digit span drops by around two digits when rehearsal is prevented by articulatory suppression (i.e., repeatedly uttering a spoken syllable such as "the"). The substantial residual performance indicates that other forms of storage are also used. One clue as to what these might be comes from studies in which participants are asked to make phonological judgments on visually presented items during articulatory suppression. While suppression does not impair simple homophone judgment (e.g. *wait – weight*; *chaos – cayoss*, Baddeley & Lewis, 1981; Besner, 1987; Besner, Davies, & Daniels, 1981), it disrupts rhyme judgement (e.g.,

weight – hate), which suggests that subvocal articulation is used when the material needs to be actively manipulated by removing the initial consonant. These results were interpreted by Baddeley and Lewis (1981) as suggesting separate systems for storing the sound characteristics of items based on auditory features (a system they refer to as the "inner ear") and on articulatory features (the "inner voice"). This distinction is analogous to the distinction between separate input and output stores discussed earlier (Huensch & Tremblay, 2015; Morton, 1979).

The concept of a multi-component working memory has been applied extensively to native and non-native language learning (Baddeley, Gathercole, & Papagno, 1998), with an early study by Service (1992) showing a link between working memory performance and the acquisition of English as a second language by Finnish school children. The role of working memory capacity in second-language learning is now well established (see Wen, 2012, 2016; Wen, Mota, & McNeill, 2015). In a meta-analysis of studies involving 3707 learners, Linck, Osthus, Koeth, and Bunting (2014) showed clear effects on second-language learning of both the executive and the phonological components of working memory.

In contrast, research on working memory and accent acquisition is sparse, although studies have demonstrated a positive association between phonological working memory and capacity to discriminate unfamiliar phonological distinctions (e.g., Aliaga-Garcia, Mora, & Cerviño-Povedano, 2011; Darcy, Park, & Yang, 2015; Flege & MacKay, 2004). In general, however, there seems to be relatively little theoretically-motivated research on optimal methods of training accent production on a broad level such as would be likely to be encountered in native-language acquisition.

Our study investigates the efficacy of repetition as a means of enhancing the acquisition of a second-language accent, while at the same time addressing the question of whether repetition needs to be overt. From a theoretical viewpoint, this links to the question

of whether effective training depends on the system for overt articulation that underpins rehearsal within the phonological loop, or whether the input buffer concerned with perception is sufficient. We therefore adopted an experimental design that contrasted two types of verbal rehearsal, one that involved overt, out-loud repetition and one in which repetition was covert, subvocal.

We measured improvements in the English accent of Chinese participants enrolled in a British university on a one-year Master's course for teachers of English as a foreign language. We compared performance across three training conditions. In a covert-repetition condition, participants repeated each sentence to themselves, subvocally. In an overt-repetition condition, participants repeated each sentence out loud immediately following sentence presentation. The third condition, which contained no training, simply involved the initial and final assessments. It served as a baseline for any improvement attributable to exposure to ambient English over the period of training. In the covert- and overt-repetition conditions, training consisted of four sessions of approximately 15 minutes each. Each session involved presentation of a spoken sentence accompanied by the sentence in written form. The sessions differed only in whether the sentence had to be repeated overtly or covertly. In all three groups, productions were recorded before and after training, or the equivalent delay for the control condition. Production was elicited in two tasks. One task required immediate sentence repetition, a task assumed to allow contributions of both the temporary short-term representation of that sentence together with any long-term contribution from prior knowledge. The second task required participants to read out written sentences, hence presumably reflecting only prior knowledge.

Pre- and post-training productions were recorded and rated by native English speakers on accent quality, fluency, and intelligibility. The accent rating was meant to capture the intuitive understanding of the concept, as defined by the Oxford English Dictionary (3rd

edition), "a distinctive way of pronouncing a language, especially one associated with a particular country, area or social class." We chose to measure accent and intelligibility separately because these have been shown to partly dissociate (e.g., Munro & Derwing, 1995), even though accent can have implications for intelligibility, especially in the early stage of second-language learning. While fluency is a multi-faceted construct (Kormos & Dénes, 2004), we included it because slow speech rate and minor disfluencies, the hallmarks of low fluency, may possibly dissociate from accent perception or intelligibility. Taken as a whole, we refer to the three dimensions as a global measure of "Accent," capitalized to distinguish it from the narrower meaning of the rated dimension. We use this term in its broadest sense with the three rated measures selected because of their relevance to the practical needs of the learner rather than to their capacity to map precisely onto underlying theoretical constructs.

If accent learning necessitates overt articulation during rehearsal, we expect improvement from pre- to post-training to be restricted to the overt-repetition condition. However, if accent learning can be achieved merely through subvocal rehearsal within the phonological store of the phonological loop, covert repetition should be effective as well. Moreover, whether the effectiveness of training depends on a combination of short-term and long-term memory or just long-term memory will be assessed by the distinction between the repeated and read-out sentences.

Methods

The study included an Accent-learning phase and an Accent-rating phase. Each phase is described in turn.

Learning

Participants. The sentences were spoken by a female native speaker of Standard Southern British English. The initial sample of learners included were 41 female native Mandarin

speakers between 20 and 28 years of age, enrolled in a one-year Master's degree in the Department of Education at the University of York. The degree offered specialization in Teaching English to Speakers of Other Languages (TESOL). The programme aimed to familiarize students with English language teaching methodology, help them develop knowledge of various areas of applied linguistics, evaluate current issues in language learning and teaching in a global context, and consider their application in classroom instruction and course assessment. For reasons described in the Procedure section, three of the 41 learners had to be discarded from the analyses. All the statistics from here onward below are based on the remaining 38 learners.

We chose a single gender as the model speaker for the learners to focus learning on Accent rather than voice differences. Among the 38 learners, the average age of acquisition of English as a second language was 10 years (range: 5-13 years). The average duration of exposure to English as a second language, regardless of the frequency of exposure, was 12 years (range: 7-17 years). All learners had resided in the UK or a country where English is spoken as a native language for less than ten months. Their average International English Language Testing System (IELTS) score was 7.00 (range: 6.5-7.5). Learners also self-assessed their proficiency in English speaking, reading, listening, and writing on a scale from 1 to 10. Averages (and ranges) were, respectively, 5.95 (3-8), 6.89 (4-9), 6.97 (4-9), and 5.79 (4-8). Learners were assigned to one of three training groups: Baseline, covert repetition, or overt repetition. To minimize disparities in English proficiency between the three groups, participants were matched as well as possible on their IELTS and self-assessment scores. Of the original sample of 41 leaners, 13 learners were assigned to the baseline condition, 13 to the covert-repetition condition, and 15 to the overt-repetition condition. The smaller number of learners for the baseline and covert-repetition conditions was the consequence of scheduling constraints and drop-out, as the study required two (baseline) to six (covert and

overt repetition) testing sessions. The initial goal was to have 15 participants in each condition. Learners were paid a small honorarium for their participation.

Materials. These were 280 semantically neutral and phonemically balanced sentences drawn from the IEEE corpus (Rothauser et al., 1969), e.g., *Glue the paper to the dark blue background*. Several sentences were slightly modified to conform to contemporary British English. Once recorded by the native English speaker, sentence average duration was 2236 ms (range: 1470-3341 ms).

Design. The learners in all three conditions (baseline, covert repetition, overt repetition) were recorded in a pre-test session and a post-test session. In the pre-test session, learners repeated out loud 20 spoken sentences and then read out loud another set of 20 written sentences. In the post-test session, learners repeated out loud the same 20 spoken sentences and read out loud the same 20 written sentences as in the pre-test session. They also repeated out loud 20 new spoken sentences and read out loud 20 new written sentences. Learners in the covert- and overt-repetition conditions underwent four training sessions between the pre- and post-test sessions. Learners in the baseline condition did not undergo any training sessions. Each training session included 50 sentences, 40 to repeat and 10 to read out. Having learners repeat some sentences and read out others was meant to increase diversity in training. The larger number of repeated sentences ensured that enough of the training was based on a native model. The only difference between the covert- and overt-repetition conditions was that participants in the covert condition were asked to repeat (or read) the sentences subvocally (i.e., using their "inner voice"), whereas participants in the overt condition were asked to repeat (or read) the sentences out loud. None of the sentences in the training sessions were used in either the pre- or post-test sessions. The sentences that the learners produced in the pre- and post-test sessions (40 pre-test + 80 post-test) were audio-recorded. The sentences

produced during training were not audio-recorded. A summary of the design for the learning part of the experiment is shown in Figure 1.

Procedure. Learners were tested one at a time in a sound-attenuated booth. On each trial of the pre- and post-test sessions, they first saw a written sentence on a computer monitor. What they had to do next depended on whether the sentences had to be repeated or read out. In the repeat condition, 3 s after the written sentence was displayed, that sentence was spoken by the female native English speaker, with the written sentence remaining onscreen. The sentence was played from loudspeakers on both sides of the monitor at approximately 68 dB SPL. The instructions were to listen and imitate the speaker as well as possible. Participants could start producing the sentence as soon as the spoken sentence was over. They then pressed a key to move on to the next sentence. If they did not press the key, the next sentence played automatically after 20 s. In the read-out condition, everything was the same, except that the spoken sentence was not played. The instructions were to read the sentence out loud. The 20 repeat sentences always preceded the 20 read-out sentences. In the post-test session, new and old sentences were randomly intermixed.

The procedural details of the trials in the covert and overt training sessions were the same as those in the pre-/post-tests, except that in the covert training sessions, learners were asked to repeat (or read) each sentence to themselves, only once, rather than repeating (reading) them out loud.

Pre-and post-tests were separated by a minimum of two weeks in all three learning groups. Training sessions in the overt- and covert-repetition groups were separated by at least two days. Each session (pre-test, training, post-test) lasted between 20 and 30 minutes.

The sentences of three of the 41 learners had to be discarded (one in the covert-repetition condition and two in the overt-repetition condition), because the recordings in both the pre- and post-test sessions contained too many extreme disfluencies, interruptions,

repetitions, and mistakes to be objectively rated. Thus, the analyses reported in the Results section were performed on the sentences produced by the remaining 38 learners.

Rating

Participants. Twenty-eight native British English speakers (21 female; average age 21, range: 18-31) were recruited as raters from the undergraduate and postgraduate populations in the University of York psychology department. None of them were familiar with Mandarin Chinese. They were paid a small honorarium for their participation.

Materials. The sentences generated by the learners during the pre- and post-test sessions constituted the materials to rate. Any background noise was removed using the Audacity software noise reduction function, with 3 dB noise reduction, signal-to-noise sensitivity set at 5 (0-24), and 4 bands of frequency smoothing. The average intensity of the sentences was normalised to 68 dB SPL across sentences.

Design. Each rater rated the sentences of three learners (or, for some raters, only two), one in each of the three conditions (baseline, covert repetition, overt repetition). Raters were paired so that they both rated the sentences of the same triplet (or pair) of learners. Thus, each sentence was rated by two raters. They rated each sentence on the quality of its accent, its fluency, and its intelligibility.

Procedure. Each rating trial started with a written presentation of the sentence to be rated. This insured that the ratings of accent and fluency was not overly influenced by intelligibility. Once they had read it, the raters pressed a key, which made the written sentence disappear and the corresponding spoken sentence start. The sentence was only played once. Then, a screen appeared containing three 5-point scales, one for accent, one for fluency, and one for intelligibility. Accent was defined as how native-British-English-sounding the sentence was. Fluency was defined as perceived effortlessness based on rate and absence of disfluencies

(hesitations, restarts). Intelligibility was defined as how easy it was to match the spoken sentence with the written sentence. Note that this procedure meant that intelligibility was assessed subjectively rather than through objective transcription. This compromise ensured that participants could distinguish between the three measures while using the same scale for all of them. Rating of 1 was on the left for poor and 5 was on the right for good. Raters clicked on a value for each of the three scales, then pressed another key to move on to the next trial. Most raters rated 360 sentences (120 x 3 learners). A few raters only rated 240 sentences (120 x 2 learners), due to the slight imbalance across the baseline, covert-repetition, and overt-repetition conditions. To encourage the raters to use the scale homogenously, the experiment started with a few examples of sentences exhibiting extreme quality (poor or good) on one scale but not on the others. Raters were told that values were expected to be around 1 (or 5) for those examples.

Results

Data analysis and inter-rater reliability

Rating data, averaged across two raters for each sentence, were analysed following a mixed-effect model-comparison approach using R (version 3.3.1) with glmer (package lme4). Statistical considerations are described in the Appendix. Average ratings are shown in Table 1.

Inter-rater reliability was calculated for each of the 14 pairs of raters who rated the same set of sentences (see Rating, Design section). Intra-class correlation coefficients (ICC), based on two-way mixed models and a consistency definition, was low to moderate, as per Koo and Lee (2016), with ICC ranging from .285 to .695 across the 14 pairs. The range of the lower bound of the 95% CI was between .076 and .583, and that of the upper bound was between .447 and .778. Despite the relatively low reliability score, all 14 coefficients were

significant at $p < .01$. Furthermore, Spearman's ρ (which was chosen over Pearson's p to account for the ordinal nature of the rating scales) was highly significant for 11 pairs ($p < .001$); $p < .01$ for the other three pairs.

Effect of training on ratings (pre-test vs. post-test)

In a first set of analyses, we only considered ratings for the sentences heard both in the pre- and post-tests, leaving aside the new sentences of the post-test, as these will be used to assess transfer in a second set of analyses. The independent variables were Time (pre-test, post-test), Modality (repeat, read-out), Group (baseline, covert, overt), and Rating Dimensions (accent, fluency, intelligibility). Of the four possible main effects, only Rating Dimension was significant, $\beta 1 = .931$, $SE1 = .019$, $\beta 2 = .055$, $SE2 = .016$, $\chi^2(2) = 2256.40$, $p < .001$ (all other main effects, ps > .15). The accent measure ($M = 3.05$, $SD = 1.06$) was rated lower than both fluency ($M = 3.72$, $SD = 1.04$), $\beta = .671$, $SE = .016$, $\chi^2(1) = 1569.30$, $p < .001$, and intelligibility ($M = 3.78$, $SD = 1.08$), $\beta = .726$, $SE = .016$, $\chi^2(1) = 1823.80$, $p < .001$. Although close to each other numerically, fluency was rated lower than Intelligibility, $\beta = .055$, $SE = .016$, $\chi^2(1) = 11.386$, $p < .001$. While ratings are difficult to compare across measures, the relatively high scores for intelligibility probably stem from the advantage of presenting the written sentence before each spoken sentence. The comparatively low score for accent is likely to reflect the obvious non-nativeness of all speakers.

Because the effect of training was the focus of this study, only interactions involving Time were investigated and the significant tests reported. Since none of the interactions involving Rating Dimension were significant, all analyses were performed on the data aggregated across the three rating dimensions, as plotted in Figure 2. This aggregated measure should be seen as an index of our global Accent construct. A significant interaction between Time and Group, $\beta 1 = -.025$, $SE1 = .178$, $\beta 2 = .162$, $SE1 = .176$, $\chi^2(2) = 14.70$, $p < .001$, revealed a trend toward increased ratings from pre- to post-test for the overt-repetition

condition, $\beta = .045$, $SE = .026$, $\chi^2(1) = 3.05$, $p = .08$, but not for either the baseline or the covert-repetition conditions (ps > .45). However, this pattern was affected by Modality, as indicated by a 3-way interaction between Time, Group, and Modality, $\beta 1 = -.172$, $SE1 = .083$, $\beta 2 = .230$, $SE2 = .081$, $\chi^2(2) = 8.66$, $p = .01$. As visible in Figure 2 (top quadrants), Time interacted with Group for the read-out sentences, $\beta 1 = -.136$, $SE1 = .058$, $\beta 2 = .262$, $SE2 = .058$, $\chi^2(1) = 20.43$, $p < .001$, but not for the repeat sentences, $\chi^2(2) = 1.14$, $p = .56$. For the read-out sentences, there was a pre-to-post improvement in the overt-repetition condition, $\beta = .080$, $SE = .026$, $\chi^2(1) = 8.13$, $p = .004$, but not in the baseline or covert-repetition conditions (ps > .40). In summary, training improved perceived nativeness (accent, intelligibility, and fluency) only if training involved repeating sentences out loud and if production was measured on read-out sentences.

However, it should be noted that, even though there was no main effect of Group, the overt-repetition group started from numerically higher aggregated ratings than the other two groups. Thus, despite matching learners on their English proficiency across the three groups, the group that showed improvement through training was also the group that started with comparatively higher ratings. To insure that this initial advantage was not somehow responsible for the learning effect, we re-ran the above analyses on a sub-group of learners matched on their pre-test aggregated ratings (Figure 2, bottom quadrants). Matching was easily achieved by removing the two learners showing the highest pre-test aggregated ratings in the overt-repetition group (averaged across modalities and rating dimensions). Average pre-test aggregated ratings before matching were: Baseline = 3.46 (N = 13), Covert repetition = 3.48 (N = 12), Overt repetition = 3.59 (N = 13). Average pre-test aggregated ratings after matching were: Baseline = 3.46 (N = 13), Covert repetition = 3.48 (N = 12), Overt repetition = 3.46 (N = 11).

The key statistical analyses showed virtually no change in significance: Time x Group, $\beta 1$ = -.060, *SE1* = .043, $\beta 2$ = .163, *SE2* = .044, $\chi^2(2) = 13.87$, $p < .001$; Time x Group x Modality, $\beta 1$ = -.178, *SE1* = .085, $\beta 2$ = .240, *SE2* = .086, $\chi^2(2) = 8.28$, $p = .02$. The Time x Group interaction was significant in the read-out modality, $\beta 1$ = -.138, *SE1* = .060, $\beta 2$ = .269, *SE2* = .061, $\chi^2(2) = 19.05$, $p < .001$, but not in the repeat modality, $\chi^2(2) = 1.37$, $p = .50$. In the read-out modality, pre- to post-test improvement was significant in the overt-repetition condition, $\beta = .080$, *SE* = .026, $\chi^2(1) = 7.10$, $p = .008$, but not in the baseline or covert-repetition conditions (*p*s > .45). Overall, these analyses confirm the patterns described earlier.

Taken together, these results show that two conditions had to be met for ratings to improve from pre-test to post-test. First, learners had to repeat sentences out loud during the training sessions. Repeating sentences subvocally was not sufficient. Second, the pre- to post-test improvement was only measureable when the recordings were read-out sentences. Sentences repeated from a spoken model did not reveal an improvement. Finally, the learning pattern was broadly comparable across the three dimensions tested, and these were highly correlated (accent-fluency, Spearman's $\rho = .458$, $p < .001$; accent-intelligibility, $\rho = .619$, $p < .001$; fluency-intelligibility, $\rho = .538$, $p < .001$), suggesting that all facets of our Accent construct responded to the learning treatment similarly.

Effect of training on transfer to new materials (post-test Old vs. post-test New)

The second set of analyses aimed to evaluate whether learning in the overt-repetition condition transferred to new sentences. To do so, we compared ratings of the old and new sentences in the post-test session. Analyses included the factors Old-New (old; new), Modality (repeat, read-out), and Group (baseline, covert repetition, overt repetition). A significant Old-New main effect showed that new sentences were rated lower than old sentences, β = -.167, *SE* = .049, $\chi^2(1) = 11.10$, $p < .001$ (matched data set: $\beta = .160$, *SE* = .049, $\chi^2(1) = 10.11$, $p = .001$). The other main effects or interactions did not reach significance in

either the full data set or the matched data set analyses. The lack of interaction with the Old-New factors suggests that the pattern of ratings transferred from the old to the new post-test sentences more or less unchanged.

However, this conclusion should be tempered in view of analyses comparing ratings of the new sentences in the post-test session with ratings of the old sentences in the pre-test session. If it is true that learning transferred to the new sentences, the contrast between the new sentences and the pre-test sentences should be similar to the contrast between the (old) post-test sentences and the pre-test sentences. This was only partly the case, however. The new sentences were rated lower than the pre-test sentences, $\beta = .074$, $SE = .026$, $\chi^2(1) = 8.03$, $p = .005$, but the key Time x Group x Modality interaction did not reach significance, $\chi^2(2) = 4.27$, $p = .12$. Analyses on the matched data showed the same patterns, respectively, $\beta = .072$, $SE = .026$, $\chi^2(1) = 7.65$, $p = .006$, and $\chi^2(2) = 2.15$, $p = .34$.

These last analyses, as well as a visual inspection of the data, show mixed evidence that the benefit of learning through overt repetition transferred to new sentences: On the one hand, the cross-group pattern of ratings was comparable in the new sentences and the old sentences, which suggests generalization. On the other hand, the cross-group pattern of ratings of the new sentences was not statistically different from the pattern in the pre-test sentences, which fails to provide confirmation for generalization. This ambiguous pattern could be partly the result of unexpectedly low ratings for the new sentences relative to the pre-test sentences. There are two potential reasons for such low ratings. First, they could be the manifestation of a contrast effect, whereby the unfamiliarity of the new sentences amid familiar ones would have led to less precise pronunciation, perhaps through decreased engagement with the task on those sentences. Alternatively, the new sentences could have been phonologically more difficult to pronounce than the old ones. This hypothesis cannot be verified, because sentence identify was not counterbalanced between the old and new sets.

This possibility seems unlikely, however, because the IEEE sentences are phonologically balanced and were randomly assigned to the old or new sets. In sum, while the overt-repetition technique led to improvement in pronunciation (in the visual read-out condition at least), evidence that the improvement transferred to new materials was mixed.

Discussion

The study of pronunciation in the second-language teaching literature is sparse (Munro & Derwing, 2015). Despite being widely used in the classroom, there is only limited empirical evidence that overt repetition, the requirement to "listen and repeat," sometimes referred to as "shadowing," is an effective teaching tool for native-accent learning (Foote & McDonough, 2017). Our study had two main aims: (1) To test empirically whether sentence repetition would enhance the acquisition of a second-language accent and (2) To provide a preliminary account of our findings within the multi-component working memory model. Specifically, we asked whether out-loud (overt) repetition or subvocal (covert) repetition would improve the English accent of native Chinese speakers over four brief training sessions. As described in the Introduction, we refer to the three rated dimensions (accent quality, fluency, and intelligibility), considered collectively, as "Accent," a general measure of effective native-sounding speech.

Within the limitations of a relatively small-scale study, the results showed that overt repetition leads to moderate improvements in native-sounding speech production whereas covert repetition does not. The improvement was visible not only in the ratings of the narrowly-defined and intuitive notion of accent but also in the ratings of fluency and intelligibility. The effect size was modest, however. Cohen's *d*, measured as the proportion between the average improvement from pre- to post-ratings for the read-out overt-repetition group (.159) divided by the pooled standard deviation of that same group (.629) was .252,

when the standard deviation was measured across raters. When the standard deviation was measured across speakers, it was .159/.446 = .356. The latter figure is probably a more realistic approximation of the gain that can be expected since it takes into account individual differences at pre-test. However, even the larger figure is smaller than the effect size of .54 reported by Sakai and Moorman's (2017) meta-analysis of accent-learning studies and smaller than the threshold of .60 for pre/post-test studies recommended by Plonsky and Oswald (2014). From a practical viewpoint, this level of improvement does not warrant a radical change in how non-native accent is taught. Nevertheless, it confirms the intuition that the "listen and repeat" method is preferable over sub-vocal repetition even after a brief training regimen, a total of about one hour to modify a lifetime's experience of Mandarin. This pattern suggests that our method might profitably be scaled up to provide a more robust test of our hypothesis using substantially more participants and a more realistic amount of training.

Within the working memory framework sketched in the Introduction, our results suggest that overt articulation through the "inner voice" is necessary for improvement to occur. This is far from an obvious outcome. It is possible, for example, that the level of accuracy of overt repetition could be so poor that it merely reinforces errors. This was clearly not the case—although the likelihood that it could be a limiting factor with beginners might have to be explored in the future. A second possibility is that improvement may depend on the capacity of the speaker to detect discrepancies between the target and his/her own utterances, something that cannot always be assumed, as for example indicated by the problems Japanese speakers have in perceiving the distinction between /r/ and /l/ (e.g., Goto, 1971). The fact that improvement occurred despite these two potential problems lends

support to the practice of encouraging learners to overtly repeat spoken utterances as suggested by some language teachers.

We were surprised that the improvement was limited to the read-out sentences, in contrast to the repeated sentences. This was unexpected since repetition priming has been demonstrated across a wide range of modalities (Tulving & Schacter, 1990). However, it is possible that performance in this condition was dominated by the short-term acoustic representation of the sentences, bypassing the long-term articulatory system. This may have led the speakers in the repeated condition to produce the sentences through imitation (vocally or subvocally), rather than by constructing an output based on long-term phonological or articulatory representations, as would have been necessary in the read-out condition. This possibility is compatible with the evidence for a direct and automatic link between speech perception and production (e.g., Pulvermüller et al., 2006; Wilson & Knoblich, 2005), even though production might lag behind perception during development (e.g., Nagle, 2018). Reliance on such a direct link in the pre- and post-test recordings does not necessarily negate the benefit of training on long-term representations, but it might have masked any training-induced improvement when the sentences were recorded using the repeat condition.

Our results raise interesting questions for the current version of the multi-component working model in general and for the phonological loop in particular. While the need to separate the acoustic and articulatory components has been accepted for many years, that separation is less clear at the level of rehearsal., where articulatory suppression is used widely as a means of controlling rehearsal. Thus far, evidence has been limited to the separate roles of the two codes in reading (Baddeley & Lewis, 1981; Besner, 1987; Besner et al., 1981). However, interest has recently been revived by Norris, Butterfield, Hall, and Page (2018) in a study that combines articulatory suppression with an ongoing task requiring short-term retention of sequences of words about which judgements of homophony or rhyme are

required. As predicted by the reading results, the capacity to make correct homophony judgements persisted.

Within the current version of the phonological loop (Baddeley & Hitch, 2018), two types of rehearsal are distinguished. One involves articulatory maintenance, whereby familiar items such as digits and words can be maintained by sub-vocalisation. The other involves the more demanding process referred to as refreshing, whereby attention continues to be focused on a limited number of items within the episodic buffer, a component that maintains bound representations and makes them available to conscious awareness. New verbal material can be maintained using this process, a process that is also available to other modalities, but one that is more limited and attentionally demanding than sub-vocalisation. We assume that this latter system, unlike overt articulation, does not require the setting up and running of speech output programs, and hence, does not lead to the long-term articulatory learning needed to allow accent modification.

A final concern from our data is with the somewhat equivocal evidence of transfer of improvement to completely new sentences. We run here into methodological considerations. On the one hand, our demonstration of improvement from overt repetition was found under carefully controlled test conditions whereby exactly the same sentences were read by the same learners and evaluated by the same judges, hence optimising sensitivity of the test. On the other hand, in the case of transfer to novel sentences, the sustained advantage of overt repetition suggests generalization to new materials, but this is essentially a weak test since it relies on the absence of an interaction. The advantage that would have been expected relative to the pre-test ratings did not reach significance. Incidentally, the data showed that the new sentences were significantly worse in pronunciation accuracy. This may be a chance effect since all sentences were selected on the same basis (or a contrast effect, as suggested in the Results section), but it is clearly a potential source of noise, possibly leading to the absence of

a clear-cut effect of generalization. From a practical viewpoint, this is an issue that needs to be resolved by means of a more powerful study with considerably larger groups and substantially more training than the four brief sessions we used.

What, therefore, are the implications of our study for second language teaching? We should begin by acknowledging the limitations of the study, involving a small number of learners, a modest degree of practice, the somewhat artificial training method, limited evidence for generality across a wider language sample, and the absence of a measure of delayed post-test performance to establish the durability of learning. Even taken at face value, it is important to remember that training studies require extensive replication before becoming a firm part of the standard curriculum, and that an adequate trial would require a much larger sample of participants and considerably longer training than we ourselves were able to achieve. With the development of web-based methods of delivery, however, this should in future become both practicable and affordable. There is of course a major limitation at this point, which concerns the provision of feedback during training. Ideally, it should be possible to accumulate data and feedback information on a learner's performance over successive learning sessions. Given a reliance on ratings by a panel of native speakers, this is not currently practicable. However, our own enquiries suggest that a number of groups are working on the capacity to evaluate accent automatically from both forensic (e.g., Brown, 2014; Chen et al., 2001, 2014) and educational viewpoints (e.g., Kim, 2006; van Dalen et al., 2015). In the meantime, the field would certainly benefit from constrained studies such as our own with a more modest aim of replication comparing instructional techniques, preferably within a theoretical context that combines contributions from both cognitive psychology and linguistics.

In conclusion, our study suggests that the perceived accent quality, fluency, and intelligibility of sentence production in a non-native language (our broad construct of

"Accent") can be improved over a relatively brief period of training and that this reflects the articulatory output stage of speech rather than its internal representation. It lends some encouragement to the practice of using overt repetition (shadowing) to improve native accent learning, although practical implications should await further research using longer training and larger samples of second language learners.

Authors' Note

This study was funded in part by an internal research grant from the University of York to S. Mattys and A. Baddeley. We thank Danijela Trenkic for recruiting the Chinese participants, sharing their language proficiency data, and contributing to discussions. We thank Imogen Birch, Lauren Larkin, Abbie Llewelyn, Alex Temple, Amandeep Sandhu, Hanna Willis, and Josh Spowage for testing the learners and raters, and David Tolman for directing our attention to the potential role of working memory in accent acquisition. The data that support the findings of this study are available on request from the corresponding author.

References

Aliaga-Garcia, C., Mora, J.C., & Cerviño-Povedano, E. (2011). Phonological short-term memory and L2 speech learning in adulthood. *Poznań Studies in Contemporary Linguistics, 47*, 1-14.

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

Baddeley, A. (2012). Working memory: Theories, models and controversies. *Annual Review of Psychology, 63*, 1-29.

Baddeley, A. D., Gathercole, S. E., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105*, 158-173.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. A. Bower (Ed.), *Recent Advances in Learning and Motivation* (Vol. 8, pp. 47-89). New York: Academic Press.

Baddeley, A. D., & Hitch, G. J. (2018). The phonological loop as a buffer store: An update. *Cortex*. doi:10.1016/j.cortex.2018.05.015

Baddeley, A. D., & Lewis, V. J. (1981). Inner active processes in reading: The inner voice, the inner ear and the inner eye. In A. M. Lesgold & C. A. Perfetti (Eds.), *Interactive Processes in Reading* (pp. 107-129). Hillsdale, N.J.: Lawrence Erlbaum.

Baddeley, A., Lewis, V., & Vallar, G. (1984). Exploring the articulatory loop. *The Quarterly Journal of Experimental Psychology Section A*, *36*, 233-252.

Besner, D. (1987). Phonology, lexical access in reading, and articulatory suppression: A critical review. *Quarterly Journal of Experiment Psychology, 39A*, 467-478.

Besner, D., Davies, J., & Daniels, S. (1981). Reading for meaning: The effects of concurrent articulation. *Quarterly Journal of Experimental Psychology, 33A*, 415-438.

Bovee, N., & Stewart, J. (2009). The utility of shadowing. In *JALT 2008 Conference Proceedings. Tokyo: JALT*.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. I. (1999). Training Japanese listeners to identify English/r/and/l: Long-term retention of learning in perception and production. *Attention, Perception, & Psychophysics*, *61*, 977-985.

Broadbent, D. (1958). *Perception and Communication.* London: Pergamon Press.

Brown, G. (2014). *Y-ACCDIST: An Automatic Accent Recognition System for Forensic Applications.* MA by research thesis, University of York.

Buss, L. (2016). Beliefs and practices of Brazilian EFL teachers regarding pronunciation. *Language Teaching Research*, *20*, 619-637.

Chen, T., Huang, C., Chang, E., & Wang, J. (2001). Automatic accent identification using Gaussian mixture models. In *Automatic Speech Recognition and Understanding, 2001. IEEE Workshop on ACRU* (pp. 343-346).

Chen, N. F., Tam, S. W., Shen, W., & Campbell, J. P. (2014). Characterizing phonetic transformations and acoustic differences across English dialects. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, *22*(1), 110-124.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, *25*, 975-979.

Darcy, I., Park H. & Yang, C-L (2015). Individual differences in L2 acquisition of English phonology: The relation between cognitive abilities and phonological processing. *Learning and Individual Differences, 40,* 63-72.

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language learning*, *48*, 393-410.

Derwing, T. M., & Rossiter, M. J. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. *Applied Language Learning*, *13*, 1-17.

Evans, B. G., & Iverson, P. (2007). Plasticity in vowel perception and production: A study of accent change in young adults. *The Journal of the Acoustical Society of America*, *121*, 3814-3826.

Flege, J.E. & McKay, I.R.A. (2004). Perceiving vowels in a second language. *Studies in Second Language Acquisition, 26*, 1-34.

Foote, J. A., Holtby, A. K., & Derwing, T. M. (2011). Survey of the teaching of pronunciation in adult ESL programs in Canada, 2010. *TESL Canada Journal*, *29*, 1-22.

Foote, J. A., & McDonough, K. (2017). Using shadowing with mobile technology to improve L2 pronunciation. *Journal of Second Language Pronunciation*, *3*, 34-56.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "L" and "R". *Neuropsychologia*, *9*, 317-323.

Henderson, A., Frost, D., Tergujeff, E., Kautzsch, A., Murphy, D., Kirkova-Naskova, A., Waniek-Klimczak, E., Levey, D., Cunningham, U., & Curnick, L. (2012). The English pronunciation teaching in Europe survey: Selected results. *Research in Language*, *10*, 5-27.

Huensch, A., & Tremblay, A. (2015). Effects of perceptual phonetic training on the perception and production of second language syllable structure. *Journal of Phonetics*, *52*, 105-120.

Hopman, E. W., & MacDonald, M. C. (2018). Production practice during language learning improves comprehension. *Psychological science*, *29*, 961-971.

Iverson, P., & Evans, B. G. (2007). Learning English vowels with different first-language vowel systems: Perception of formant targets, formant movement, and duration. *The Journal of the Acoustical Society of America*, *122*(5), 2842-2854.

Kim, I. S. (2006). Automatic speech recognition: Reliability and pedagogical implications for teaching pronunciation. *Educational Technology & Society*, *9*, 322-334.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*, 155-163.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*, 145-164.

Lambacher, S., Martens, W., Kakehi, K., Marasinghe, C., & Molholt, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics*, *26*, 227-247.

Lee, J., Jang, J., & Plonsky, L. (2014). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, *36*, 345-366.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT press.

Levis, J. M. (2015). The journal of second language learning pronunciation: An essential step towards a disciplinary identity. *Journal of Second Language Learning: Pronunication*, *1*, 1-10.

Linck, J.A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review*, *21*, 861-883.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify */r/* and */l/*: A first report. *Journal of the Acoustical Society of America*, *89*, 874-886.

Mori, Y. (2011). Shadowing with oral reading: Effects of combined training on the improvement of Japanese EFL learners' prosody. *Language Education & Technology*, *48*, 1-22.

Morton, J. (1979). Facilitation in word recognition: Experiments causing change in the logogen model. In *Processing of visible language* (pp. 259-268). Springer.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*, 73-97.

Munro, M. J., & Derwing, T. M. (2015). A prospectus for pronunciation research in the 21st century: A point of view. *Journal of Second Language Pronunciation*, *1*, 11-42.

Murray, D. J. (1968). Articulation and acoustic confusability in short-term memory. Journal of Experimental Psychology, 78, 679-684

Nagle, C. (2018). Examining the temporal structure of the perception-production link in SLA: A longitudinal study. *Language Learning*, *68*, 234-270.

Norris, D., Butterfield, S., Hall, J., & Page, M. P. (2018). Phonological recoding under articulatory suppression. *Memory & Cognition*, *46*, 173-180.

Piske, T, Flege, J.E. McKay, I.R.A. & Meador, D. (2002). The production of English vowels by fluent early and late Italian-English bilinguals. *Phonetica, 59*, 49-71.

Piske, T., MacKay, I.R.A., Flege, J.E. (2001). Factors affecting degree of foreign accent in an L2. A review. *Journal of Phonetics, 29,* 191-215.

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

Pulvermüller, F., Huss, M., Kherif, F., del Prado Martin, F. M., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences*, *103*, 7865-7870.

Rothauser, E., Chapman, W., Guttman, N., Silbiger, H., Hecker, M., Urbanek, G., Nordby, K., & Weinstock, M. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics, 17*, 225-246.

Sakai, M., & Moorman, C. (2017). Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics*, 1-38.

Service, E. (1992). Phonology, working memory and foreign-language learning. *Quarterly Journal of Experimental Psychology, 45A*, 21-50.

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science, 247*, 302-306.

van Dalen, R. C., Knill, K. M., & Gales, M. J. (2015). Automatically grading learners' English using a Gaussian process. *SLaTE 2015: Workshop on Speech and Language Technology in Education*, 7-12.

Wen, Z., Mota, M. B., & McNeill, A. (Eds.). (2015). *Working memory in second language acquisition and processing* (Vol. 87). Multilingual Matters.

Wen, Z. (2012). Working memory and second language learning. *International Journal of Applied Linguistics*, *22*, 1-22.

Wen, Z. E. (2016). *Working memory and second language learning: Towards an integrated approach* (Vol. 100). Bristol, UK: Multilingual matters.

Wilson, M. & Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychological Bulletin*, *131*, 460-473.

Appendix

Statistical Analyses

The two-level categorical factors were coded as follows: Time (pre-test: -.5; post-test = .5), Modality (repeat: -.5; read-out: .5). The three-level categorical factors (Group and Rating Dimension) were each coded as two dummy variables with contrasted level-coding structures. Group followed an effect coding structure since it included a baseline level: Group1 (baseline = -.5; covert = .5; overt = 0) and Group2 (baseline = -.5; covert = 0; overt = -.5). Rating Dimension followed a Helmert coding structure since it did not have an obvious baseline level or an ordered structure: Rating Dimension1 (accent = -.5; fluency = .25; intelligibility = .25) and Rating Dimension2 (accent = 0; fluency = -.5; intelligibility = .5). For each test, we provide the β and *SE* values originating from the more complex model of the comparison. For tests involving three-level categorical factors, the β and *SE* values of each of the two dummy variables are reported as $\beta 1$ and *SE1* and $\beta 2$ and *SE2*, respectively.

Following a model-comparison approach, factors were added incrementally to a base model, and improved fit was assessed using the likelihood ratio test. Interaction terms were assessed by comparing models containing all main effects and interaction terms with models in which the critical interaction term was removed. For all models, we included the most complete random structure that led to successful convergence. This structure consisted of by-learner, by-rater, and by-sentence intercepts, as well as by-rater and by-item random slopes for Time. Interaction terms were not included because they prevented the models from converging. The effect of each fixed factor on rating was assessed against a model that only included the random structure.

Table 1: Average ratings (and standard deviations) as a function of Time (pre-test, post-test), Modality (repeat, read-out), Group (baseline, covert repetition, overt repetition), and Rating Dimensions (accent, fluency, intelligibility). Rating scales are from 1 (poor) to 5 (good).

	Repeat			Read-out		
	Pre	Post	Post (new)	Pre	Post	Post (new)
Baseline						
Accent	3.09 (0.97)	3.04 (0.97)	2.85 (0.98)	2.96 (0.98)	2.92 (0.99)	2.85 (0.97)
Fluency	3.68 (0.90)	3.70 (0.93)	3.60 (0.92)	3.50 (1.06)	3.51 (1.02)	3.29 (1.09)
Intelligibility	3.80 (0.97)	3.79 (0.98)	3.55 (1.05)	3.72 (1.06)	3.68 (1.05)	3.43 (1.08)
Covert Repetition						
Accent	3.06 (1.05)	3.02 (1.17)	3.03 (1.31)	2.96 (1.06)	2.91 (1.10)	2.82 (1.08)
Fluency	3.75 (1.09)	3.79 (1.06)	3.69 (1.07)	3.66 (1.14)	3.69 (1.13)	3.48 (1.16)
Intelligibility	3.72 (1.17)	3.70 (1.11)	3.55 (1.18)	3.71 (1.21)	3.65 (1.19)	3.55 (1.17)
Overt Repetition						
Accent	3.18 (1.13)	3.16 (1.09)	3.05 (1.04)	3.09 (1.07)	3.21 (1.05)	2.98 (1.11)
Fluency	3.85 (1.02)	3.91 (0.99)	3.76 (1.03)	3.70 (1.06)	3.92 (1.00)	3.68 (1.09)
Intelligibility	3.88 (1.08)	3.90 (1.01)	3.71 (1.06)	3.82 (1.05)	3.95 (1.00)	3.73 (1.10)

Figure 1: Number of sentences per sessions (pre-test, training, post-test) as a function of modality (repeat, read-out) and learning groups (baseline, covert repetition, overt repetition).

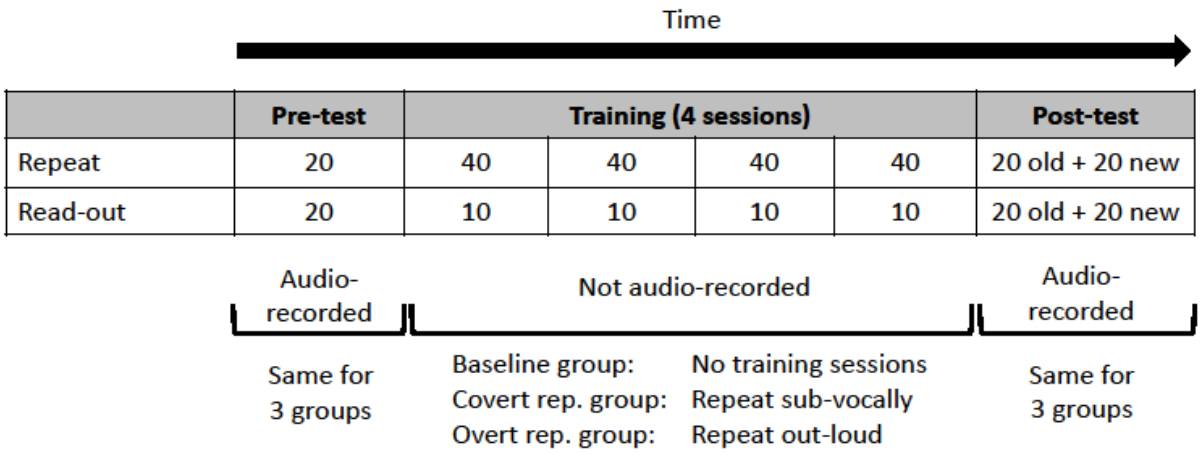

Figure 2: Average ratings of sentences (aggregated across accent, fluency, and intelligibility) as a function of Time (pre-test, post-test), Modality (repeat, read-out), and Group (baseline, covert repetition, overt repetition). Since the Post (New) condition includes sentences unused in the pre-test and post-test, it is shown as disconnected from those conditions. Top figures include all data. Bottom figures show data matched on average ratings in the pre-test. Error bars are standard errors of the mean.