Shaping Phonetic Performance in Second Language Learners

by

Hiu-Nam Jaime Leung

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfilment of the requirements of the degree of

MASTER OF ARTS

Department of Psychology
University of Manitoba
Winnipeg

ASTRACT

This study aimed to evaluate the efficacy of a software-administered shaping procedure in guiding English monolinguals to acquire accurate Mandarin pronunciation. A single-subject reversal ABAB design was used to evaluate treatment effects. A purposely-developed algorithm generated an accuracy score defined as the similarity between a participant's utterance and the target pronunciation. The shaping procedure provided performance-dependent reinforcement, while the control condition provided performance-independent reinforcement at a density yoked to the shaping procedure. A no-feedback condition assessed spontaneous language learning ability prior to treatment. Data were evaluated via visual analysis and complemented with effect size analyses and repeated-measures ANOVAs. There were no overall treatment effects. However, three individuals demonstrated a statistically significant difference between treatment and control. A follow-up study compared shaping to no feedback using a simplified procedure and simpler stimuli. A multiple-baseline design was used. The results showed no treatment effects. Possible contributing factors and directions for future research are discussed.

Keywords: second language learning, computer-assisted language learning, shaping.

ACKNOWLEDGMENTS

I would like to thank my co-advisors, Dr. Javier Virues-Ortega and Dr. Lorna Jakobson, for their indispensable guidance for the last three years. I would also like to thank my committee members, Drs. Joseph Pear and Zahra Moussavi, for their valuable feedback. In addition, I would like to thank Zi Ye for creating and troubleshooting the comparison algorithm and software used in this thesis.

Finally, I want to acknowledge the financial support I have received from the University of Manitoba as well as from Canadian Institutes of Health Research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Correcting an accent is an important component of language learning. It not only affects the intelligibility of language (Munro & Derwing, 1999), but also other's perceptions of the speaker (Bresnahan, Ohashi, Nebashi, Liu, & Morinaga Shearman, 2002; Lindemann, 2003). Statements spoken in non-native accents are perceived as less credible than those spoken in the native accent, even when the non-native speaker is merely passing on information (Lev-Ari & Keysar, 2010). Further, compared with native-sounding speakers reading the same passage, people with accents are consistently perceived as less intelligent, less competent, and less friendly than those who sound native (Bresnahan et al., 2002; Lindemann, 2003). These effects are especially evident when the accent renders the speech unintelligible (Bresnahan et al., 2002). The development of effective methods to establish more native-like (i.e., more accurate) pronunciation would be highly valuable to second language learners, and could serve as a model for approaches used to improve pronunciation in clinical populations with limited phonetic performance (Munro & Derwing, 1999; Schmid & Yeni-Komshian, 1999).

The role of accents in communication is highly relevant today with the rising trend in global mobility. More people are travelling to foreign countries for work and education, which requires them to learn foreign languages quickly and accurately. In Canada, the number of foreign students increased from 178,000 to 265,000 between 2008 and 2012 (Government of Canada, 2012). In addition, the number of foreign workers in Canada has increased by 25% within the last four years (Government of Canada, 2012). Similar trends have been reported in other countries (e.g., Independent Administrative Institution Japan Student Services Organization, 2011).

Although accurate pronunciation is clearly important, current language learning materials and programs do not focus on pronunciation training but rather on vocabulary and grammar. If pronunciation training is conducted, feedback is often limited to the presentation of a simple "correct" or "incorrect" response (Burleson, 2007), on occasions followed by an opportunity for self-correction. Often, beginners struggle to chance upon the correct pronunciation. Occasionally, this correct/incorrect feedback is supplemented with visual information such as waveforms (e.g., Rosetta Stone). For example, Hirata (2010) used visual depictions of how native speakers perceive speech (prosody graphs) to improve Japanese language pronunciation in English speakers. Participants were able to increase their performance, but only after receiving frequent assistance and detailed instructions on how to interpret the prosody graphs. Although this study found that it is possible to use visual feedback to train pronunciation, this method still requires one-to-one training to detect discrepancies between the subject's own pronunciation and the model.

Detecting differences between accurate and inaccurate second language (L2) pronunciation relies on the learner's ability to: (a) discriminate the mismatch between what they said and the native pronunciation, and (b) correct their own performance in order to minimize this discrepancy. This is even more challenging in programs where the learner's pronunciations are not repeated to them (e.g., CD programs, free websites), because they would need to accurately recall their own pronunciation to make the comparison. Moreover, research shows that a person's native language (L1) interferes with the acquisition of correct L2 pronunciation (Aoyama, Flege, Guion, Akahane-Yamada, & Yamada, 2004; Guion, Felge, & Loftin, 2000). The speech learning theory (Flege, 1995) suggests that it is easier to learn L2 phonemes that are

dissimilar to those in L1 than those that are more similar (Aoyama et al., 2004). On the other hand, L2 learners may find similar sounding phonemes difficult to differentiate.

Language learning research has shown that infants who have been exposed to specific distributions of similar sounds are more likely to discriminate these sounds. For example, Maye, Werker, and Gerken (2002) exposed infants to a continuum of pronunciation between "da" and "ta." Infants who were exposed to a unimodal frequency distribution of these sounds (one peak in the mean) were unable to discriminate between "da" and "ta." Conversely, infants who were exposed to a bimodal frequency distribution (peaks at the two extremes of the phonetic continuum between "da" and "ta"), which clearly differentiated "da" and "ta" as unique, were able to do so. Exposure to discriminating information is necessary to establish that similar phonemes are distinct; this learning process alone can be lengthy. Without receiving performance-specific feedback for gradual improvement the learning process may reach an asymptote or deteriorate.

Shaping

A systematic method to guide the acquisition of a complex skill that requires gradual approximations to a target performance is known in the behavioural literature as *shaping*. Shaping is the process of reinforcing attempts at a final target behaviour that are equal to or better than the previous best attempt. For quantifiable behaviours such as reaction time or duration, it can be relatively simple to determine if the current response is better than the previous best. However, target behaviours are often complex and need to be broken down into simpler versions of the behaviour (i.e., precursors). These precursors are then ranked into a hierarchy that increase in approximation to the target behaviour. The hierarchy begins with a behaviour that is in the existing response repertoire of the individual and approximates the target

behaviour in some way. For example, shaping a target behaviour of pressing a certain lever at the back of a room might involve reinforcing taking a step in the direction of the lever as a first step in the hierarchy. In this way, the previous best attempt is the precursor response highest in the hierarchy that the individual has emitted. As the individual learns and improves, the criterion for a correct response increases, and is determined by the person's own learning pace.

Shaping works by combining the effects of reinforcement and extinction delivered through differential reinforcement for progression towards a desired target behaviour. This can be conceptualized as four different types of responses, a) criterional responses that are reinforced, b) criterional responses that are not reinforced, c) noncriterional responses that are reinforced, and d) noncriterional responses that are not reinforced (Galbicka, 1994; Savage, 2001). Typically, shaping only involves the first and last type of responses; criterional responses are reinforced and those that do not are not reinforced. Occasionally, experimenters alter these proportions to implement a more complicated shaping procedure. For example, an experimenter can provide criterional responses with constant reinforcement, and fade out older responses with intermittent reinforcement (Midgley, Lea, & Kirby, 1989; Pear & Legris, 1987; Savage, 2001).

Regardless of experimenter intention, extinction is inherent in the shaping process; responses that do not meet the current criterion do not receive reinforcement. As the criterion moves towards the target behaviour, previously reinforced behaviour no longer meet with reinforcement and, therefore, become extinguished. The extinction aspect of shaping serves to eliminate earlier precursors as well as other irrelevant variations in behaviour (Savage, 2001). In addition, extinction of a previously reinforced behaviour increases the variability of responses (Neuringer, Kornell, & Olufs, 2001; Savage, 2001). This has been observed in humans (Jensen, Stokes, Paterniti, & Balsam, 2014; Lalli, Zanolli, & Wohn, 1994) and nonhumans (e.g., rats; see

Roberts & Gharib, 2006). Jensen and colleagues (2014) found that response variability in key presses greatly increased when responses that were previously reinforced with 10 points were met with 0 points. In fact, even a "downshift" of reinforcement which brought the reinforcer from 10 points to 1 point demonstrated this increase in response variability. The authors attributed the increase in variability to the unexpected decrease in reinforcement because they did not observe the same increase in variability within a condition where the reinforcer was unexpectedly increased from 10 points to 100 points. This variability increases the efficiency of shaping as it often elicits previously unobserved responses, which increases the likelihood that a closer approximation is emitted. Variability preceding learning is an established phenomenon in literature (Jensen et al., 2014).

Extinction is not the only contingency that can be implemented concurrently to selective reinforcement during shaping. Shaping is not necessarily a straightforward linear progression through the steps towards establishing the target behaviour; in some cases, backtracking may be implemented. Backtracking is a lowering of task difficulty, often achieved by reverting the criterion to a simpler response "step" that has been previously emitted by the individual. Backtracking is necessary when an individual is not able to meet with reinforcement within a set time period (Midgley et al., 1989). Midgley and colleagues (1989) observed that although all of their rats were successfully shaped to deposit a ball bearing into a designated hole in the floor, all had backtracked and moved downward in the response hierarchy during the learning process.

Shaping is a well-established behaviour modification procedure in establishing behaviours in nonhumans (Eckerman, Hienz, Stern, & Kowlowitz, 1980; Galbicka, 1994; Midgley et al., 1989; Pear & Legris, 1987; Preston, Umbricht, Wong, & Epstein, 2001; Savage, 2001; Stokes & Balsam, 1991). But this technique has also been used in many studies involving

humans; for example, shaping has proven to be an effective way to increase the attention span of individuals with schizophrenia (Silverstein, Menditto, & Stuve, 2001), decrease agoraphobia (Everaerd, Rijken, & Emmelkamp, 1973), and treat stuttering (Ryan, 1971).

In one study, shaping was used to achieve abstinence from cocaine in typical adults (Preston et al., 2001). Individuals for whom urinalysis demonstrated a decrease in cocaine use of at least 25% across sampling periods received money vouchers that could be redeemed for items supplied by the experimenters. Consecutive samples that met reinforcement criterion resulted in a voucher of a higher value in increments of $1.50, and every three consecutive vouchers received an additional $10 bonus voucher. The shaping procedure was implemented for three weeks, followed by five weeks of abstinence criterion (urinalysis showing no traces of cocaine use). This was compared to a control group that operated under the abstinence criterion for the same eight weeks. Preston and colleagues found that although both groups demonstrated similar decreases in cocaine use in the first three weeks, when the more stringent abstinence criterion was implemented in the last 5 weeks of the intervention phase, the shaping group demonstrated a further decrease in cocaine use while the control group did not. In fact, participants in the shaping group maintained a lower rate of cocaine use than those in the control group throughout the abstinence criterion as well as through a maintenance phase. The authors suggest that these results were due to participants meeting reinforcement at a higher rate in the three weeks of shaping than in the corresponding weeks of the control condition. Lastly, although there was no group difference in the overall number of vouchers earned, there were significantly fewer participants who earned no vouchers in the shaping group, compared to the control group. This last finding is pertinent to my study as it illustrates how, in establishing a new behaviour, shaping

can increase contact with reinforcement by gradually increasing the reinforcement criterion based on the individual's performance.

Shaping and verbal behaviours. Shaping has also been shown to be effective for training different vocal behaviours such as making requests, answering questions, and adjusting speech volume (Newman, Reinecke, & Ramos, 2009; Patterson, Teigen, Liberman, & Austin, 1975). In addition, shaping has been shown to be effective in establishing speech in mute individuals (Ayllon, & Kelly, 1974; Isaacs, Thomas, & Goldiamond, 1960) and in training pronunciation. For example, Hung (1976) successfully trained three non-verbal children with severe intellectual disability to mimic single syllable speech sounds using a shaping technique.

Shaping is useful for complex behaviours such as these because during shaping, even if the learner is far from the target or native pronunciation, beating their personal best will still result in reinforcement. Because reinforcement is not contingent on being "correct" but rather on showing improvement, shaping procedures produce a relatively constant rate of reinforcement throughout the learning process. Further, because reinforcement is usually provided for making modest steps towards the goal, the presence of the reinforcement can convey some information about the direction of the target behaviour, allowing the learner to discriminate the course of action that would lead to further increments in performance. This would shorten the learning process leading to the target pronunciation.

The idea of automating the shaping of pronunciation is not entirely new (Desrochers, Kinsner, & Pear, 1988; Pear, Kinsner, & Roy, 1987). Pear and colleagues (1987) used a computer to train four children with intellectual disability to pronounce the sound "ah." The shaping procedure used involved reinforcing any response falling within a "criterion region," and withholding reinforcement for responses outside of this region. Each criterional response

resulted in the criterion region shrinking by a certain amount, while each noncriterional response resulted in the criterion region expanding. This is similar to the procedure used by Pear and Legris (1974) which manipulated the dimensions of a virtual "reinforcement sphere" to shape behaviour in pigeons. The authors reported a small trend toward improvement, but no clear effects of the procedure.

Criticisms of shaping. Some researchers have expressed the view that shaping – particularly shaping "by hand" – is more of an "art" than a "science" due to the large amount of experimenter judgement that is involved (Galbicka, 1994; Pear & Legris, 1987). For example, the precursors of a final behaviour need to be determined *a priori*, and it is up to the experimenter to determine how much one step must differ from its precursor, and how many steps are necessary to reach the desired behaviour. The experimenter also determines when to move on to the next criterion. This could be done simply by determining if the current response has surpassed the current criterion, or it could involve assessing if an individual's response distribution has shifted enough (i.e., is being emitted at a consistent enough rate) to warrant moving on to the next criterion (Silverstein et al., 2001).

Shaping is not necessarily a straightforward process as not all individuals progress at the same, or at a constant, pace (Midgley et al., 1989). Occasionally backtracking is necessary. If an individual's progress deteriorates and the emitted responses do not meet reinforcement criterion over a certain amount of time, the experimenter may need to backtrack to a previous criterion of appropriate difficulty in order for the learner to maintain contact with reinforcement. Similarly, if an individual quickly masters an intermediate step, the experimenter needs to move the criterion to the next "step."

Overall, there are no specific guidelines regarding establishing precursors or how to progress through them, and the administration of the shaping procedure requires the experimenter to "stick to [these steps] and yet be flexible" (Martin & Pear, 2003, p. 128). Making these judgements requires a lot of experience and training. In addition, it is difficult for an experimenter to make judgements that are consistent over time due to factors such as fatigue, rater drift, or becoming more skilled over time. It is even more difficult for different experimenters to administer the same procedure consistently. Without consistency, it is difficult to maximize the efficacy of shaping, and "progress is likely to be retarded" (p. 128).

There has been some research investigating ways of automating the progression of reinforcement schedules in response to an individual's performance, particularly in terms of backtracking. "Algorithmic" shaping involves the implementation of rules that affect the process of shaping, often by a computerized process. A simple application of this would be to predetermine an acceptable length of time in which the individual does not make contact with reinforcement, and the degree to which the criterion would decrease if this time limit were to be exceeded. Pear and Legris (1987) used this type of approach to establish an arbitrary behaviour in pigeons. They used a computer to calculate a "virtual sphere" around the target area (a lower back corner of the chamber) that shrank after the pigeon's head made contact with the sphere (a behaviour that resulted in the delivery of reinforcement) or grew if a pre-determined amount of time passed without such contact. As the virtual sphere shrank, contact with it brought the pigeon's head closer to the target until finally the pigeon's head itself made contact with the target area (Pear & Legris, 1987).

A more complicated example of algorithmic shaping was seen in a study conducted by Midgley and colleagues (1989). These investigators taught rats to deposit ball bearings into a

hole in the floor by implementing six rules that dictated how the rate of reinforcement of a given response should decrease, how previous responses would be subject to extinction, and how backtracking would be implemented. Precursor responses were defined *a priori* from their previous experiences of shaping this behaviour by hand, and then ordered into a hierarchy that increased in approximation to the target behaviour. In Experiment 1, Midgley and colleagues (1989) were able to establish the efficacy of algorithm shaping, and in Experiment 2, they found that algorithmic shaping was equally effective when compared to hand shaping at establishing the behaviour, but resulted in less variable responses. Despite its proven efficacy, this approach is uncommon.

In sum, while traditional shaping procedures are effective, administering shaping to its maximum efficacy requires a great deal of training, practice, and skill (Martin & Pear, 2003). Further, because the "rules" of shaping are vague, it adapts well to the individual, but comparisons between studies are difficult as parameters (reinforcement density, the number of precursors, difficulty of criterion) cannot be held constant across studies.

Percentile schedules. To address some of the criticisms of traditional shaping, Platt (1973) proposed a "percentile reinforcement schedule." The efficacy of percentile shaping has been demonstrated in the literature and has been argued to be superior to that of traditional shaping schedules for behaviours that have a low occurrence, or are subject to fluctuations (Athens et al., 2007; Galbicka, Smurthwaite, Riggs, & Tang, 1997). Percentile schedules have been employed in a variety of settings with neurotypical adults, including shaping short-term pain sensitization and pain habituation (Hölzl, Kleinböhl, & Huse, 2005), and smoking cessation (Lamb, Kirby, Morral, Galbicka, & Iguchi, 2010; Lamb, Morral, Kirby, Iguchi, & Galbicka, 2005).

In contrast to traditional shaping by hand which uses a reinforcement criterion of the "previous best" response, a percentile schedule sets the reinforcement criterion as a certain percentile of the individual's distribution of possible responses, based on previously emitted responses. This decreases the likelihood that an extreme score that happened by chance would unduly increase the criterion and necessitate backtracking. In addition, using percentiles allows for a more moderate criterion to be applied, because the criterion can be a response that is well within the individual's ability (e.g., a criterion set at the 60th percentile of possible responses), which provides the experimenter with more flexibility. Of course, the experimenter can choose to establish the reinforcement criterion as the "previous best" with a percentile (e.g., 95th percentile). To determine the actual reinforcement criterion with a percentile schedule, first the "criterion rank" that corresponds to the desired criterional percentile (e.g., 60th percentile) is calculated, then previous responses are ranked from lowest to highest, and the previous response that is at the "criterion rank" is the current reinforcement criterion.

This "criterion rank" can be calculated with a formula (Galbicka, 1994), but prior to discussing the formula, I must reconceptualise the criterional percentile (e.g., responses surpassing the 60th percentile of the response distribution) as the inverse of the reinforcement probability (w). For example, the experimenter who wishes to set the criterional percentile at the 60th percentile would be implementing a likelihood of reinforcement of 0.4 because the top 40 percent of the response distribution would qualify for reinforcement. As the individual improves, the responses that comprise the top 40 percent of the distribution will change, although the criterional percentile will remain the same.

Another parameter of percentile schedules is the number of previous trials that is considered when ordinally ranking previous performances (m). This can be conceptualized as

the sample size of responses from which I infer the theoretical response distribution of the individual. The experimenter who wants to reinforce the top 40 percent of responses of the last nine trials would ordinally rank the performance (i.e., scores) of the last nine trials and compute the criterion rank (k) using the formula:

$$k = (m+1)(1- w)$$

where k is the criterion rank, m is the size of the window, and w is the probability of reinforcement. In this example:

$$k = (9+1)(1- 0.4) = 6$$

Therefore, the sixth ranked score (from lowest to highest) from the previous nine trials is the reinforcement criterion. In other words, if the participant's performance in the current trial surpasses six of the nine preceding performances, then the response is reinforced. With each successive trial, the scores are re-ranked and a new reinforcement criterion is established (see Table 1 for an example).

Using this approach, in which the participant's performance trend over the last m number of trials determines the actual reinforcement criterion, allows one to create a highly individualized and flexible learning environment. The criterion can be adjusted to the individual's performance whether it improves or declines, resulting in a more automated approach to determining when the reinforcement criterion should be lowered for backtracking and when the progress through criteria should be expedited. When an individual's response distribution has shifted, the reinforcement criterion is shifted along with it. There is no need for an experimenter to decide when or how much backtracking is necessary. The percentile schedule approach changes the criterion while attempting to maintain the desired reinforcement probability throughout the procedure. This approach keeps the task at a challenging but

manageable level, which is personalized to the individual and facilitates a steadier reinforcement density by ensuring that the opportunity to contact reinforcement is not unduly minimized.

Another advantage of percentile schedules is that they are formalized, with a formula to dictate which ordinal rank (*k*) is to be the reinforcement criterion based on the desired probability of reinforcement (*w*) and the number of previous trials used to sample performance (*m*). These variables can be manipulated experimentally to determine their impact on learning (Galbicka, Kautz, & Jagers, 1993; Galbicka, 1994; Lamb et al., 2005). The ability to specify the parameters of shaping also makes it easier to maintain consistency between experimenters (Athens et al., 2007) and replicate experimental protocols.

An experimenter could manipulate the responsiveness of the criterion and the difficulty of the task by adjusting the "sample window" (*m*) and the reinforcement probability (*w*) of the percentile schedule. But whether using a smaller *m* or a larger *m* is more effective is unclear. With smaller sample windows, the reinforcement criterion would be more responsive to changes in the participants' performance. However, with larger sample windows, it should be less likely that the reinforcement criterion would change due to performance variability or an extreme score. Lamb and colleagues investigated the effects of manipulating the sample size of previous trials (*m*) on shaping smoking cessation in neurotypical adults (Lamb, et al., 2005). Participants were adults who smoked more than 15 cigarettes a day and had no plans to quit in the next six months. A baseline was first established with 10 trials where no contingencies were implemented, after which participants were randomly assigned into one of two conditions: $m = 4$ or $m = 9$. The reinforcement probability (*w*) was the same across groups at 40%. Breath samples that indicated less smoking than either three out of the previous four trials or six out of nine trials, respectively, were met with reinforcement. As well, breath samples that were indicative of a smoking

abstinence for a day were also reinforced. Attendance and amount of reinforcement received did not differ between the two groups. While both groups demonstrated a significant decrease in smoking, participants in the $m = 4$ group achieved the target behaviour of abstaining from smoking for a day significantly more quickly than the $m = 9$ group. More participants were able to reach the target behaviour in the $m = 4$ group, although the difference was not significant.

Lamb and colleagues suggested that the smaller "sample window size" allowed the criterion to be more responsive to decreases in smoking, which required participants to maintain the decrease in order to keep a constant rate of contact with reinforcement. Similarly, any relapses would lead to the establishment of a criterion that is more easily met by the participants, thus maintaining contact with reinforcement. Some investigators have suggested that a smaller *m* is more effective when trying to decrease the frequency of behaviour, but that a larger *m* is more suitable for increasing behaviour frequency or duration (Athens et al., 2007). A larger *m* is also considered to be advantageous when sequential responses are not independent, because the larger "sample window" is less sensitive to variability in responses (Athens et al., 2007; Galbicka, 1994; Lamb, et al., 2005; see Table 1 for example). The sample window of previous trials used in the literature ranges from 4 to 20, although $m = 4$ or $m = 9$ are often used in order to end up with a whole number for the criterion rank (k), which makes it easier to select the reinforcement criterion (Lamb et al., 2005).

The Present Study

As expected, percentile schedules are more easily implemented with readily quantifiable behaviours such as frequency of response, or duration (e.g., attention span, task engagement). In the present thesis, I used a quantitative variable to measure pronunciation (accuracy score) in order to evaluate the efficacy of computer-administered shaping on pronunciation training. A

computer program assessed the accuracy of a speaker's pronunciation (defined as similarity to target native pronunciation) and provided reinforcement contingent upon performance improvement. Through this approach, I evaluated the effects of a shaping procedure in which the delivery of reinforcement was guided by a phonetic comparison algorithm. I hypothesized that the use of shaping would increase phonetic accuracy beyond what would be expected from exposure to the target pronunciation alone.

It should be noted that after a description of the main study (Study 1) and its results, a follow-up study is described (Study 2). The follow-up study aimed to address some of questions that arose during analysis of the data from Study 1.

STUDY 1

Method

Design

I used a single-subject ABAB reversal design (Cooper, Heron, & Heward, 2007; Miller, & Neuringer, 2000) composed of a pre-baseline phase, baseline (A) phases, and intervention (B) phases. Phonetic accuracy generated by the phonetic comparison algorithm was the dependent variable. The logic of a reversal design rests on the assumption that changes in the dependent variable are a direct result of exposure to the levels of the independent variable. However, irreversibility may be possible in skill acquisition paradigms or on occasions when the individual comes into contact with natural sources of reinforcement (Virues-Ortega & Martin, 2010).

During the pre-baseline condition, no programmed consequences were presented. Subsequently, I administered the baseline condition, in which I implemented a behaviour-independent (noncontingent) reinforcement schedule (NCR). Finally, during the intervention phases, I implemented a shaping, or behaviour-dependent (contingent) reinforcement schedule

(CR). The order of the conditions was the same for all participants (see Table 3). For simplicity, I will be referring to the first and second NCR sessions as NCR1 and NCR2, respectively. Similarly, the first and second CR conditions will be referred to as CR1 and CR2, respectively.

Often, an extinction—no reinforcement – condition is used as the baseline condition, as a form of control. However, because the intervention condition involves both reinforcement and a contingent relationship between reinforcement and an individual's response, a difference between the conditions could be attributed to the presence of reinforcement and not to the contingent relationship. Using NCR, which provides reinforcement independent of an individual's behaviour, as the baseline condition can eliminate the possibility that any observed effects are due to the presence of reinforcement. If there is a difference between the NCR and CR conditions, then it is likely that the effect is due to the contingent relationship between reinforcement and the behaviour (Thompson & Iwata, 2005).

Materials

Language History Questionnaire. Second language exposure was an important consideration in the present study as it could (a) increase an individual's ability to differentiate between similar speech sounds, and (b) affect accurate L2 pronunciation acquisition. Participants filled out a short online questionnaire, the *L2 Language History Questionnaire 2.0* (LHQ 2.0), to assess their language history, their language competency, and their exposure to other languages (Li, Zhang, Tsai, & Puls, 2013; see Appendix A). The questionnaire was developed by aggregating the questions most frequently asked in research relating to languages, and has a split-half reliability of .85 for the quantitative variables (Li, Sepanski, & Zhao, 2006). Although, traditionally, respondents are instructed to stop at question 4 ("*Do you speak a second language?"*) if a response of "no" is selected to this item (thus preventing self-identified

monolinguals from proceeding), participants in the current study were asked to complete the entire questionnaire. All participants reported exposure to a language other than English (see Table 2 for list of reported languages), but only those who reported "*limited*" fluency (i.e., a rating of 3) or less in all domains of their other language(s) were included in the final sample. For the purposes of this study, these individuals were considered to be English-speaking monolinguals; the average reported language ability in any language other than English for this group was between *"very poor*" and "*poor*" (*M* = 1.6 out of 7).

Assessment of oral praxic function. A rapid syllable repetition task was administered to assess individual differences in oral praxis. Oral praxis is the ability to organize sequenced oral movements, and is a possible confounding factor in assessing phonetic skills. Further, participants differing in oral praxis may perform differently in terms of their starting point or learning rate. The inclusion of this information could strengthen the external validity of our procedure. The task used in the present study is commonly used in assessing motor speech (Mateer, & Kimura, 1977; Wong, Murdoch, & Whelan, 2012). Participants were first asked to repeat a single syllable (*ba*) as quickly as possible for a 5 s period, then the procedure was repeated for another single syllable (*ga*). Next, the participants were asked to repeat a series of three syllables (*badaga*) as quickly as possible for a 5 s period. They were given three practice trials before the recorded test trial. Participants' attempts were recorded and scored by the experimenter according to the instructions by Kimura and Watson (1989). For the single syllables phase, the score was equal to the total number of correctly pronounced syllables. For multiple syllables, one point was added for each correctly pronounced and positioned syllable, while one point was subtracted for every omission. According to the normative score distribution provided in Kimura (1993), all participants performed within two standard

deviations of the mean (M = 27.38, 23.25, 28.5, for "ba", "ga", "badaga" respectively). One participant received a very low score – more than one standard deviation away from the norm. However, as this participant had normal speech and hearing ability, and his performance in the study did not differ from that of other participants, he was not excluded from the final sample.

Words used as stimuli. The target words were from Mandarin Chinese, which is unrelated to English. Mandarin has a high reliance on tones (pitch changes). Since pitch is more easily measured than other aspects of pronunciation, I chose to use Mandarin as the target language to maximize the algorithm's performance. I anticipated that the large difference between the two languages would make it easier for me to document gradual gains in phonetic accuracy. Ten natively-pronounced Mandarin Chinese words were used as target words. All target words had three syllables. There are four tones in Mandarin Chinese, and across the ten target words, each tone occurred in each position (first, middle, or last syllable) at least twice. Tones 1 and 4 appeared seven times in total while Tones 2 and 3 appeared eight times (see Appendix B for a list of words and tones). No tone was repeated within a word.

In order to prevent word-specific effects, five groups were created to counterbalance the words used in each phase (pre-baseline, NCR1, CR1, NCR2, CR2) across participants (see Appendix C for counterbalancing chart). Each word appeared once in the pre-baseline across word-order groups, and each word appeared twice in each of the reinforcement conditions. No two consecutive words appeared more than once across word-order groups.

Due to the differences in voice quality between male and female voices, there were two target pronunciations of each stimulus word: one spoken by a male voice and one by a female voice. Both of these individuals were native speakers and pronounced the words in Standard Mandarin, which is used as lingua franca across China.

Phonetic comparison algorithm. The phonetic comparison algorithm developed for this study compared multiple aspects of the participants' input pronunciation of a target word across trials (Ye, Leung, & Virués-Ortega, 2014). Through this comparison, the algorithm generated an estimation of pronunciation accuracy. According to the source filter model of human speech (Fant, 1960), speech sounds can be divided into two interacting components: a *source signal* and a *filter signal*. The source signal is the sound created in the vocal tract, and thus contains information pertaining to pitch, volume, and quality. The filter signal is the transformation of the source signal by speech organs. As such, it contains information used for vowel and consonant discrimination. Our comparison algorithm used information about both the source and filter signals to compare the participant's pronunciations with the target pronunciations. All sound manipulations involved in this process were conducted using Praat software for phonetic processing (version 5.3.56; Boersma & Weenink, 2013). The filter signal and source signal were extracted and analyzed according to the methods in Childers and Kesler (1978) and Boersma (1993).

Because the absolute pitch of voices can vary greatly due to individual differences, relative changes in pitch are more relevant than absolute pitch for pronunciation. Therefore, we transformed the pitch data to be relative to the mean pitch of the entire pronunciation. This relative pitch information was utilized when making comparisons.

Subsequently, we used an optimization algorithm to correct the timing differences between the input and the target pronunciations. Because the filter signal carries most of the phonetic information, it would allow us to more easily match the syllable onset of the two pronunciations. This was done by dividing the target filter signal into segments of 200 ms and then dividing the input filter signal into the same number of segments (Figure 2). Each segment

of the target was compared to its corresponding segment within the input filter signal, and the amount of similarity was calculated. Similarity is defined as the total difference between input and target divided by the total possible difference, subtracted from one. The duration of each of the segments in the input recording were generated by the algorithm using an exhaustive search that maximizes the similarity between target and input.

Similarity was also calculated for the pitch signal. No new segmentations were calculated for the pitch. In other words, the input pitch signal was segmented according to the final divisions that were used for the filter signal.

The weighted sum of filter similarity and pitch similarity was then computed, to yield a final accuracy score. The specific weights used were determined during preliminary validation of the algorithm (see below). The degree to which each input segment deviated from the "ideal" length of each corresponding target segment was also taken into account in the calculation of the final similarity score as the amount of time correction applied. In other words, an extreme amount of time correction was detrimental to the final accuracy score. The final similarity score ranged between 0 (lowest similarity) and 100 (highest similarity).

The algorithm was validated with an in-house validation process (Ye et al., 2014). We recorded five English monolinguals pronouncing four trisyllabic Mandarin words eight times, totalling 32 pronunciation clips. These clips were sent to four native Mandarin speakers to be rated for accuracy. The clips were also analyzed via the algorithm to obtain a similarity score. The similarity scores and the ratings from human raters were then analyzed for agreement. The agreement between the algorithm's ratings and those of the human raters was comparable to the agreement between the human raters.

Participants

The initial sample included 15 self-identified English monolinguals between the ages of 17 and 27 years (M = 18.8 years; five males, ten females) who were students at the University of Manitoba. Participants who were recruited through the PSYC 1200 Subject Pool were compensated with credit toward a course requirement. Based on responses to a language history questionnaire (described above), I excluded four participants who disclosed a fluency rating of 4 ("*average*") or above in any domain (e.g., listening, reading) for a language other than English. One additional participant was excluded due to software malfunction. Finally, data from two participants were excluded because of a small change in the procedure for recording responses (see Table 2 for excluded participants). The final sample, then, included eight participants (three males, five females) with a mean age of 18.7 years. All participants reported normal hearing and speaking abilities. For descriptive purposes, I also evaluated the articulatory fluency (oral praxis) of participants in the intervention study.

The study protocol was approved by the Psychology/Sociology Research Ethics Board of the University of Manitoba. Participants provided informed consent and were given an opportunity to ask questions before the study began as well as before every task.

Procedure

Participants were tested individually in a quiet (approximately 35 dB) 60 sq ft research room on the University of Manitoba premises. The oral praxis task and the LHQ 2.0 were administered at the beginning of the session, and any participants who failed to meet the inclusion criteria were excused. The remaining participants went on to complete the pronunciation training, which took approximately 100 minutes. All testing was completed in a single session.

Participants sat in front of a computer with a microphone and were instructed to repeat the words they heard on each trial, to the best of their ability. On-screen instructions informed participants when to listen to the recording and when to speak. One of the trisyllabic Mandarin words was presented at the beginning of the trial. Then a screen with a button labelled "*Record"* appeared with the instructions "*Press the button to record*." I implemented a 150 ms delay between pressing the button and the recording screen in order to avoid recording the sound of the button press. The 150 ms duration was determined during preliminary evaluations of the protocol where it was demonstrated to drastically reduce the likelihood of recording the sound of the button press, while still allowing me to record the participant's complete utterance.

During the recording screen, the participant was allowed up to 3 s to respond. If the participant did not respond within 3 s, the trial was presented again. Participants were instructed to press the button only when they were ready to make their response. In addition, participants were instructed to respond only once per recording and otherwise remain silent during the recording. Participants were given the opportunity to take a short break between words. The stimulus word was presented trial after trial until the pre-determined 10 trials were presented.

As previously discussed, even successful shaping of a behaviour often involves backtracking when the individual is unable to make consistent contact with reinforcement, and reverting to a previous (lower) reinforcement criterion. In this study, the backtrack procedure was designed to lower the difficulty of the task by breaking up the trisyllabic word into single syllables and providing participants an opportunity to practice each syllable on its own. After the presentation of the full word for 10 trials, each word was broken up into its constituent syllables, which were presented individually for 10 trials each. Next, both pairs of two consecutive syllables were presented to help participants generalize the pronunciation of individual syllables.

A secondary purpose of adding the backtrack phases was to ensure a relatively steady rate of reinforcement during the shaping procedure, which would mitigate the feelings of frustration reported during preliminary evaluations of our protocol. Repeating the same trisyllabic word for 25 consecutive trials quickly grew monotonous. By grouping the trials into phases of ten and introducing variety through the backtrack procedure, it was less monotonous and frustration and boredom were reported less often at the end of the session. After the backtrack procedure, the full word was presented again. Each of the seven phases (full word, three single syllables, two bisyllabic pairs, full word) contained 10 trials, totalling 70 trials per word. Then, I presented the next word and the experiment proceeded in the same manner.

In order to allow for individual differences in learning pace, shaping procedures often lack a pre-established number of trials. Also, as previously mentioned, the number of precursors required to shape a behaviour and the method of increasing or recombining initial precursors in a backtracking procedure are often based on the learner's performance. By contrast, I used a fixed number of trials, which allowed us to automate stimuli presentation and data recording.

This procedure was the same across pre-baseline, baseline, and intervention conditions (see Table 3). There were two words in each condition (Pre-baseline, NCR1, CR1, NCR2, CR2) totalling 10 words. The order of the words was counterbalanced across participants to prevent order and word-specific effects. The conditions of the study are described below.

Pre-baseline condition. A pre-baseline assessment was conducted to determine each participant's pre-existing language learning ability before the intervention. Participants were given the same instructions as in the other conditions, but no feedback was provided. The pre-baseline assessment allowed us to rule out the possibility that the effects of the shaping

procedure were caused by the presence of reinforcement (Thompson, Iwata, Hanley, Dozier, & Samaha, 2003).

Baseline condition. During baseline conditions, I presented positive feedback independently of the participant's performance in the task. The use of a percentile schedule allowed us to yoke the probability of NCR to the probability of reinforcement in the CR conditions (Galbicka, Fowler, & Ritch, 1991; Miller & Neuringer, 2000) which had been set *a priori* at 0.6 (see below).

In the baseline condition, the participants received noncontingent performance feedback after every trial in the form of a vertical green "performance bar" that filled up from the bottom. Labels "*Excellent"* at the top and "*Poor*" allowed the participants to interpret the feedback. Further, a horizontal line labelled as the participant's "*average so far"* indicated the criterion. Reinforcement was provided in the form of an auditory cue, social praise ("Good job!" message), and a visual depiction of the "performance bar" exceeding the criterion line. In the NCR condition, the visual depiction did not accurately reflect the participant's performance. The criterion line was accurate, but on trials where positive feedback was randomly provided the green bar depicting performance on the current trial surpassed the criterion line, while on trials where no positive feedback was provided the bar stopped just below this line. I did not conduct formal preference or reinforcer assessments. I assumed that performance-dependent feedback was reinforcing for our achievement-focused target population (university students). I did not expect to see differences between noncontingent reinforcement and pre-baseline assessment.

Intervention condition. In pilot testing, the intervention (CR) condition followed a traditional shaping procedure, with the reinforcement criterion being the participant's previous best within a phase. One of the main difficulties with shaping is that if a participant arbitrarily

produced a response that raised the reinforcement criterion by chance (an outlier), it becomes difficult for the participant to meet that criterion and access reinforcement again. During preliminary evaluation of the protocol, I found that pronunciation varied greatly from one trial to the next, and often participants would attain a high score due to chance but then were unable to meet this criterion again. This resulted in a low reinforcement density, decreased performance, and self-reported frustration and boredom. In light of this, I decided that a criterion that would be more forgiving of score fluctuations would be more suitable to our study. Thus, I decided to employ percentile schedules. When a larger number of previous trials (*m*) is taken into consideration to determine the reinforcement criterion the resulting reinforcement criterion is less susceptible to variation and extreme scores (i.e., change less frequently and less extreme scores).

From my pilot data, it was apparent that pronunciation scores varied greatly and frequently, so a larger *m* would be beneficial. While I wanted to decrease the effect of random extreme scores on the reinforcement criterion, I still needed a criterion that would be responsive to changes in the participants' performance given the number of trials. Thus, I chose to use a "sample window" of $m = 4$. This number has been used in the literature and has been found to be more effective than larger values for performances with high variability (Lamb, Morral, Galbicka, Kirby, & Iguchi, 2005). In addition, it is a small enough window for the criterion to be changed several times in the 10 trials and thus effect shaping, and yet it is large enough to have lowered sensitivity to variations in performance. For each phase where there were fewer than four trials, the criterion rank (*k*) was calculated by substituting the available number of previous trials for *m* in the equation $k = (m+1)(1- w)$ to determine the reinforcement criterion (Galbicka, 1994).

As for reinforcement probability (*w*), I decided to use $w = 0.6$ to increase the likelihood of reinforcement. Because *w* is only the *expected* average reinforcement probability and the *actual* average reinforcement density is dependent on the individuals' performance, oftentimes the actual reinforcement density is lower than that of *w* for difficult tasks. Using a $w = 0.6$ ensured that participants would be reinforced approximately half the time. Another boon to using percentile schedules is that the formalized parameters allow comparison between studies and participants, and facilitate yoking to a control condition.

As in the baseline condition, participants received visual feedback of their performance after every trial. In the intervention condition, however, feedback was contingent on performance, and every response that was more similar to the target pronunciation than the criterion was reinforced. The reinforcement used was the same as that described in the baseline condition. When a participant had completed all seven phases of a word, she or he received the following message: "Excellent! You have now mastered this word. Let's try another one!"

Analysis

Rarely, the end of a mouse click or heavy breath was captured in a recording. However, from preliminary comparisons between the original data and a "cleaned" version of the data that had clicks and tones (extraneous or otherwise) removed, this appeared to have little impact on the data. Given this, I did not expect the rare occurrence of prolonged mouse clicks to pose a problem to the data analysis. As such, the data were analyzed without cleaning.

I evaluated the results of the intervention study through visual analysis (Cooper et al., 2007, p. 248). Specifically, I analysed changes in trend, level, and variability of pronunciation accuracy. As is the norm for single-subject design, performance data (accuracy) were plotted on a line graph with the beginning of each new word and each new condition clearly marked. Data

recording for the dependent variable was automated; therefore, it was not necessary to assess interobserver reliability.

To complement the traditional method of visual analysis, I also calculated effect sizes to quantify the differences between conditions. The effect size calculation was developed specifically for use with single-subject designs and is comparable to the typical effect size *d* that is used in between-group analyses (Hedges, Pustejovsky, & Shadish, 2012; Shadish, et al., 2014).

This model conceptualizes each j^{th} observation from the i^{th} individual in the control condition to be:

$$Y_{ij} = \mu^C + \eta_i + \varepsilon_{ij}$$

where μ^C is the mean performance of the control condition, η_i is the individual level effect, and ε_{ij} is the amount of change between observations within one participant. The variance of η_i is the between-subject variance, denoted by τ^2. The variance of ε_{ij} is the within-subject variance, denoted by σ^2. Each Y_{ij} observation in the treatment condition is likewise conceptualized as:

$$Y_{ij} = \mu^T + \eta_i + \varepsilon_{ij}$$

where μ^T is the mean performance in the treatment condition. This model assumes that Y_{ij} is normally distributed, that there is no time trend within the data, and that there is only a weak first-order autocorrelation φ. Furthermore, the total variance can be calculated by summing the within-subject variance (σ^2) and the between-subject variance (τ^2).

The commonly used between-group effect size, Cohen's *d*, is calculated by determining the difference between the mean of the control group and the mean of the treatment group. The difference is divided by the standard deviation:

$$\delta = \frac{\mu_t - \mu_c}{\sigma}$$

Using $\sigma^2 + \tau^2$, this model can also calculate the difference between control and treatment, then divides it by the standard deviation:

$$\delta = \frac{\mu^T - \mu^C}{\sqrt{\sigma^2 + \tau^2}}$$

However, this method of calculation was vulnerable to biases and an unbiased estimator of this effect size (Hedges *g*) was formulated (for more information, please refer to Hedges, Pustejovsky, & Shadish, 2012). This analysis was found to be suitable for as few as three independent cases and therefore, is an appropriate analysis for me to use with my sample size of eight independent cases. I used an SPSS macro (DHPS, version 1.11; Marso & Shadish, 2014) to calculate this effect size (*g*) and its variance (*Var[g]*). According to Cohen, an effect size of 0.2 is considered a small effect, and effect sizes of 0.5 and 0.8 are considered medium and large, respectively (Howell, p. 235).

In addition to the above, supplemental repeated-measures ANOVA analyses were conducted to determine if there were statistically significant differences between conditions that were too small to observe through visual analysis. An $\alpha = 0.05$ was used for all of these statistical analyses.

Results

The purpose of this study was to test the efficacy of computer-administered shaping on pronunciation accuracy. My hypothesis was that shaping would be an effective pronunciation training technique, and I expected the results to show clear gains in phonetic accuracy during shaping relative to baseline.

In general, visual analysis of full word pronunciations suggested that there were no noticeable differences in performance between the NCR conditions and the CR conditions. The average full word performance of all participants (both before and after the backtrack procedure)

did not differ in terms of accuracy, variability or trend between conditions or between phases (see Figure 3). However, this may have been due to the averaging of performance across participants. When looked at individually, I found a small difference in the expected direction between CR and NCR conditions for three participants (Figure 4). Conversely, two participants exhibited the reverse effect where performance in NCR conditions was slightly better than performance in CR conditions (Figure 5). Lastly, for three participants, there appeared to be little difference between the two conditions (Figure 6).

Individual participant results are discussed below. Supplementary graphs depicting each participant's performance with individual words are provided in Appendix D.

I will first discuss the participants who demonstrated the hypothesized result. Overall, participant 306 performed marginally worse during full word presentations in the NCR conditions than in either the pre-baseline or the CR conditions (Figure 4). The average performance across words in each condition did not differ in terms of trend or accuracy (Figure 7). Performance was more variable in the pre-baseline than in the other two conditions for the single syllable presentation phases, as well as for the final full word presentation, but that could be due to the fact that the pre-baseline was averaged over two words while NCR and CR contained four words each. Interestingly, the second word presented in each CR condition and in NCR2 (see performance of *yurongyi, dalishi, maigangqin* in Figures D1 and D2) were better overall than the first word presented in those same conditions; however, no learning trend was observed. Performance did not differ before and after the backtrack procedure in any of the conditions.

For 310, the average full word performance in NCR2 was lowest of all the conditions (Figure 4). Performance in NCR1 was also marginally worse than the other conditions. Unlike

other participants, there were no visible differences in variability between the pre-baseline, CR, and NCR (Figure 11). A learning trend was observed for one word in the pre-baseline condition (see *hengaoxin* in Figure D8). Performance of a word in CR1 dropped after the backtrack procedure (see *yurongyi* in Figure D8). 310's performance in the backtrack phases suggests that she was able to pronounce the individual syllables of *yurongyi*, but it was the generation of bisyllabic or full word utterances that she had difficulty with. The variability of the second full word presentation of *dalishi* increased after the backtrack procedure (see *dalishi* in Figure D9).

Participant 315 appeared to perform worse in NCR1 than in all other conditions, particularly when compared to CR1, which immediately followed it (Figure 4). However, there were no other notable differences between conditions. There were no visible differences in variability between the pre-baseline, CR, and NCR (Figure 14). No learning trends were observed. Performance did not differ before and after the backtrack procedure in any of the conditions.

Of the two participants who unexpectedly demonstrated superior performance in NCR compared to CR conditions, participant 308 appeared to perform much better in the two NCR conditions than pre-baseline or CR conditions (Figure 5). In particular, she had the most accurate pronunciation in NCR2. Once again, performance in the pre-baseline condition varied more than that in the NCR and CR conditions (Figure 9). When performance with individual words was inspected, 308's performance during the backtrack procedure varied greatly (Figures D4 and D5). An increasing learning trend was observed during the first full word presentation of one word in the pre-baseline condition (see *tanglaoya* Figure D4), suggesting that even in the absence of feedback 308 was able to learn. Performance did not differ before and after the backtrack procedure in any of the conditions.

Participant 314 appeared to perform marginally worse in the two CR conditions relative to the other conditions (Figure 5). There were no visible differences in variability between the pre-baseline, CR, and NCR (Figure 13). The second word of NCR1 (*dalishi*) was pronounced most accurately overall, but a learning trend was not evident with the full word. Inspection of its backtrack phases revealed a small increasing trend in the single syllable phase of "*da*" (see *dalishi* in Figure D12).

The first of the three participants who showed little variability in performance across NCR and CR conditions was participant 307. Interestingly, her average full word performance in the pre-baseline condition was lower than that seen in either NCR and CR (Figure 6), suggesting that the mere presence of reinforcement is beneficial to some individuals, even when it is not contingent on performance. Performance was also more variable in the pre-baseline than in the other two conditions (Figure 8). No learning trends were observed. Performance did not differ before and after the backtrack procedure in any of the conditions.

For 309, the average full word performance was lowest in NCR2, and there were no differences in full word performance between any of the other conditions (Figure 6). Performance in the pre-baseline condition varied greatly in comparison to the other two conditions where it was much more stable (Figure 10). No learning trends were observed. Overall, performance did not differ before and after the backtrack procedure in any of the conditions. However, variability in full word pronunciation decreased after the backtrack procedure for *dalishi* and *sanshiwu* (Figure D4). Variability decreased across the backtrack phases with minimal variability in the bisyllabic phase. This supports the assumption that the backtrack procedure can make a difficult task easier by breaking down a difficult word into

simpler parts and allowing the participant to practice joining a few of these parts together before trying the full word again.

For 313, the average full word performance in CR1 was marginally higher than the other conditions, but there were no other differences (Figure 6). Performance in the pre-baseline condition varied greatly in comparison to the other two conditions where it was much more stable (Figure 12). The variability of the second full word presentation of *dalishi* increased after the backtrack procedure (see *dalishi* in Figure D10). A learning trend was observed during the first full word presentation of *jinzita* in NCR1 which disappeared after the backtrack procedure and the variability of the word increased (Figure D10). This suggests that in cases where an individual is successfully learning a word, interrupting that process by breaking up the word could introduce confusing information that undermines the progress made by the individual and thus lead to lower performance.

Effect sizes and complementary analyses.

The effect size calculations indicated that there was little difference between NCR and CR for full word presentations (g = 0.15, s = 0.066, power = .63), individual syllables (g = -0.06, s = 0.065, power = .12), and bisyllabic pairs (g = 0.09, s = 0.074, power = .25).[1] However, compared to the pre-baseline assessment (where no reinforcement was provided) there was a moderate positive effect of providing CR on the accuracy of full word pronunciations (g = 0.41, s = 0.069, power = 1). This effect was much smaller when syllables were presented individually

[1] Note that the *post hoc* power analyses conducted here were performed using an SPSS macro that was developed specifically for use with this effect size calculation, and takes into account the autocorrelation and intraclass correlation of single-subject design data (Shadish, et al., 2014).

(g = 0.27, s = 0.058, power = .93) or in pairs (g = 0.26, s = 0.061, power = .97). Finally, compared to pre-baseline, there was also a small positive effect of NCR for full word presentation (g = 0.27, s = 0.061, power =.99), individual syllables (g = 0.27, s = 0.054, power = .97), and bisyllabic pairs (g = 0.21, s = 0.059, power = .88).

A 2 (Condition: NCR, CR) x 2 (Order: first, second) repeated-measures ANOVA was conducted on accuracy data from all eight participants to complement the results from visual analysis. No interaction was found (F(1, 7) = 0.03, p = .87, power = .053) and, consistent with visual analysis, no main effects of condition (F(1, 7) = 0.28, p = 0.61, power = .075) or order (F(1, 7) = 0.04, p = .85, power = .053) were found. However, as previously mentioned, this may have been due to the averaging of results across participants. Because of this, I repeated this analysis using only the data obtained from the three participants who demonstrated the expected difference between conditions. Once again, the interaction between condition and order was not significant (F(1, 2) = 0.01, p = .92, power = .051), and no main effect of Order was found (F(1, 2) = 0.07, p = .82, power = .053). However, consistent with the results from the visual analysis, in these three participants a significant main effect of Condition was observed (F(1, 2) = 67.98, p = 0.014, power = .97), indicating that performance in the CR conditions (M = .71, SD = .04) was statistically higher than performance in the NCR conditions (M = .64, SD = .06).

Word effects.

As part of preliminary data analyses, I discovered that performance was not consistent across words, and that several participants performed best when pronouncing the word *dalishi*. Due to the small sample size, the average performance in each condition could be greatly affected if one stimulus word is easier to pronounce than others. Therefore, I inspected the data to assess participants' accuracy in pronouncing each word, in each condition (Figure 15). It is

important to keep in mind that due to the sample size, the average performance of each word was drawn from one or two sets of data. I found that performance on *dalishi* and *yurongyi* was typically high, suggesting that they were relatively easier stimulus words, while performance on *jinzita* was generally poor, suggesting that it is a more difficult word to pronounce.

Because each participant was presented with only two words in each condition, scores obtained in a given condition could be inflated by the presence of a relatively easy-to-pronounce word. For example, 308's performances in the NCR conditions were higher than in CR1 (Figure 5), which is a finding contrary to what would be expected given what is known about CR and NCR. However, this pattern may have arisen because the relatively easy words *dalishi* and *yurongyi* were used in NCR1 and NCR2 respectively, while the more difficult words *jinzita* and *maigangqin* were both presented in CR1 (Figure 16). The same sequence occurred for 314 who performed in the same manner. Further, the reverse of this effect occurred for 310 who encountered *yurongyi* and *dalishi* in CR1 and CR2, while both *jinzita* and *maigangqin* were in NCR2 (Figure 17). Unsurprisingly, 310 performed better in the two CR conditions than in NCR2.

Discussion

The purpose of this study was to test the efficacy of computer-administered shaping on pronunciation accuracy by comparing manipulating feedback. I expected to find little difference between the pre-baseline condition (where no feedback was provided) and the NCR conditions. Further, I expected participants to perform the best when feedback was contingent on performance (CR).

Overall, no clear differences were found between CR and NCR conditions; therefore, the hypothesis was rejected. This was supported by the effect size analysis where no differences

between CR and NCR conditions were found, as well as by the results of the two-way repeated-measures ANOVA conducted on data obtained from all eight participants.

One limitation of the present study was that power was generally low in my statistical analyses. I utilized a single-subject experimental design, and in single-case research, sample sizes are typically low (three to five participants), and visual analysis of data is preferred over statistical analysis. For these reasons, I did not conduct an *a priori* power analysis to determine and recruit the number of participants that would have given my analyses more power, as I would have done had I planned to use a group design.

However, the absence of differences between conditions may have been due more to the collapsing of differing results across participants than to a lack of power. Indeed, three participants demonstrated small effects where accuracy was significantly higher in CR conditions than in NCR conditions (Figure 5), and this subset analysis did have high power (.97). Although the difference between conditions seen in this exploratory analysis is promising, the effect was small. Given that there were only three participants who showed the predicted effect, the findings from this analysis provide only preliminary support for the conclusion that the software program developed for the current project can lead to improvements in L2 pronunciation in some learners.

When the two reinforcement conditions were compared to the pre-baseline (where no reinforcement was provided), the effect sizes of NCR phases suggest that the mere presence of reinforcement can have a small, positive effect on learning, even if it is response-independent (g = 0.27, 0.21, 0.27, for trisyllabic, bisyllabic, and individual syllable presentation, respectively). Interestingly, CR seemed to be differentially beneficial for learning the full word (g = 0.41), while effect sizes for backtrack phases in CR conditions were similar to NCR (g = 0.26, 0.27, for

bisyllabic and individual stimuli, respectively). The larger effect size for full word presentations suggests that CR may have a positive effect on learning over and above the effects of the presence of reinforcement.

However, the effects found in these comparisons may be biased. The pre-baseline assessment only contains two words while both CR and NCR contain four words each. This asymmetrical comparison may be more susceptible to the effects of variance. For many participants, pre-baseline performance was more variable than that seen in either CR or NCR conditions, which, when coupled with the small sample size, may have resulted in the effect sizes found. I believe the comparison between CR and NCR conditions (which revealed no effect) to be more accurate as these conditions had an equal number of words, and performance in both conditions showed minimal variability.

Although, my single-subject reversal design did not provide a large amount of statistical power and cannot demonstrate the generalizability of my protocol with only eight participants, it allowed me to observe learning at an individual level while still evaluating treatment effects. My study used purposely–developed software to explore a specific research question. The hypotheses and research rationale were drawn from similar areas (e.g., computer-assisted pronunciation training, hand-shaping of pronunciation in clinical populations), but it was unclear how individuals would respond to the protocol. Using the established tradition of visual analysis allowed for the possibility of observing different trends within conditions, and observing whether individuals differ in their response patterns in different phases of the implementation of this protocol. In addition, as the purpose of my study was to evaluate the efficacy of computer-administered shaping in training pronunciation, I was more interested in assessing clinically

significant (rather than statistically significant) differences between conditions. Assessing differences observed by way of visual analysis was more pertinent to this goal.

There were several possible reasons why no clear differences were evident between CR and NCR conditions in the present study. The first of these relate to the stimuli and the number of training trials that were used. It was possible that the trisyllabic words selected as stimuli were too difficult for participants to learn in the short time of the study (floor effect). Indeed, word analysis revealed that some stimulus words were quite difficult to learn – a fact that may have unduly affected performance in certain conditions. In addition, participants were only exposed to the full word for 20 trials in total, and these trials were split into two sets of 10 trials, separated by five phases (50 trials) of backtrack procedure. While breaking each trisyllabic word down into its constituent parts may have been beneficial, giving participants only 10 final trials in which to practice the full word pronunciation after this procedure ended may not have allowed them to fully utilize the newfound information gained from the backtrack procedure. Further, shaping is typically administered without a limit to the number of times a certain precursor target is presented, and the number of trials for each precursor is dependent on the individual's performance. I had set a limit to facilitate the automation of the procedure, but that may have affected the efficacy of the procedure.

A second factor that may have contributed to the lack of shaping effects was that the *differences* in pronunciation between reinforced and not-reinforced trials may not have been salient enough. In the present study, the task was subtle and involved finer motor skills, and participants were making changes to their pronunciation that they themselves may not have been able to detect. It is possible that the changes in pronunciation (whether in tone or in muscles involved) that resulted in reinforcement were so miniscule that it was not clear what behaviours

would be met with reinforcement. In other words, if the differences between reinforced and not-reinforced trials were too subtle, participants may not have discerned any differences at all and would have operated as if exposed to a non-contingent environment.

Third, the lack of effect may have been due to ineffective reinforcers. I did not perform formal preference and reinforcer assessments in order to select and evaluate the reinforcing effects of the consequent events used during the study. Further, I only reinforced criterional responses, and did not "fade out" previously criterional responses with intermittent reinforcement. This may have resulted in precursors that were not firmly established as participants progressed, making performance more prone to backtracking and variable responses.

Fourth, it is also possible that strong order effects caused by the NCR phase precluded the effect of the intervention – an analogous effect in the associative learning literature is known as learned irrelevance. Participants experienced the NCR condition first and, through this, they may have learned that the reinforcement was independent of their performance and was uninformative. This may have led them to disregard reinforcement information in subsequent conditions. While participants could (potentially) have learned new information in the CR condition to override this, this may have been difficult to do given that each condition targeted only two words. Participant 306 appeared to have learned the contingency in CR, as he scored higher on the second word than the first in each CR condition. While this pattern lends some support to this possibility, this may have been due to chance or idiosyncratic factors because no other participant demonstrated a similar pattern of learning. A potential strategy to address this issue in a future study would be to counterbalance condition order across subjects and expose participants to the CR condition first. Another strategy might be to increase the number of words

in each condition to allow more time for participants to learn if the reinforcement is informative and to make use of it.

Fifth, my attempts to make the task easier may have introduced information that interfered with learning. I chose to break down the trisyllabic stimuli into simpler syllables to provide participants with an opportunity to practice individual syllables that were particularly difficult. The backtrack procedure I employed may have created an opportunity for poor pronunciations to be reinforced because the criterion for reinforcement was reset for each phase. In other words, the pronunciation of a syllable during a backtrack phase could be worse than during the full word presentation, yet still be reinforced if it qualified as an improvement within that phase. It was observed that participants occasionally received reinforcement when a syllable was mispronounced during the backtrack procedure, and that this mispronunciation was then carried through to the second full word presentation.

While making the task easier is undoubtedly important in the shaping of a difficult task, using a percentile schedule already fulfills that goal. Percentile schedules adjust the reinforcement criterion based on the performance of a set number of previous trials; if a participant begins to perform worse than before, the criterion is lowered accordingly. By trying to break down the word and make the task easier in this way, I may have unnecessarily complicated matters and introduced an opportunity for confusion. In future research, it may be useful to compare the effectiveness of a percentile schedule and a backtrack procedure separately as well as in various combinations; doing so may help us to determine whether percentile schedules decrease task difficulty as effectively as backtracking, and if there are advantages to using both combined. Repeating this study using a reinforcement criterion that carries over between phases would also be advantageous.

Sixth, it is possible that participants may have followed their own strategies in an attempt to improve their performance, which may have made them insensitive to the feedback provided. This is supported by the fact that some participants (307, 309 and 313) showed similar performance across all three conditions (pre-baseline, baseline, shaping). This could indicate that these participants were not utilizing the feedback that was provided. While participants have reported that they found the reinforcement used in this study to be "motivating", it may have been useful to inquire about their usage of the feedback.

Lastly, the fact that the majority of participants did not show the predicted effect may be related to their language learning ability, which most participants rated as being "limited." It may be that the task was too difficult for individuals who have had limited experience in learning a second language. The stimulus language may have been too dissimilar to English, making the task too difficult. Using a language from a more similar language family, such as French, may have elicited different results. Age could also have been a factor in language learning, which may in term have increased the difficulty of the task. It is well accepted that in general, children (lower age of acquisition) are more adept at learning second languages than adults (higher age of acquisition; Birdsong, 2006; Dixon, et al., 2012). In fact, a L2 speaker with a lower age of acquisition is more likely to be perceived as native-like (Abrahamsson & Hyltenstam, 2009). Because my study recruited only adults to learn L2 pronunciation, this may have increased the difficulty of the task and diminished any learning trends. A future study with younger participants may be better suited to test my hypothesis

Finally, the high responding during the baseline conditions may have contributed to the lack of effect found in my study. While it is true that the stimuli could have been too difficult to adequately demonstrate learning in the limited time of one study session, participants also

performed at relatively high levels from the very beginning, often at 0.6 or above. This leaves me with a window of 0.6 to 1.0 to observe an effect, and given the difficulty of the task, a high score such as 0.9 or above was rare. The majority of participants performed at a limited range of 0.6 to 0.85, and it is very difficult to observe effects in such a narrow range. This high responding in baseline could be due to the innate language learning ability of participants or the program being insufficiently sensitive. Scaling the scores to take into account the fact that scores cannot be as low as 0 because any utterance will at least result in some sort of similarity in terms of pitch may help with this issue.

STUDY 2

To address some of the issues encountered in Study 1, I conducted a follow-up study in which I utilized a simpler version of the protocol and simpler (single syllable) stimuli. Specifically, the backtrack procedure was removed to eliminate the possibility of introducing confusing information to participants which may have interfered with learning. A related issue in the Study 1 was that targets were broken up into phases of 10 trials, which may have been too few trials for participants to acquire the skills to pronounce the target before the target changed. In the follow-up study, participants were given at least 25 uninterrupted trials to learn the target pronunciation which reduced task difficulty.

The stimuli in this follow-up study were likewise simplified. In Study 1, trisyllabic words may have been too difficult to learn given the time constraints of the study. The follow-up study instead used single syllables as targets, further reducing task difficulty. Lastly, this study addressed the issue of order effects, namely the possibility that participants may have learned that the feedback was irrelevant due to the condition order (NCR then CR). To address this,

participants in this study began with the CR condition prior to the control (no reinforcement) condition.

Method

Design

As the exposure to the protocol in Study 1 might have provided irreversible learning experiences to participants, in this follow-up study I used a multiple baseline design, which is considered to be an alternative to the reversal design when the behaviour under investigation is irreversible. It is a widely used experimental design for evaluating treatment effects in the field of applied behaviour analysis, and it is highly flexible (Cooper, Heron, & Heward, 2007).

The multiple baseline procedure involves a "time-lagged" application of treatment where the treatment is administered to one participant after a baseline phase, but the other participants remain in the baseline phase. After an effect has been demonstrated in the treatment phase for the first participant, treatment is administered to the second participant, and so on. While two of such "tiers" are sufficient, three to five tiers are typically used.

The basic logic of the multiple baseline is that the effects of treatment can be evaluated on an individual basis by contrasting a participants' performance in the baseline and treatment phases, but it can also be assessed across individuals by comparing the performance of participants who have or have not undergone treatment at a particular point in time. If the pattern of responding from the participant in the treatment condition deviates from baseline, and from that of non-treated participants, then the experimenter can conclude that the change in the participant's responses is due to the treatment. If a treatment effect can also be demonstrated in the second participant, relative to the third participant, and so on, this conclusion will be strengthened.

In the present study, I used a modified multiple baseline design. A multiple baseline typically has all participants begin with the same condition, and after baseline is established, treatment is administered in a staggered manner across participants. Due to the limitations of the software, I needed to complete data collection with a participant in one sitting, and so I decided to modify the procedure and use a more automated approach. Instead of administering treatment depending on performance of other participants, I manipulated the number of trials in each condition to simulate the longer baseline for each subsequent participant in the traditional multiple baseline procedure. The first participant had 25 trials in each of the two conditions, the second had 35, and the third had 45 trials in each condition. While this approach departs from the traditional method by imposing a limited number of trials, it still potentially allows for demonstration of a delayed condition effect, which (when present) would illustrate the efficacy of the treatment. For reasons outlined below, during the "baseline" phase I delivered CR, while during the "treatment" phase participants received no feedback.

Materials

Language History Questionnaire. As in the Study 1, participants completed the LHQ 2.0 on-line. Data from this questionnaire were used to assess participants' language history, language competency, and exposure to other languages. All participants reported exposure to a language other than English (see Table 4 for list of reported languages), but only those who reported "*limited*" fluency (i.e., a rating of 3) or less in all domains of their other language(s) were included in the final sample. The average reported language ability in any language other than English for this group was between *"poor*" and "*limited*" (*M* = 2.25 out of 7).

Assessment of oral praxic function. The same rapid syllable repetition task used in Study 1 was administered to assess individual differences in oral praxis. All participants

performed within two standard deviations of the mean on this task (*M* = 31.33, 29.33, 31.67, for "ba", "ga", "badaga" respectively).

Words used as stimuli. In this follow-up study, stimuli included two single syllables from one of the words included in Study 1. Specifically, the syllables *da* and *li* were taken from the word *dalishi* . In order to prevent syllable-specific effects, presentation order was counterbalanced across participants.

Phonetic comparison algorithm. The phonetic comparison algorithm used in this study was the same as that described in Study 1.

Participants

The initial sample included six self-identified English monolinguals between the ages of 17-20 years (*M* = 18.5 years; one male, five females) who were students at the University of Manitoba. Participants who were recruited through the PSYC 1200 Subject Pool were compensated with credit toward a course requirement. Based on responses to the language history questionnaire, I excluded three participants who disclosed a fluency rating of 4 ("*average*") or above in any domain (e.g., listening, reading) for a language other than English. The final sample, then, included three participants (three females) with a mean age of 18.7 years. All participants reported normal hearing and speaking abilities. For descriptive purposes, I also evaluated the articulatory fluency (oral praxis) of participants in the intervention study.

The study protocol was approved by the Psychology/Sociology Research Ethics Board of the University of Manitoba. Participants provided informed consent and were given an opportunity to ask questions before the study began as well as before every task.

Procedure

Participants were tested individually in a quiet (approximately 35 dB) 60 sq ft research room at University of Manitoba premises. The oral praxis task and the LHQ 2.0 were administered at the beginning of the session, and any participants who failed to meet the inclusion criteria were excused. The remaining participants went on to complete the pronunciation training, which took between 25 to 45 minutes. All testing was completed in a single session.

Participants sat in front of a computer with a microphone and were instructed to repeat the sounds they heard on each trial, to the best of their ability. On-screen instructions informed participants when to listen to the recording and when to speak. One of the target syllables was presented at the beginning of the trial. Then a screen with a button labelled "*Record"* appeared with the instructions "*Press the button to record*." The same 150 ms delay used in Study 1 was implemented between pressing the button and the recording screen in order to avoid recording the sound of the button press. During the recording screen, the participant was allowed up to 3 s to respond. If the participant did not respond within 3 s, the trial was presented again. Participants were instructed to press the button only when they were ready to make their response. In addition, participants were instructed to respond only once per recording and otherwise remain silent during the recording. The stimulus syllable was presented trial after trial until the pre-determined number of trials was presented. Then, the second syllable was presented. There was no backtrack procedure in this study.

In this study, I employed a simpler protocol, with only two conditions. The study began with the contingent reinforcement condition (CR) in order to address the issue of learned irrelevance. This condition used the same percentile schedule and reinforcement as Study 1. The control condition was presented after the CR condition, and it provided no feedback to

participants. I chose to use no reinforcement as the control condition rather than NCR because the purpose of this follow-up study was to determine if the software can implement a very basic shaping protocol. To evaluate this, I needed to maximize the effect of the CR condition relative to the control condition. Effect size analyses from Study 1 suggested that it was possible that the mere presence of positive feedback could have an effect on performance, which would result in a smaller difference between the two conditions and make it more difficult to draw the conclusion that the protocol performed as intended.

Analysis

I evaluated the results with visual analysis in the same manner as in the Study 1. Similarly, I used the same SPSS macro (DHPS, version 1.11; Marso & Shadish, 2014) to calculate the effect size (*g*) and its variance (*Var[g]*). However, I used a variant of the analysis that has been developed for multiple baseline designs (Hedges, Pustejovsky, & Shadish, 2013; Shadish et al., 2014). Finally, a repeated-measures *t*-test was conducted to complement the results from the visual analysis. Although it would be informative to conduct further statistical analyses on the subset of participants who did respond to treatment, as in Study 1, it was not possible to conduct such an analysis with data from only one participant without violating assumptions of the test, which would result in biased and inaccurate results.

Results

The purpose of this follow-up study was to address some of the plausible problems with Study 1, to determine if the software was able to shape the pronunciation of a single syllable if given enough trials. I hypothesized that performance in the CR condition would be better than that seen in the no-feedback control condition. The effect size analysis appeared to support this hypothesis. Thus, there was a moderate effect of condition, with pronunciation accuracy being

higher with CR than in the control (no feedback) condition (g = -0.46, s = 0.16, power = .67).[2] This conclusion was not entirely supported by the visual analysis, as only one participant demonstrated a drop in performance in moving from the CR to the control condition (404, see Figure 18). However, there did appear to be slightly less variability in the CR than in the control condition for all three participants.

A supplemental repeated-measures *t*-test comparing mean performance in the CR condition (M = .73, SD = .08) to mean performance in the control condition (M = .72, SD =.05) revealed no significant difference between conditions, $t(2) = 0.07$, $p = 0.95$, power = .05.

Discussion

While the effect size analysis suggested a moderate difference between the CR and control conditions, visual analysis of the data did not fully support this conclusion. By way of visual analysis, I observed the expected difference in performance in only one participant, and although the low variability in her data suggests that, the difference was due to treatment and that more participants may have revealed the same pattern of performance, this singular case still provides a limited degree of functional control to render the effect believable.

The results of the supplemental *t*-test corroborates the observation that there was no difference between CR and control conditions. However, this study had a small sample size of three participants, meaning that power of this analysis was low. As well, the multiple baseline design resulted in an unequal number of trials in each condition between participants, which may

---

[2] As with Study 1, the *post hoc* power analysis was conducted using an SPSS macro developed specifically for use with this effect size calculation. The analysis takes into account the autocorrelation and intraclass correlation of single-subject design data (Shadish, et al., 2014).

have affected the validity of this analysis. The effect size analysis may have been similarly affected by the limited sample size and by the small number of data points. This may have left the effect size analysis vulnerable to bias and the effects of variability.

In addition, as is often the case in statistical analyses, effects found in calculations may not amount to discernable differences in practice. As the objective of my thesis was to evaluate the efficacy of a pronunciation training program, a clinical difference between conditions is more important than a difference found only in calculations. Therefore, my discussion will focus on the results from the visual analysis, which did not reveal a clear difference between the two conditions.

This follow-up study shed some light on some of the issues from Study 1. First, I had hypothesized that the backtrack procedure in the main study may have hindered participants by introducing confusing information in the middle of the learning process, which resulted in no increase in accuracy after the backtrack procedure. The results from the follow-up study did not support this hypothesis. The follow-up study did not implement the backtrack procedure, and yet participants did not perform better than in Study 1.

Second, a potential limitation of Study 1 was the difficulty of learning trisyllabic stimuli within the given time constraints. In Study 2, I investigated whether adjusting target word difficulty through decreasing the target word length would elicit the expected results by using single syllables instead of trisyllabic words as stimuli. Furthermore, to address the issue of time constraint, participants had at least 25 trials of uninterrupted learning of one syllable – more than twice the number used in each phase of Study 1. While the task was easier and participants were given more trials to learn the target pronunciation, no learning trend was observed. Both participants 403 and 404 were able to obtain fairy high scores (0.7 and above by the second trial)

which may indicate that the chosen stimuli were too easy (ceiling effect). Perhaps with syllables that are more difficult I would have observed a learning trend or a more pronounced effect between conditions. However, participant 407 did not demonstrate such a ceiling effect, and, even when provided 45 trials to learn each syllable, did not demonstrate a learning trend or difference between conditions. This suggests either that the software itself was ineffective and needs to be fine-tuned, or that the number of trials necessary to induce a learning effect was much higher than anticipated and may not have been practical to implement within one session without greatly simplifying the initial protocol.

Lastly, this study looked at condition order and learned irrelevance as possible explanations for the lack of an effect in Study 1. Here, the treatment condition preceded the control condition; as such, participants did not have the opportunity to learn that the feedback was independent of performance and discard it as being irrelevant. Further, the control condition did not provide feedback to differentiate between the two conditions. Despite this, there was no observed difference between the two conditions, which suggests that the contingent reinforcement was just as uninformative as no feedback.

GENERAL DISCUSSION

My thesis evaluated the efficacy of a purposely-developed software to shape second language pronunciation. Unfortunately, no clear differences between CR and NCR conditions were found. However, for some participants in Study 1 and one participant in the follow-up study, CR resulted in a small increase in performance. This coupled with the effect sizes found between CR and no-reinforcement conditions is promising. It suggests that there may be a CR effect that my protocol is unable to demonstrate, or that people may differ in their language learning ability and that it is difficult to shape pronunciation without a computer program that is

more flexible in terms of adjusting the difficulty and the number of trials. Future studies with a larger sample size could clarify this issue.

A problem with both studies was that statistical power was limited. This is a common problem with single-subject experimental designs, which, due to their focus on individual-level behaviours, typically involve recruitment of a small number of participants. Addressing my research questions using a group design would allow me to perform *a priori* power calculations and determine the number of participants needed to achieve an acceptable level of power.

Although the lack of clear results in the present work differs from findings in other computerized language training studies that have demonstrated a treatment effect (e.g., Hirata, 2004), I used a different approach in determining accuracy and delivering feedback. Rather than segmenting utterances into phonemes for comparison or comparing participant utterances to expected errors (e.g., rope vs. robe; Burleson, 2007), pronunciations from participants were compared to the target pronunciation; the similarity between the two was the accuracy score. This allowed for the possibility of training pronunciation with individuals with intellectual or developmental disabilities -- populations in which typical methods are not effective.

Another strength of my study was that I implemented a computer-administered shaping procedure to deliver feedback, and there is little research in the efficacy of computer-administered shaping, particularly in the area of pronunciation. Pear and colleagues (1987) implemented a computer-shaping protocol to shape single syllable utterances in individual with intellectual and developmental disabilities and, as in the present work, they did not find a clear improvement as a result of the shaping procedure.

Further investigation is needed to determine the appropriate task difficulty level needed to test the hypothesis. While Study 1 revealed that trisyllabic stimuli were too difficult for

participants to learn within the constraints of the study, the follow-up study demonstrated that decreasing the task difficulty may not be as simple as shortening the stimuli; utilizing a single syllable resulted in possible ceiling effects in some participants, but for another even 45 trials were insufficient to learn a syllable. Difficulty could be adjusted by manipulating the phonetic similarity of the target words to the participants' native language (English) – removing novel phonemes, or selecting a language more similar to English, would make stimuli easier to reproduce. Mandarin is very different from English, and shaping the pronunciation of a more similar language may be easier. Replication with two samples of monolinguals comparing the learning of languages from two different language families (e.g., Italian and Korean) could effectively test the theory regarding language family.

In addition, one could investigate the effect of implementing two approaches thought to optimize the effects of shaping. First, one could incorporate prompts, or additional precursors that lead the learner temporarily to the right answer (e.g., cues or demonstrations on how to position lips, tongue, and throat muscle to facilitate correct performance); this has been shown to facilitate shaping in other work (Ray & Ray, 2008). It is possible that the backtrack procedure implemented in Study 1 (individual syllables) was still too difficult for participants, and that a backtrack procedure that can deliver these precursors may be able to provide a wider range of task difficulty and can better adapt to individuals.

Second, one could employ more gradual shaping goals. For example, instead of placing the backtrack procedure in the middle of a word presentation, I could begin with it and gradually transition to a full trisyllabic word. Alternatively, I could try shaping only single phonemes as an initial goal that, once mastered, could be followed by shaping syllables containing the mastered phonemes, and so forth. While no learning trend was observed when the number of learning

trials was increased to 45, increasing the number of study sessions may be beneficial. In particular, if sessions continued until progress was made (as in traditional shaping protocols), it could inform us of the number of trials necessary for individuals to learn a syllable, and the protocol could be altered accordingly.

A potential limitation of this study relates to the fact that it included only English monolinguals, which may have further increased task difficulty. Future studies could examine the effect of the current procedure among individuals who are experienced at learning languages. This procedure may have been ineffective on monolinguals because they have no second language learning experience, and thus have not developed strategies or a sensitivity to cues to acquire accurate pronunciation of foreign words. Bilinguals could be more sensitive to language learning cues and the difference between the CR and NCR procedures, resulting in a more pronounced effect. As well, they may be better able to differentiate between phonemes and may have had experience controlling their oral muscles to mimic foreign sounds, thus making them more aware of their own pronunciations and the differences between pronunciations that were and were not reinforced. Conversely, if the bilingual individual already had a strategy in place, shaping could interfere with language learning by introducing confusing information. Again, replication with bilinguals could address this issue.

One of the greatest strengths of this study is its multi-disciplinary nature. This study investigated the application of behaviour modification procedures through an electronic medium and its effect on second language pronunciation training on typical adults. Studies utilizing behaviour modification procedures are seldom conducted with typical adults, so this extends existing shaping literature to this population. Similarly, shaping is typically administered by trained therapists in a one-to-one format, and this study furthers research on electronically-

administered shaping. Future research should investigate this further because, if successful, computer-assisted shaping could prove to be a highly effective form of intervention.

In particular, I expect that pronunciation training would benefit greatly from a computer-administered shaping procedure by making it much more efficient and accessible. Not only could it facilitate second language learning in typical adults, but it has potential applications with clinical populations as well. Accurate pronunciation is a prominent problem affecting many individuals with autism spectrum disorder (ASD; Newman et al., 2009; Rutter & Bartak, 1971). The few procedures focusing on vocal shaping as a component of early language acquisition in children with ASD are labour intensive (requiring one-to-one intervention), and rely on the ability of the therapist to identify pronunciation progress (Koegel, O'Dell, & Dunlap, 1988; Newman et al., 2009). A computer-administered procedure could lower the number of hours of instruction required for pronunciation training with individuals with ASD, freeing up valuable therapist time for other interventions. However, further research is needed to fine-tune the computer-administered shaping protocol before the potential of using this software with clinical populations can be determined.

SUGGESTIONS FOR FUTURE RESEARCH

As previously mentioned, there are many ways for future research to improve upon this study. Changing the protocol so that participants are first introduced to simpler single-syllable stimuli before gradually introducing longer syllables would likely yield a different result. A similar modification would be to add more precursors, including demonstrations (e.g., animations of tongue position and movement) of correct pronunciation. Additionally, increasing the number of trials or the number of sessions would increase the likelihood of observing a treatment effect.

In terms of stimuli, future studies could lower the task difficulty by using simpler stimuli. In addition, modifications to the target pronunciation, such as slowing down the pronunciation and exaggerating tonal differences, may help participants acquire skills to produce novel utterance and improve their ability to perceive the correct pronunciation. Another method of presenting stimuli is to embed the word within a sentence. Pronunciation of a word in isolation is often different from the pronunciation of the same word within a sentence, and using in-context pronunciation for training may further increase the accuracy of second language pronunciation, particularly for tonal languages such as Mandarin Chinese.

Improving participants' ability to perceive the correct pronunciation would help them discern differences between their own pronunciation and the target pronunciation, which may aid participants in discerning what behaviours would be met with reinforcement. This could increase participants' sensitivity to the contingencies of each condition, and increase the likelihood of observing a treatment effect. Future studies could also conduct a preference assessment to ensure that the reinforcement used in the study is indeed reinforcing to the participants.

Lastly, this study only considered and screened for second language exposure and oral praxis as potential confounds. In future research, it would be useful to assess participants' attention (selective and sustained), short-term/auditory memory, and phonetic or pitch discrimination skills, as individual differences in any of these areas could influence performance.

References

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning, 59*(June), 249–306. doi:10.1111/j.1467-9922.2009.00507.x

Ayllon, T., & Kelly, K. (1974). Reinstating Verbal Behavior in a Functionally Mute Retardate. *Professional Psychology, 5*(4), 385–393. doi:10.1037/h0021326

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., & Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: the case of Japanese /r/ and English /l/ and /r/. *Journal of Phonetics*, *32*, 233–250. doi:10.1016/S0095-4470(03)00036-6

Athens, E. S., Vollmer, T. R., & St. Peter Pipkin, C. C. (2007). Shaping Academic Task Engagement with Percentile Schedules. *Journal of Applied Behavior Analysis*, *40*(3), 475–488. doi:10.1901/jaba.2007.40-475

Birdsong, D. (2006). Age and second language acquisition and processing: A selective overview. *Language Learning, 56*, 9–49. doi:10.1111/j.1467-9922.2006.00353.x

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences,17,* 97-110.

Boersma, P. & Weenink, D. (2013): Praat: doing phonetics by computer [Computer program]. Version 5.3.56, retrieved from http://www.praat.org/

Bresnahan, M. J., Ohashi, R., Nebashi, R., Liu, W. Y., & Morinaga Shearman, S. (2002). Attitudinal and affective response toward accented English. *Language & Communication*, *22*(2), 171–185. doi:10.1016/S0271-5309(01)00025-8

Burleson, D. (2007). Improving Intelligibility of Non-Native Speech with Computer-Assisted Phonological Training. *IULC Working Papers Online, 7*, 1–18. Retrieved from https://www.indiana.edu/~iulcwp/pdfs/07-Burleson5.pdf

Childers, D. G., & Kesler, S. B. (Eds.). (1978). *Modern spectrum analysis* (Vol. 331). New York: IEEE Press.

Cooper, J, O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). New Jersey: Pearson.

Desrochers, M., Kinsner, W., & Pear, J. J. (1988). Evaluation of the assessment component of a system for shaping vocal behavior in severely speech-disabled individuals. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 4*, 1647–1648. doi:10.1109/IEMBS.1988.94831

Dixon, L. Q., Zhao, J., Shin, J.-Y., Wu, S., Su, J.-H., Burgess-Brigham, R., … Snow, C. (2012). What We Know About Second Language Acquisition: A Synthesis From Four Perspectives. *Review of Educational Research*. doi:10.3102/0034654311433587

Eckerman, D., Hienz, R., Stern, S., & Kowlowitz, V. (1980). Shaping the location of a pigeon's peck: Effect of rate and size of shaping steps. *Journal of the Experimental Analysis of Behavior, 3*(3), 299–310. Retrieved from http://onlinelibrary.wiley.com/doi/10.1901/jeab.1980.33-299/abstract

Everaerd, W. T. A. M., Rijken, H. M., & Emmelkamp, P. M. G. (1973). A comparison of "flooding" and "successive approximation" in the treatment of agoraphobia. *Behaviour Research and Therapy, 11*(1), 105–117. doi:10.1016/0005-7967(73)90073-9

Fant, G. (1960). *Acoustic Theory of Speech Production: With calculations based on x-ray studies of Russian articulations*. The Hague: Mouton and Co.

Flege, J. E. (1995). Second Language Speech Learning Theory, Findings, and Problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 233–277). Timonium, MD: York Press.

Galbicka, G. (1994). Shaping in the 21st century: Moving percentile schedules into applied settings. *Journal of Applied Behavior Analysis, 4*(27), 739–760. Retrieved from http://onlinelibrary.wiley.com/doi/10.1901/jaba.1994.27-739/abstract

Galbicka, G., Fowler, K., & Ritch, Z. (1991). Control over response number by a targeted percentile schedule: Reinforcement loss and the acute effects of d-amphetamine. *Journal of the Experimental Analysis of Behavior, 56*(2), 205–215. Retrieved from http://onlinelibrary.wiley.com/doi/10.1901/jeab.1991.56-205/abstract

Galbicka, G., Kautz, M. A., & Jagers, T. (1993). Response acquisition under targeted percentile schedules: a continuing quandary for molar models of operant behavior. *Journal of the Experimental Analysis of Behavior*, *60*(1), 171–184. doi:10.1901/jeab.1993.60-171

Galbicka, G., Smurthwaite, S., Riggs, R., & Tang, L. W. (1997). Daily rhythms in a complex operant: targeted percentile shaping of run lengths in rats. *Physiology & Behavior, 62*(5), 1165–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9333214

Guion, S. G., Flege, J. E., & Loftin, J. D. (2000). The effect of L1 use on pronunciation in Quichua–Spanish bilinguals. *Journal of Phonetics, 28*, 27–42. doi:10.1006/jpho.2000.0104

Government of Canada, Citizenship and Immigration Canada. (2012). *Temporary foreign workers present on December 1st by province or territory and urban area, 2008-2012.* Retrieved from http://www.cic.gc.ca/english/resources/statistics/facts2012-preliminary/04.asp

Government of Canada, Citizenship and Immigration Canada. (2012). *Temporary foreign students present on December 1st by province or territory and urban area, 2008-2012.* Retrieved from http://www.cic.gc.ca/english/resources/statistics/facts2012-preliminary/04.asp

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods, 3*(3), 224–239. doi:10.1002/jrsm.1052

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods, 4*, 324–341. doi:10.1002/jrsm.1086

Hirata, Y. (2004). Computer Assisted Language Learning Computer Assisted Pronunciation Training for Native English Speakers Learning Japanese Pitch and Durational Contrasts for Native English Speakers Learning Japanese. *Computer Assisted Language Learning*, *17*(3-4), 357–376. doi:10.1080/0958822042000319629

Hölzl, R., Kleinböhl, D., & Huse, E. (2005). Implicit operant learning of pain sensitization. *Pain, 115*(1-2), 12–20. doi:10.1016/j.pain.2005.01.026

Howell, D. C. (2013). *Statistical Methods for Psychology* (Eighth Ed.) Belmont, CA: Cengage Learning.

Hung, D. W. (1976). Teaching mute retarded children vocal imitation. *Journal of Behavior Therapy and Experimental Psychiatry, 7*(1), 85–88. doi:10.1016/0005-7916(76)90052-5

Independent Administrative Institution Japan Student Services Organization. *International Students in Japan 2011*. Retrieved from http://www.jasso.go.jp/statistics/intl_student/data11_e.html

Isaacs, W., Thomas, J., & Goldiamond, I. (1960). Application of operant conditioning to reinstate verbal behavior in psychotics. *Journal of Speech and Hearing Disorders, 25*(1), 8-12.

Jensen, G., Stokes, P. D., Paterniti, A., & Balsam, P. D. (2014). Unexpected downshifts in reward magnitude induce variation in human behavior. *Psychonomic Bulletin & Review, 21*(2), 436–44. doi:10.3758/s13423-013-0490-4

Kimura, D. (1993). *Neuromotor mechanisms in human communication*. New York: Oxford University Press.

Kimura, D., & Watson, N. (1989). The relation between oral movement control and speech. *Brain and Language, 37*(4), 565–590. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=2479446

Koegel, R. L., O'Dell, M., & Dunlap, G. (1988) *Journal of Autism and Developmental Disorders, 18*, 525-538. doi: http://dx.doi.org/10.1007/BF02211871

Lalli, J., Zanolli, K., & Wohn, T. (1994). Using extinction to promote response variability in toy play. *Journal of Applied Behavior Analysis, 27*(4), 735–736. Retrieved from http://onlinelibrary.wiley.com/doi/10.1901/jaba.1994.27-735/abstract

Lamb, R. J., Kirby, K. C., Morral, A. R., Galbicka, G., & Iguchi, M. Y. (2010). Shaping smoking cessation in hard-to-treat smokers. *Journal of Consulting and Clinical Psychology, 78*(1), 62–71. doi:10.1037/a0018323

Lamb, R. J., Morral, a R., Galbicka, G., Kirby, K. C., & Iguchi, M. Y. (2005). Shaping reduced smoking in smokers without cessation plans. *Experimental and Clinical Psychopharmacology, 13*(2), 83–92. doi:10.1037/1064-1297.13.2.83

Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology, 46,* 1093–1096. doi:10.1016/j.jesp.2010.05.025

Li P, Sepanski S, Zhao X. (2006). Language history questionnaire: A web-based interface for bilingual research. *Behaviour and Brain Research, 38*, 202-210.

Li, P., Zhang, F., Tsai, E., & Puls, B. (2013). Language history questionnaire (LHQ 2.0): A new dynamic web-based research tool. *Bilingualism: Language and Cognition, 17*(03), 673–680. doi:10.1017/S1366728913000606

Lindemann, S. (2003). Koreans, Chinese or Indians? Attitudes and ideologies about non-native English speakers in the United States. *Journal of Sociolinguistics, 7,* 348–364. doi:10.1111/1467-9481.00228

Marso, D. & Shadish, W.R. (2014): Software for Meta-analysis of Single-Case Design [Computer program]. Version 1.11, retrieved from http://faculty.ucmerced.edu/wshadish/software/software-meta-analysis-single-case-design

Martin, G., & Pear, J. (2003). *Behavior modification what it is and how to do it* (7th edition). Upper Saddle River, NJ: Prentice Hall

Mateer, C., & Kimura, D. (1977). Impairment of nonverbal oral movements in aphsia. *Brain and Language, 4,* 262-276. doi: 10.1016/0093-934X(77)90022-0

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82,* B101–11. doi:10.1016/S0010-0277(01)00157-3

Midgley, M., Lea, S. E., & Kirby, R. M. (1989). Algorithmic shaping and misbehavior in the acquisition of token deposit by rats. *Journal of the Experimental Analysis of Behavior,*

52(1), 27–40. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1338941&tool=pmcentrez&rendertype=abstract

Miller, N., & Neuringer, A. (2000). Reinforcing variability in adolescents with autism. *Journal of Applied Behavior Analysis*, *33*(2), 151–65. doi:10.1901/jaba.2000.33-151

Munro, M. J., & Derwing, T. M. (1999). Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning*, 49(2), 73–97. doi:10.1111/j.1467-1770.1995.tb00963.x

Neuringer, A., Kornell, N., & Olufs, M. (2001). Stability and variability in extinction. *Journal of Experimental Psychology: Animal Behavior Processes*, *27*(1), 79–94. doi:10.1037/0097-7403.27.1.79

Newman, B., Reinecke, D., & Ramos, M. (2009). Is a reasonable attempt reasonable? Shaping versus reinforcing verbal attempts of preschoolers with autism. *The Analysis of Verbal Behavior*, *25*, 67–72. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2779071&tool=pmcentrez&rendertype=abstract

Patterson, R. L., Teigen, J. R., Liberman, R. P., & Austin, N. K. (1975). Increasing speech intensity of chronic patients ("mumblers") by shaping techniques. *The Journal of Nervous and Mental Disease*, *160*, 182–187. doi:10.1097/00005053-197503000-00004

Pear, J. J., Kinsner, W., & Roy, D. (1987). Vocal shaping of retarded and autistic individuals using speech synthesis and recognition. *Proc. IEEE Engineering in Medicine and Biology Soc*, 1787–1788.

Pear, J. J., & Legris, J. A. (1987). Shaping by automated tracking of an arbitrary operant response. *Journal of the Experimental Analysis of Behavior*, *2*(2), 241–247. Retrieved from http://onlinelibrary.wiley.com/doi/10.1901/jeab.1987.47-241/abstract

Platt, J. R. (1973). Percentile reinforcement: Paradigms for experimental analysis of response shaping. *Psychology of Learning and Motivation, 7*, 271-296.

Preston, K. L., Umbricht, A., Wong, C. J., & Epstein, D. H. (2001). *Shaping cocaine abstinence by successive approximation. Journal of Consulting and Clinical Psychology*, *69*, 643–654. doi:10.1037/0022-006X.69.4.643

Ray, J. M., & Ray, R. D. (2008). Train-to-code: A adaptive expert system for training systematic observation and coding skills. *Behavior Research Methods, 40*, 673-693. doi: 10.3758/BRM.40.3.673

Roberts, S., & Gharib, A. (2006). Variation of bar-press duration: Where do new responses come from? *Behavioural Processes, 72*, 215–223. doi:10.1016/j.beproc.2006.03.003

Rosetta Stone [computer program]. Retrieved from http://www.rosettastone.com

Rutter, M., & Bartak, L. (1971) Causes of infantile autism: some considerations from recent research. *Journal of Autism and Childhood Schizophrenia, 1*, 20-32

Ryan, B. P. (1971). Operant procedures applied to stuttering therapy for children. *Journal of Speech and Hearing Disorders*, *36*(2), 264-280.

Savage, T. (2001). Shaping: A multiple contingencies analysis and its relevance to behaviour-based robotics. *Connection Science*, *13*(3), 199–234. doi:10.1080/09540090110096196

Schmid, P., & Yeni-Komshian, G. (1999). The effects of speaker accent and target predictability on perception of mispronunciations. *Journal of Speech, Language and Hearing*, *42*, 56–64. Retrieved from http://jslhr.asha.org/cgi/content/abstract/42/1/56

Shadish, W. R., Hedges, L. V., Pustejovsky, J. E., Boyajian, J. G., Sullivan, K. J., Andrade, A., & Barrientos, J. L. (2014). A d-statistic for single-case designs that is equivalent to the usual between-groups d-statistic. *Neuropsychological Rehabilitation, 24*(3-4), 528–53. doi:10.1080/09602011.2013.819021

Silverstein, S., Menditto, A., & Stuve, P. (2001). Shaping attention span: an operant conditioning procedure to improve neurocognition and functioning in schizophrenia. *Schizophrenia Bulletin, 27*(2), 247–257. Retrieved from http://psycnet.apa.org/journals/szb/27/2/247/

Stokes, P., & Balsam, P. (1991). Effects of reinforcing preselected approximations on the topography of the rat's bar press. *Journal of the Experimental Analysis Behavior, 55*(2), 213–231. Retrieved from http://onlinelibrary.wiley.com/doi/10.1901/jeab.1991.55-213/abstract

Thompson, R. H., & Iwata, B. A. (2005). A review of reinforcement control procedures. *Journal of Applied Behavior Analysis, 38*(2), 257–78. doi:10.1901/jaba.2005.176-03

Thompson, R. H., Iwata, B. A., Hanley, G. P., Dozier, C. L., & Samaha, A. L. (2003). The effects of extinction, noncontingent reinforcement and differential reinforcement of other behavior as control procedures. *Journal of Applied Behavior Analysis, 36*, 221-238.

Virués-ortega, J., & Martin, G. L. (2010). Guidelines for sport psychologists to evaluate their interventions in clinical cases using single-subject designs. *Journal of Behavioral Health and Medicine, 3*, 158–171.

Wong, M. N., Murdoch, B. E., & Whelan, B. M. (2012). Lingual kinematics during rapid syllable repetition in Parkinson's disease. *International journal of language & communication disorders / Royal College of Speech & Language Therapists, 47*, 578–88. doi:10.1111/j.1460-6984.2012.00167.x

Ye, Z., Leung, J., & Virués-ortega, J. (2014). An algorithm for evaluating word pronunciation accuracy relative to a single exemplar. Manuscript submitted for publication.

Tables and Figures

Table 1

Example of Percentile Schedule With Differing Sample Size Parameter, w = 0.4

Current trial	$m = 4, k = 3$ $w = 0.4$		$m = 7, k = 5$ $w = 0.4$		$m = 10, k = 7$ $w = 0.4$	
	Previous trials[a]	Criterion[b]	Previous trials	Criterion	Previous trials	Criterion
1	2, 3, 4, 9	4				
0	1, 2, 3, 9	3				
12	0, 1, 2, 3	2				
8	0, 1, 2, 12	2	0, 1, 2, 3, 4, 9, 12	4		
4	0, 1, 8, 12	8	0, 1, 2, 3, 8, 9, 12	8		
11	0, 4, 8, 12	8	0, 1, 2, 3, 4, 8, 12	4		
4	4, 8, 11, 12	11	0, 1, 2, 4, 8, 11, 12	8	0, 1, 2, 3, 4, 4, 8, 9, 11, 12	8
7	4, 4, 8, 11	8	0, 1, 4, 4, 8, 11, 12	8	0, 1, 2, 3, 4, 4, 8, 9, 11, 12	8
6	4, 4, 7, 11	7	0, 4, 4, 7, 8, 11, 12	8	0, 1, 2, 3, 4, 4, 7, 8, 11, 12	7
4	4, 6, 7, 11	7	4, 4, 6, 7, 8, 11, 12	8	0, 1, 2, 4, 4, 6, 7, 8, 11, 12	7
12	4, 4, 6, 7	6	4, 4, 4, 6, 7, 8, 11,	7	0, 1, 4, 4, 4, 6, 7, 8, 11, 12	7
5	4, 6, 7, 12	7	4, 4, 4, 6, 7, 11, 12	7	0, 4, 4, 4, 6, 7, 8, 11, 12, 12	8
	4, 5, 6, 12	6	4, 4, 5, 6, 7, 11, 12	7	4, 4, 4, 5, 6, 7, 8, 11, 12, 12	8

[a]Scores from the specified *m* number of previous trials ordered from smallest to largest
[b] Criterion is determined by the k^{th} previous score from smallest to largest

Table 2

Known Languages Reported by Participants in Main Study

Languages	Reported by participants
English	all
French	301[a], 304[a,b], 306, 308, 309, 310, 311[b], 313, 315
Tagalog	304[a,b], 305[b], 314
Polish	307
Icelandic	309
Traditional Chinese	309
Hausa	312[b]
Yoruba	312[b]
Arabic	312[b]
Tamil	302[a]
Bengali	303[a]
Ukranian	304[a,b]

[a] Excluded due to software malfunction or procedural changes
[b] Excluded due to language exposure

Table 3

Phases of Stimuli Presentation Across Conditions Within One Participant.

Condition	Full word	Backtrack					Full word
		Single syllables			Bisyllabic pairs		
Pre-baseline	Yigeren	Yi	Ge	Ren	Yige	Geren	Yigeren
	Hengaoxin	Hen	Gai	Xin	Hengao	Gaoxin	Hengaoxin
NCR1	Tanglaoya	Tang	Lao	Ya	Tanglao	Laoya	Tanglaoya
	Liangtiaolu	Liang	Tiao	Lu	Liangtiao	Tiaolu	Liangtiaolu
CR1	Sanshiwu	San	Shi	Wu	Sanshi	Shiwu	Sanshiwu
	Yurongyi	Yu	Rong	Yi	Yurong	Rongyi	Yurongyi
NCR2	Jinzita	Jin	Zi	Ta	Jinzi	Zita	Jinzita
	Maigangqin	Mai	Gang	Qin	Maigang	Gangqin	Maigangqin
CR2	Bowuguan	Bo	Wu	Guan	Bowu	Wuguan	Bowuguan
	Dalishi	Da	Li	Shi	Dali	Lishi	Dalishi

Table 4

Known Languages Reported by Participants in Follow-up Study

Languages	Reported by participants
English	all
French	401[a], 402[a], 403, 404, 405[a]
Yoruba	402[a], 405[a]
Afrikaans	401[a]
German	401[a]
Hausa	405[a]
Spanish	406

[a] Excluded due to language exposure

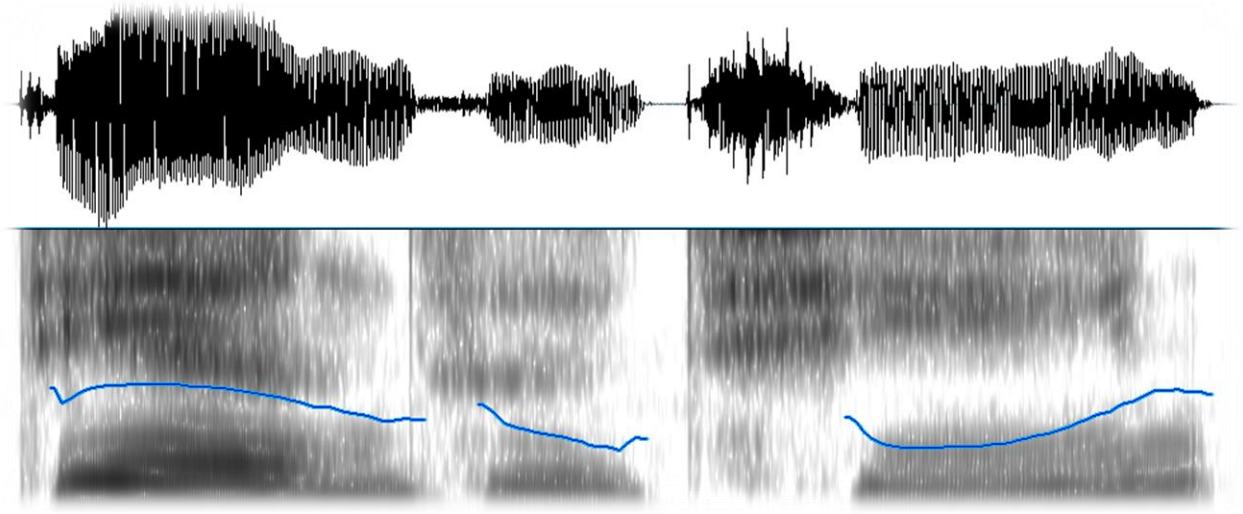

Figure 1. Spectrogram (top) and pitch graph (bottom) of a female voice speaking Mandarin

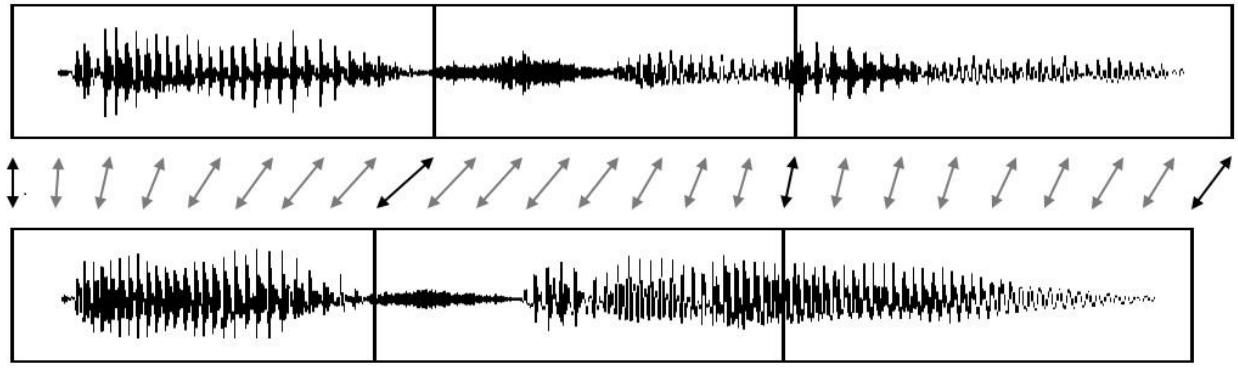

Figure 2. Example of timing correction applied to input recordings to match syllable onset of target recording

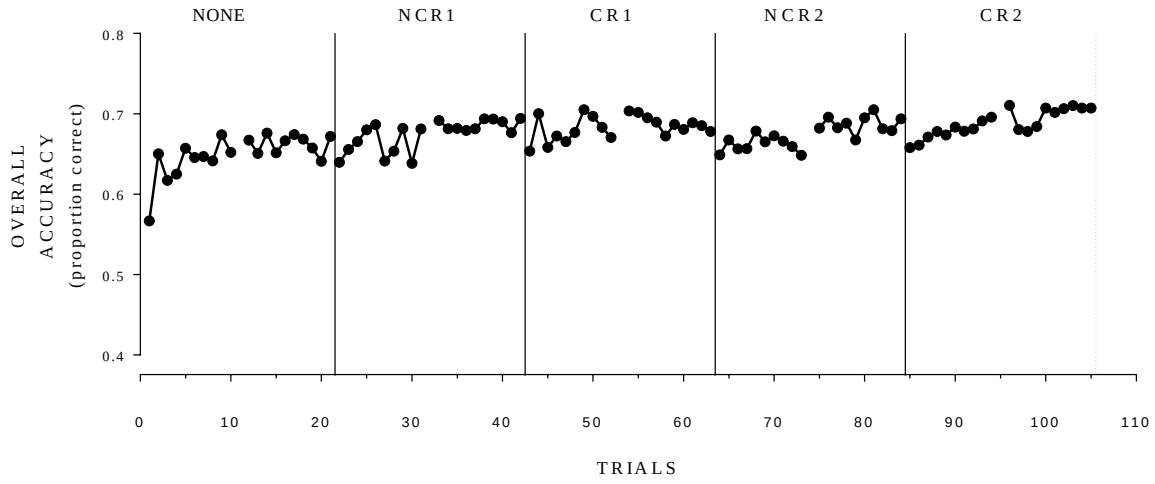

Figure 3. Average full word accuracy before and after backtrack (separated by the space) in each condition

.

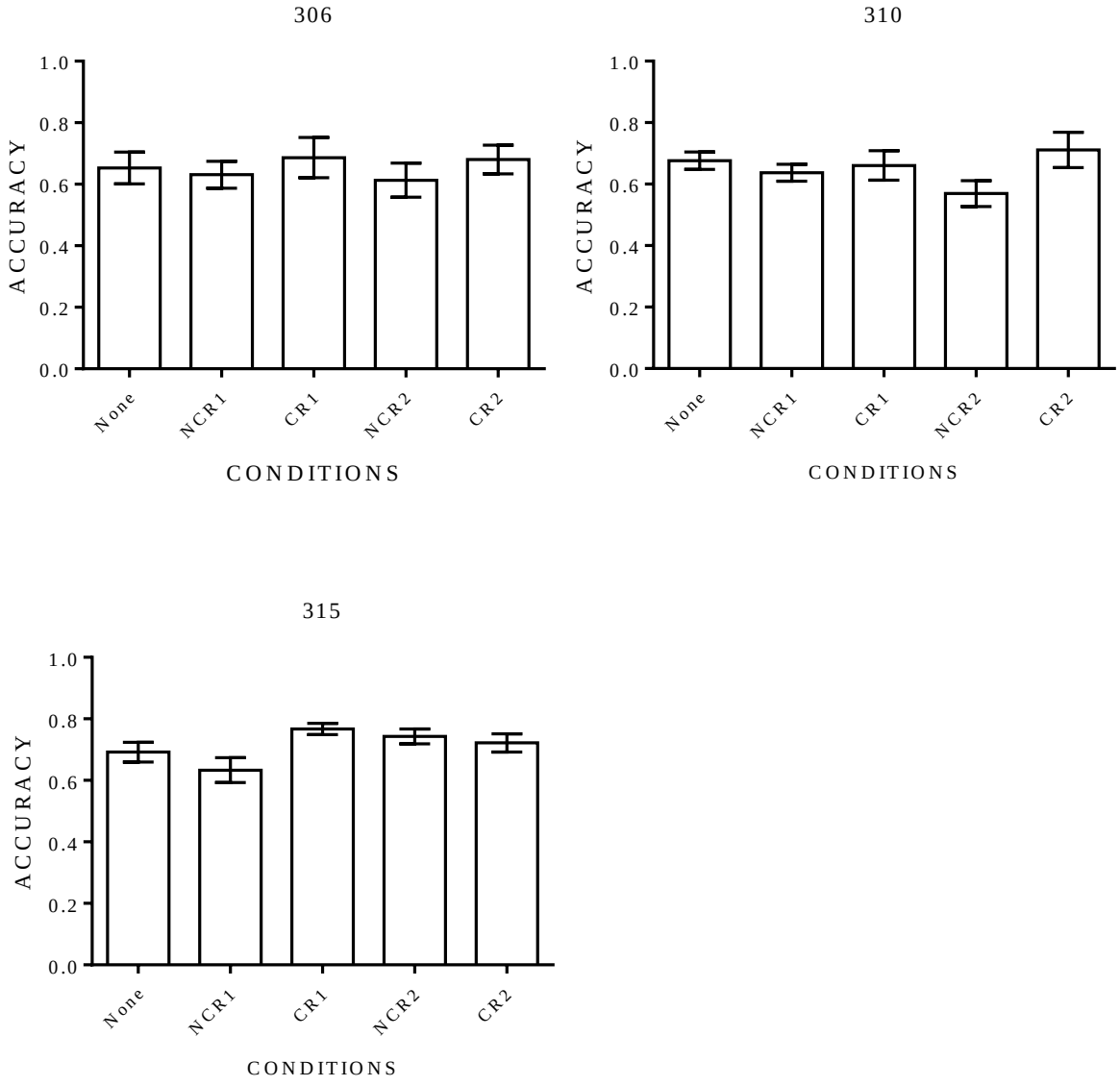

Figure 4. Average full word performance in each condition for participants that performed better in CR conditions than NCR. Standard deviations are represented by the error bars attached to each column.

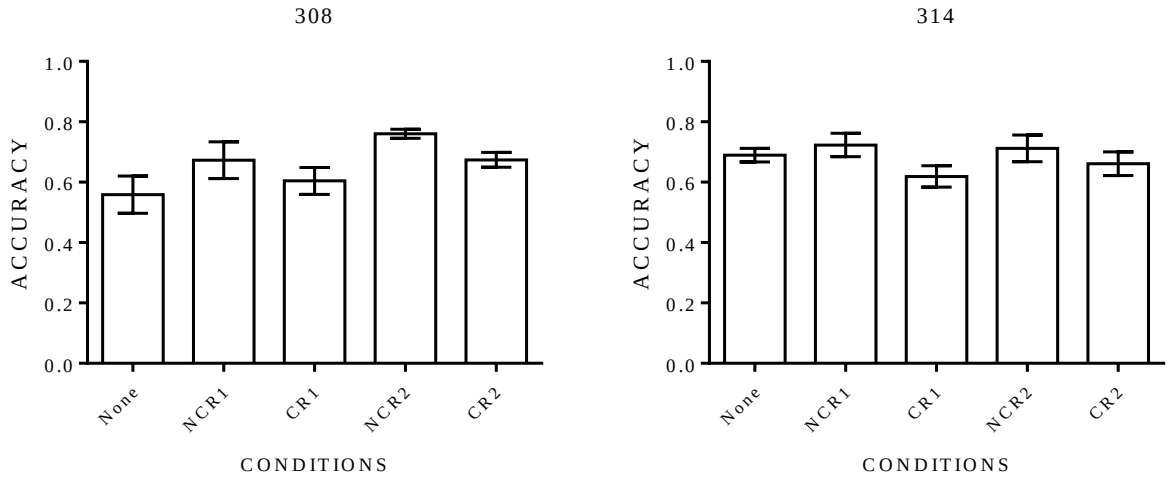

Figure 5. Average full word performance in each condition for participants that performed better in NCR conditions than CR. Standard deviations are represented by the error bars attached to each column.

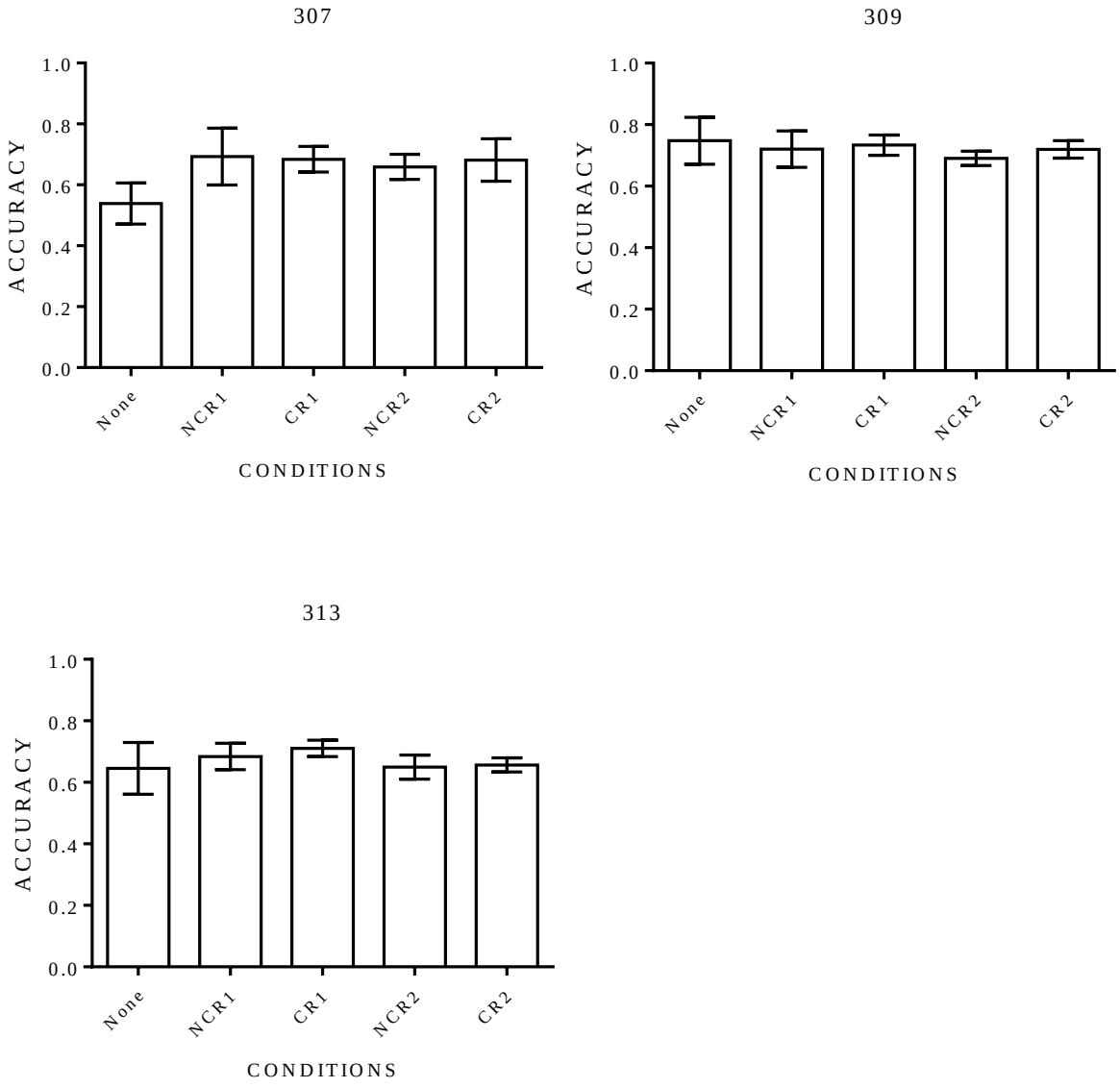

Figure 6. Average full word performance in each condition for participants that did not perform differently between CR and NCR conditions. Standard deviations are represented by the error bars attached to each column.

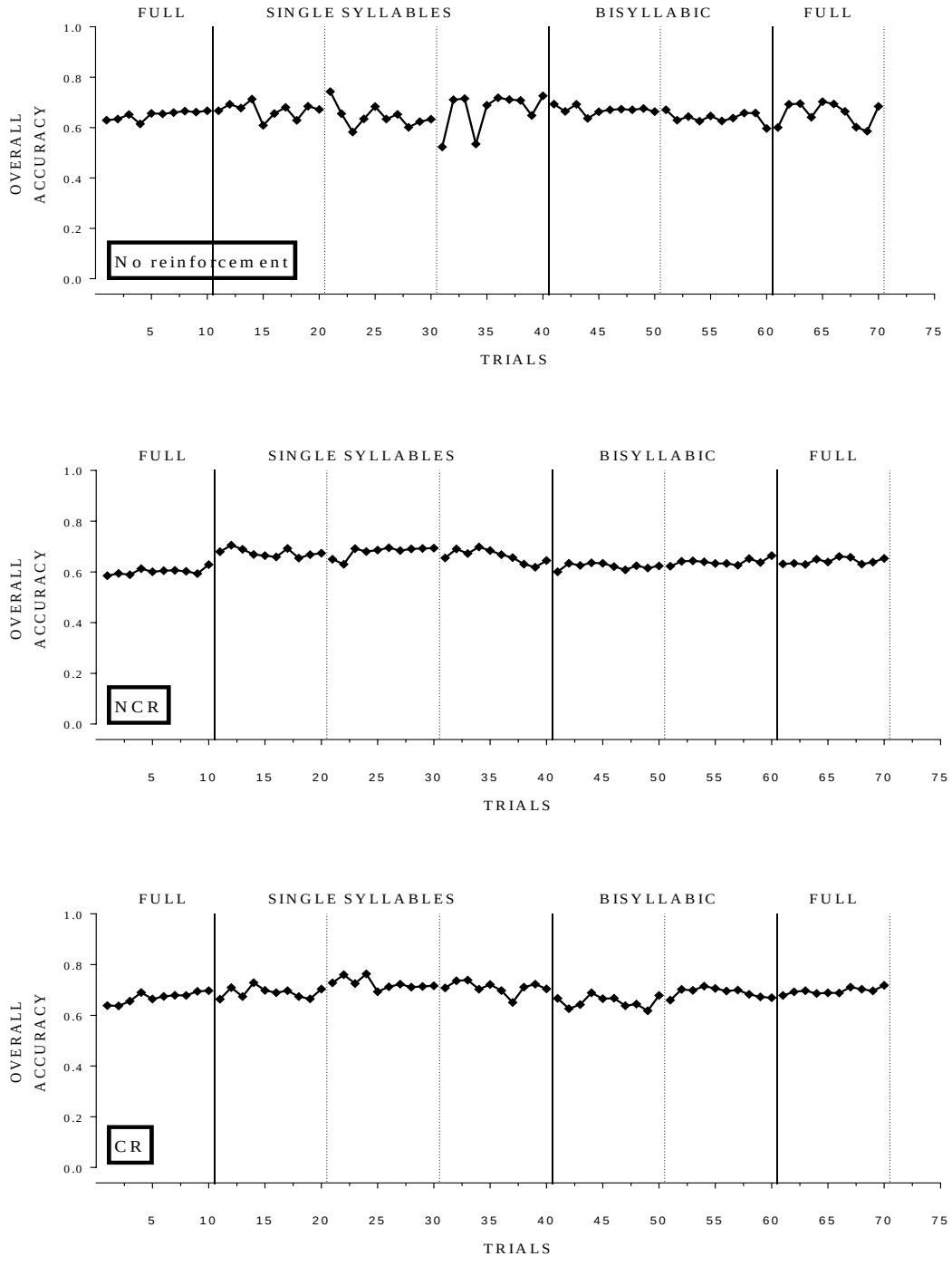

Figure 7. Averaged performance of 306 in each condition

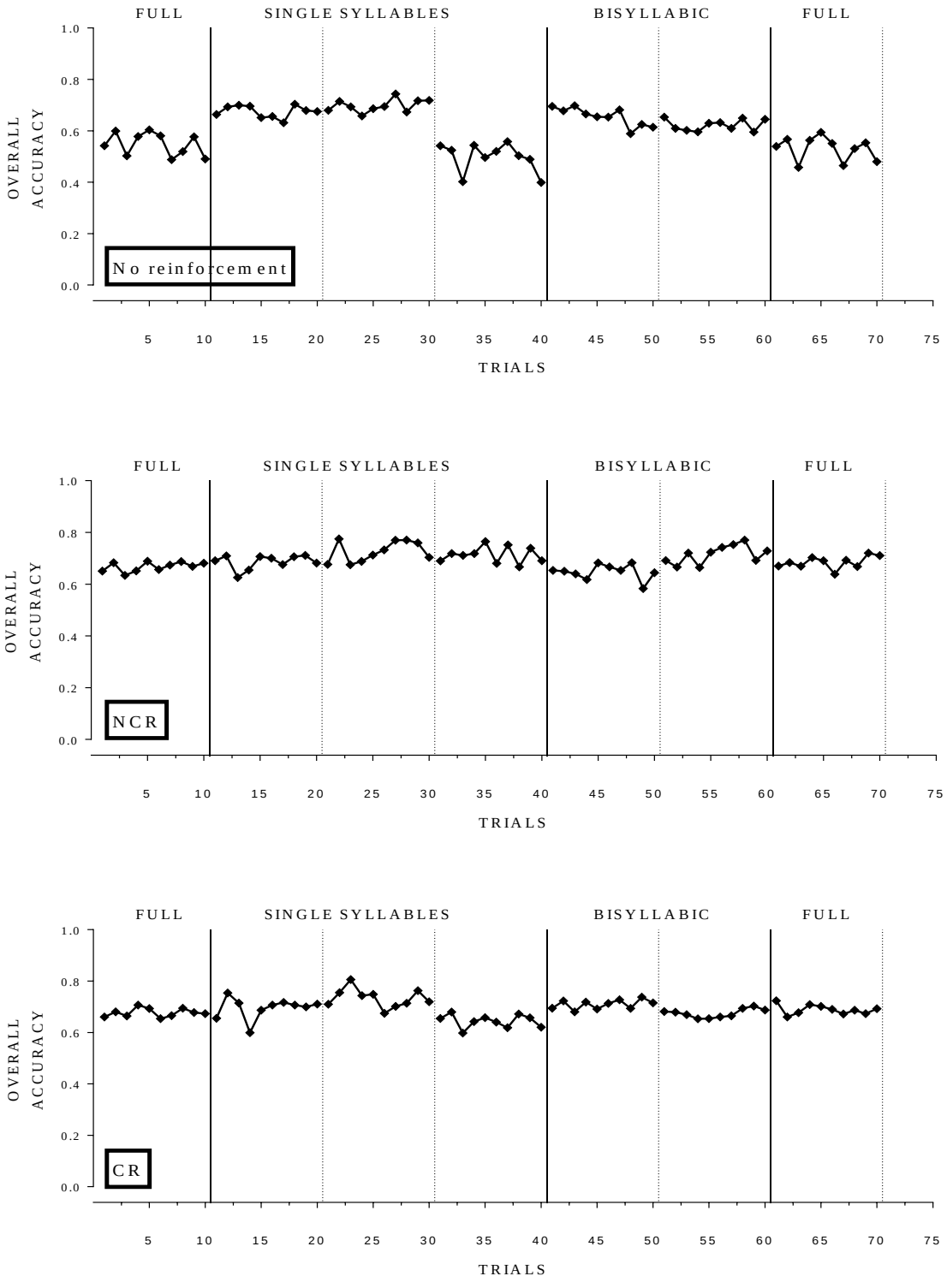

Figure 8. Averaged performance of 307 in each condition

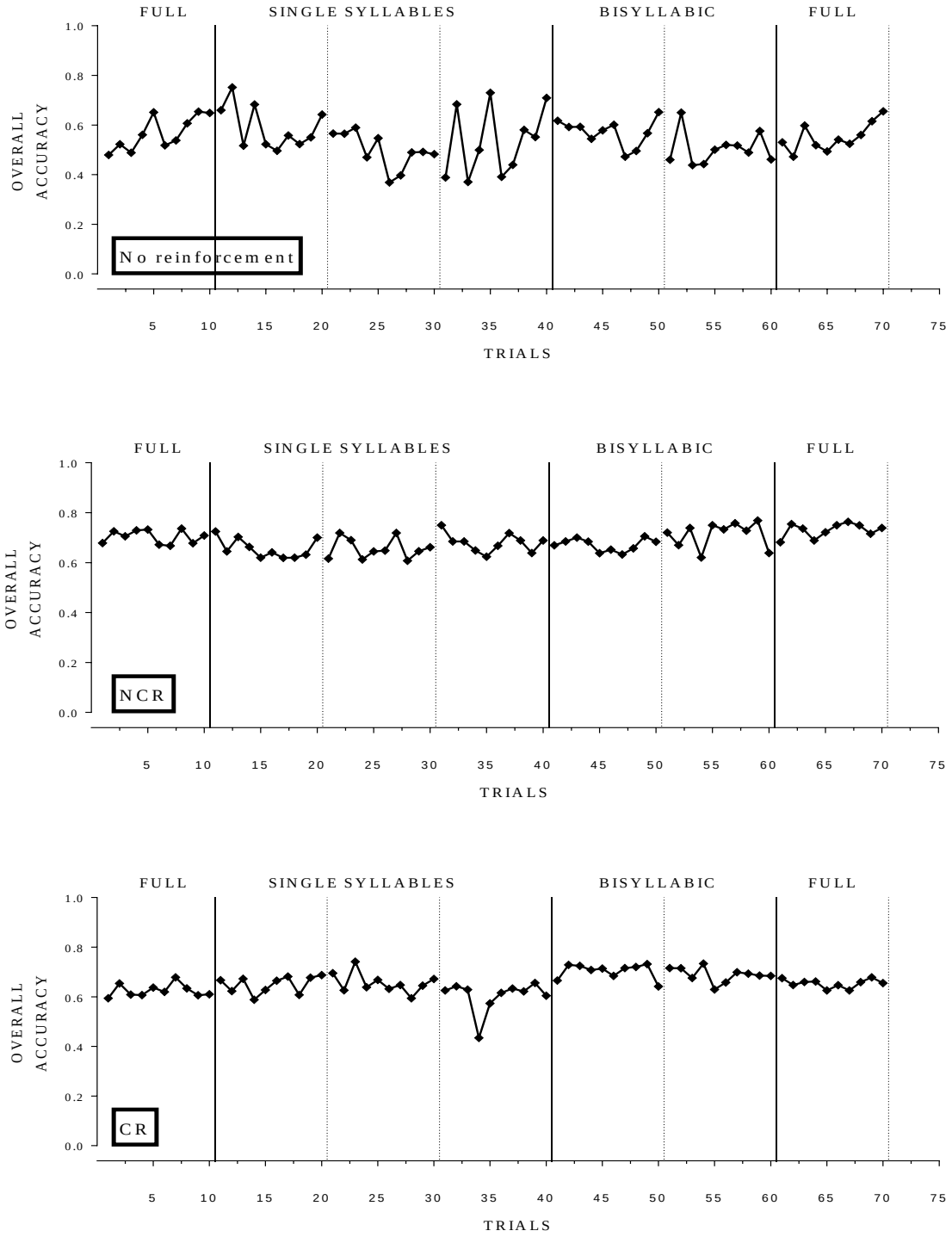

Figure 9. Averaged performance of 308 in each condition

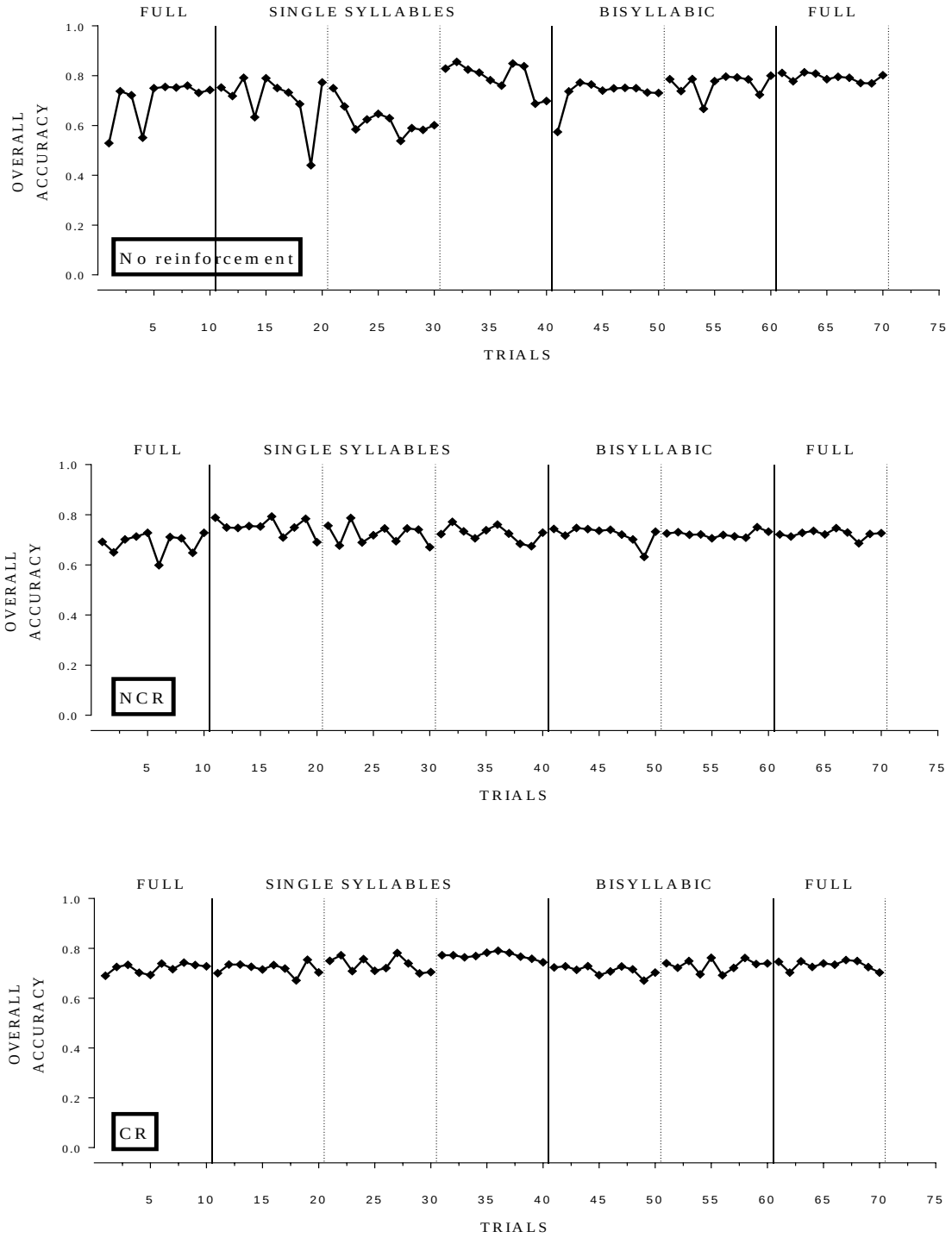

Figure 10. Averaged performance of 309 in each condition

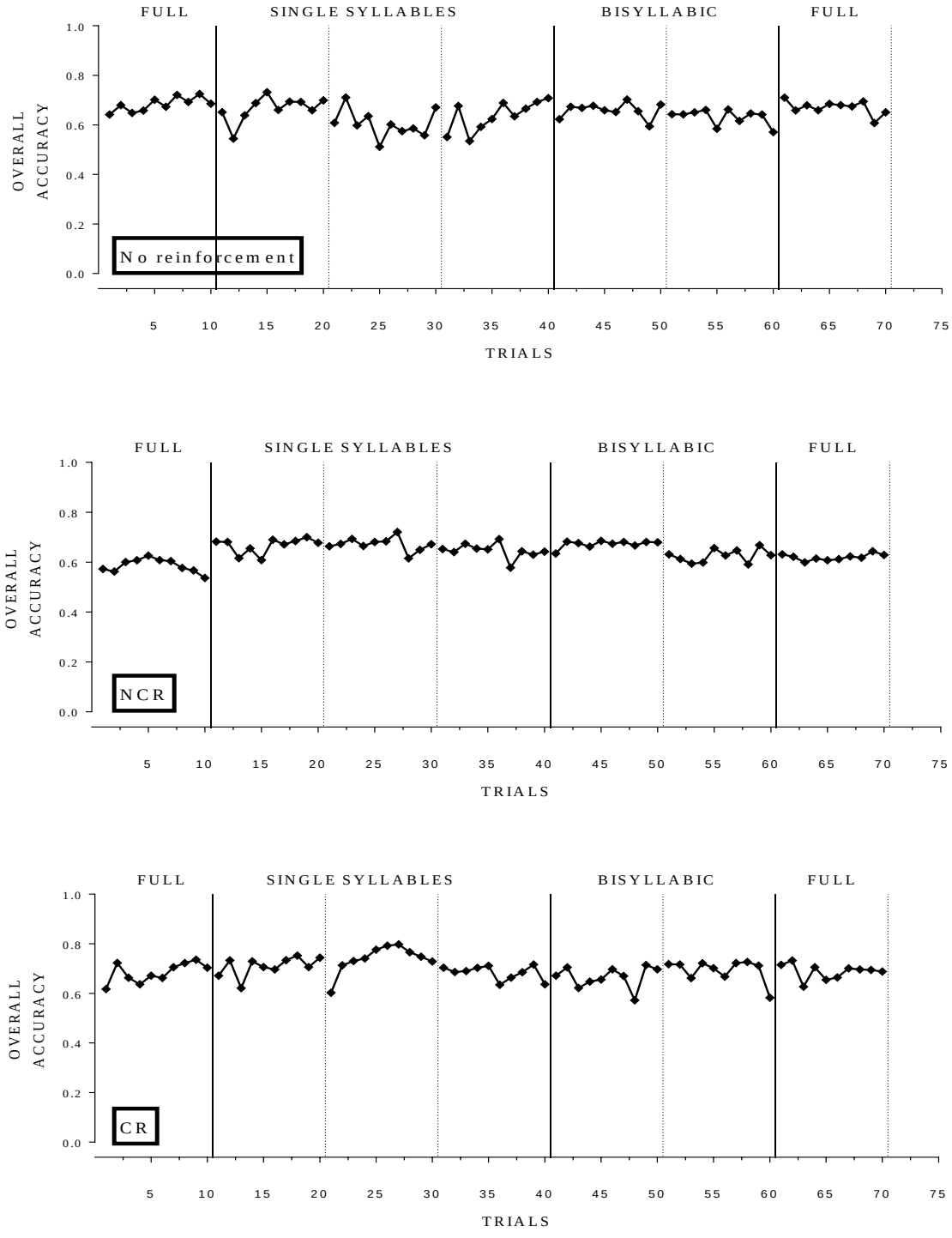

Figure 11. Averaged performance of 310 in each condition

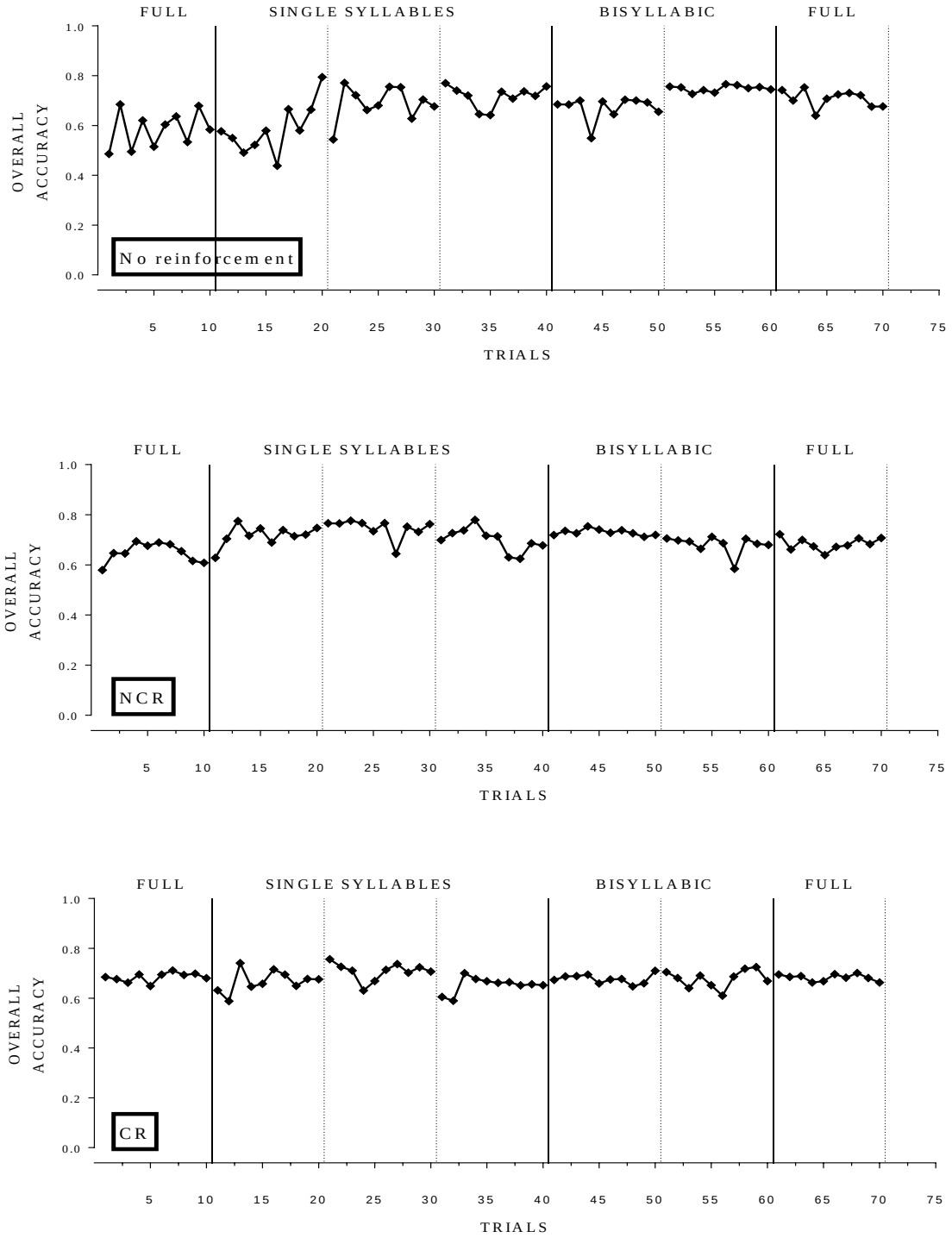

Figure 12. Averaged performance of 313 in each condition

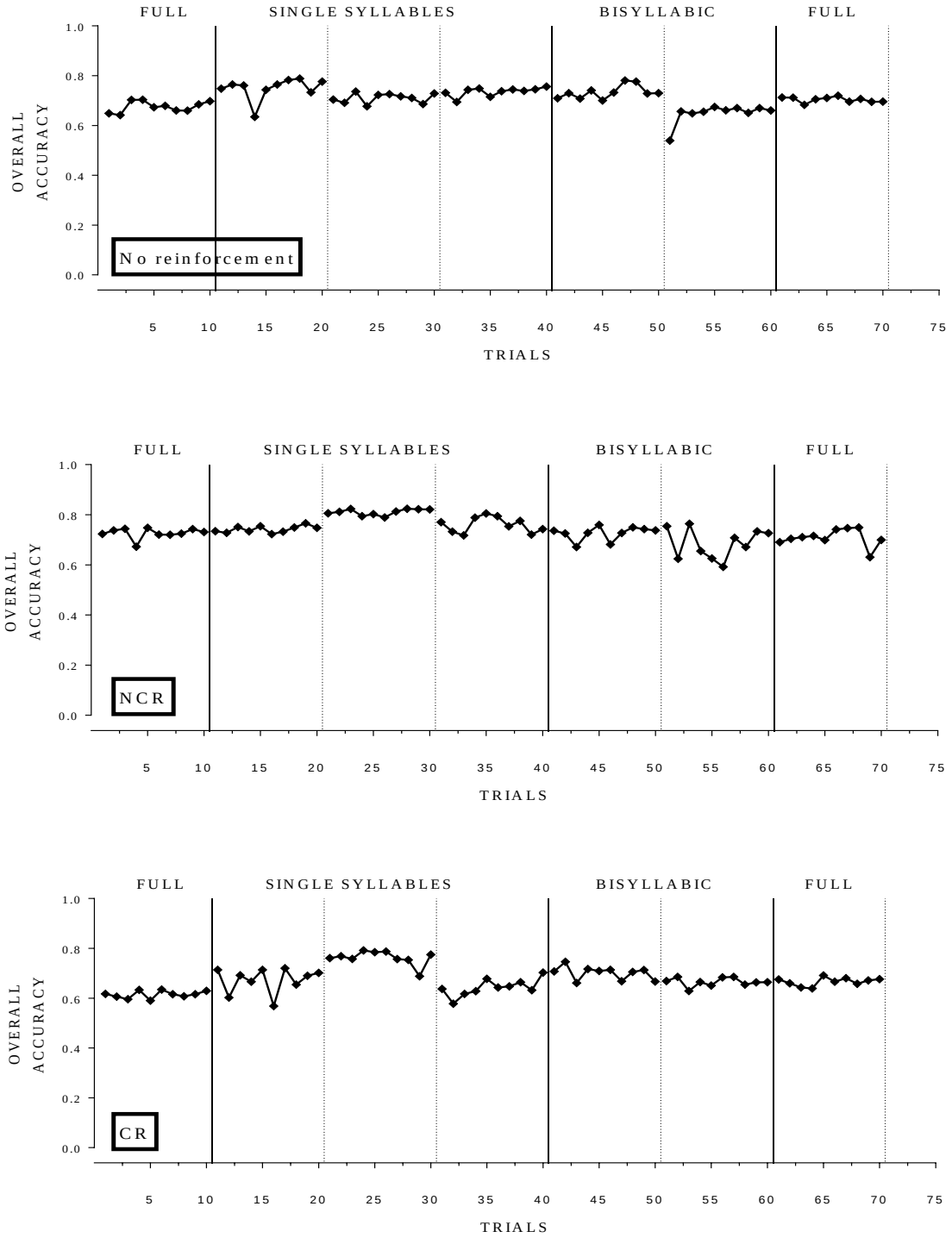

Figure 13. Averaged performance of 314 in each condition

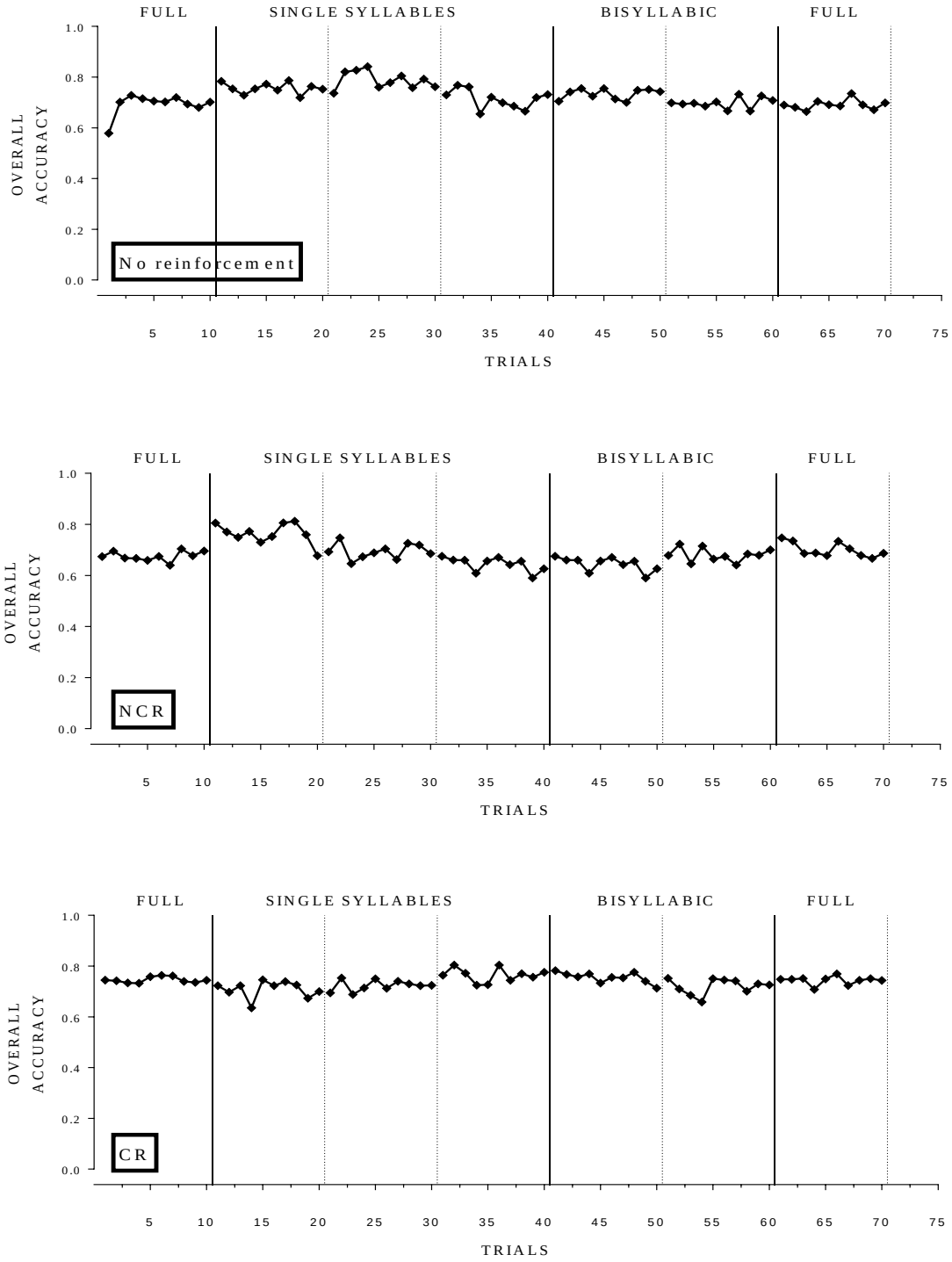

Figure 14. Averaged performance of 315 in each condition

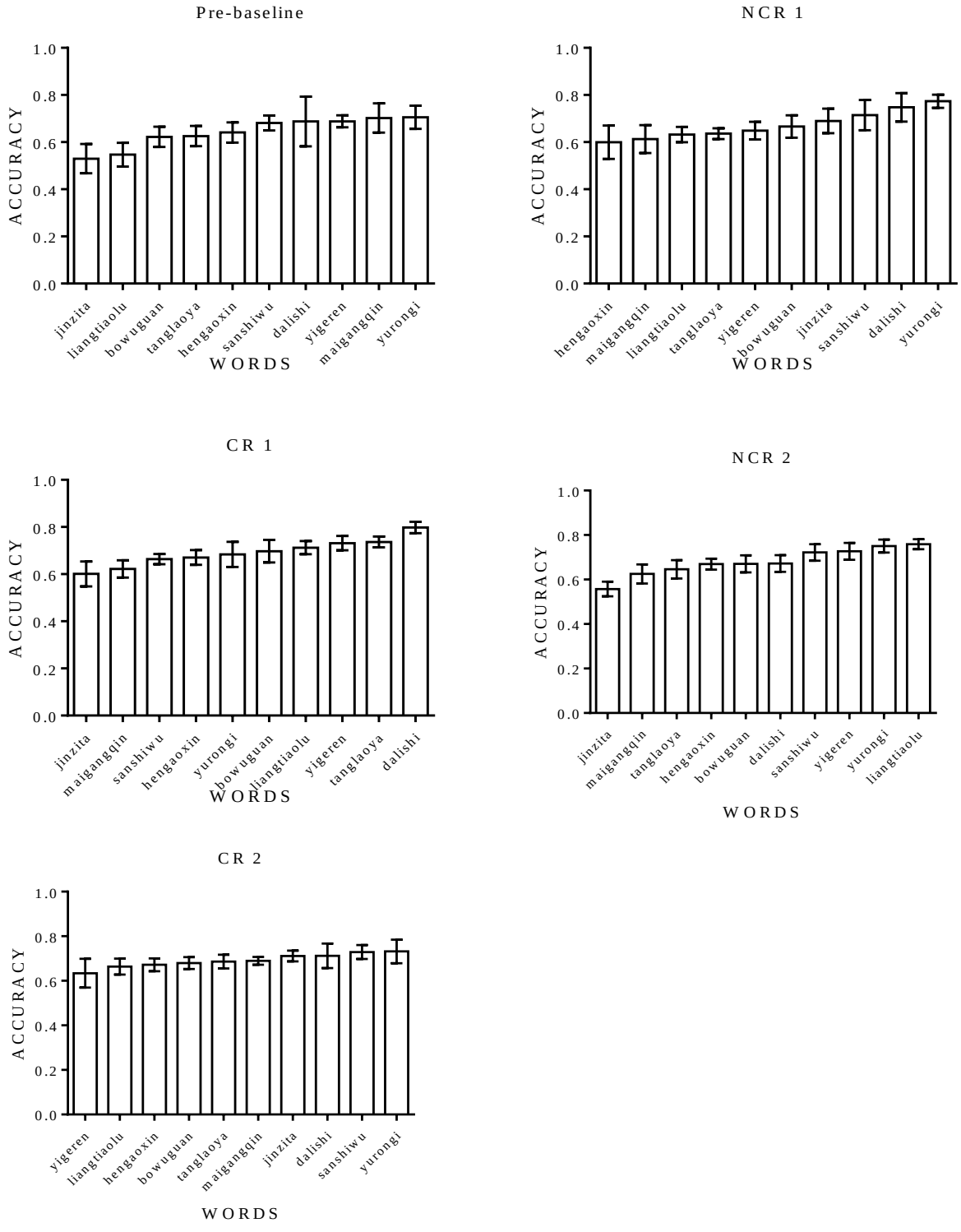

Figure 15. Average performance of words within each condition, from least to most accurate. Standard deviations are represented by the error bars attached to each column.

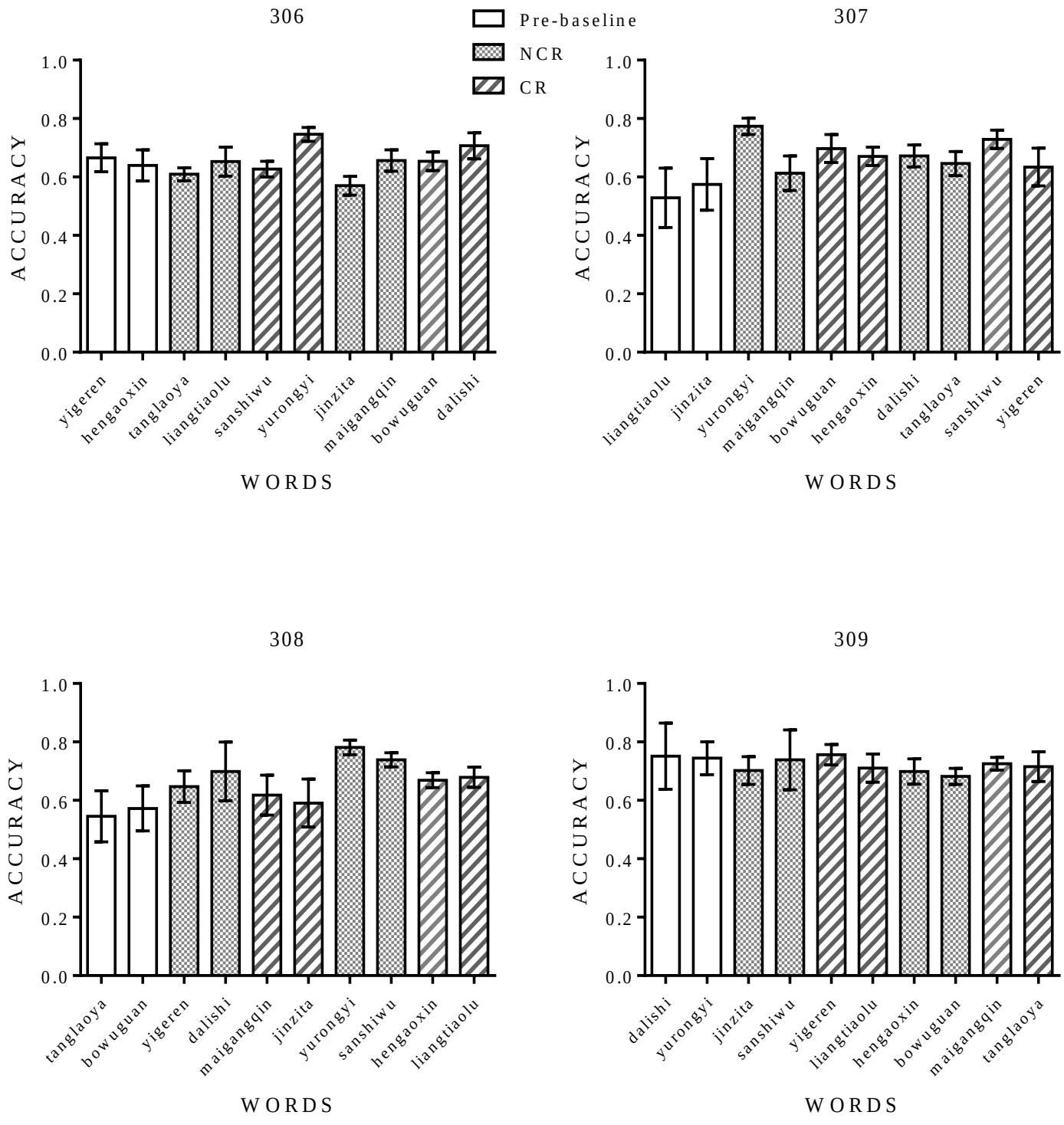

Figure 16. Individual participant performance for each word (participants 306 to 309). Standard deviations are represented by the error bars attached to each column.

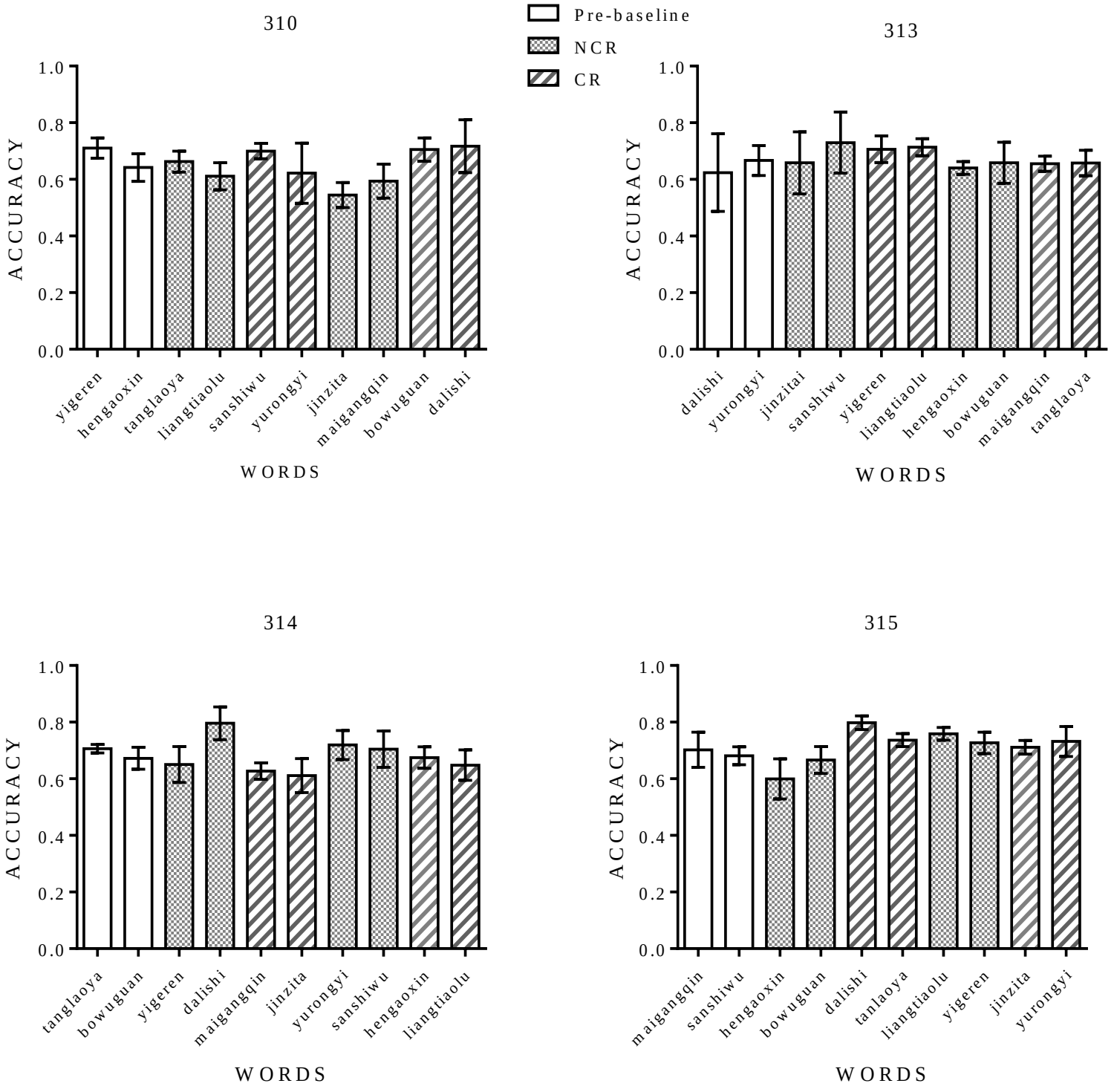

Figure 17. Individual participant performance for each word (participants 310 to 315). Standard deviations are represented by the error bars attached to each column.

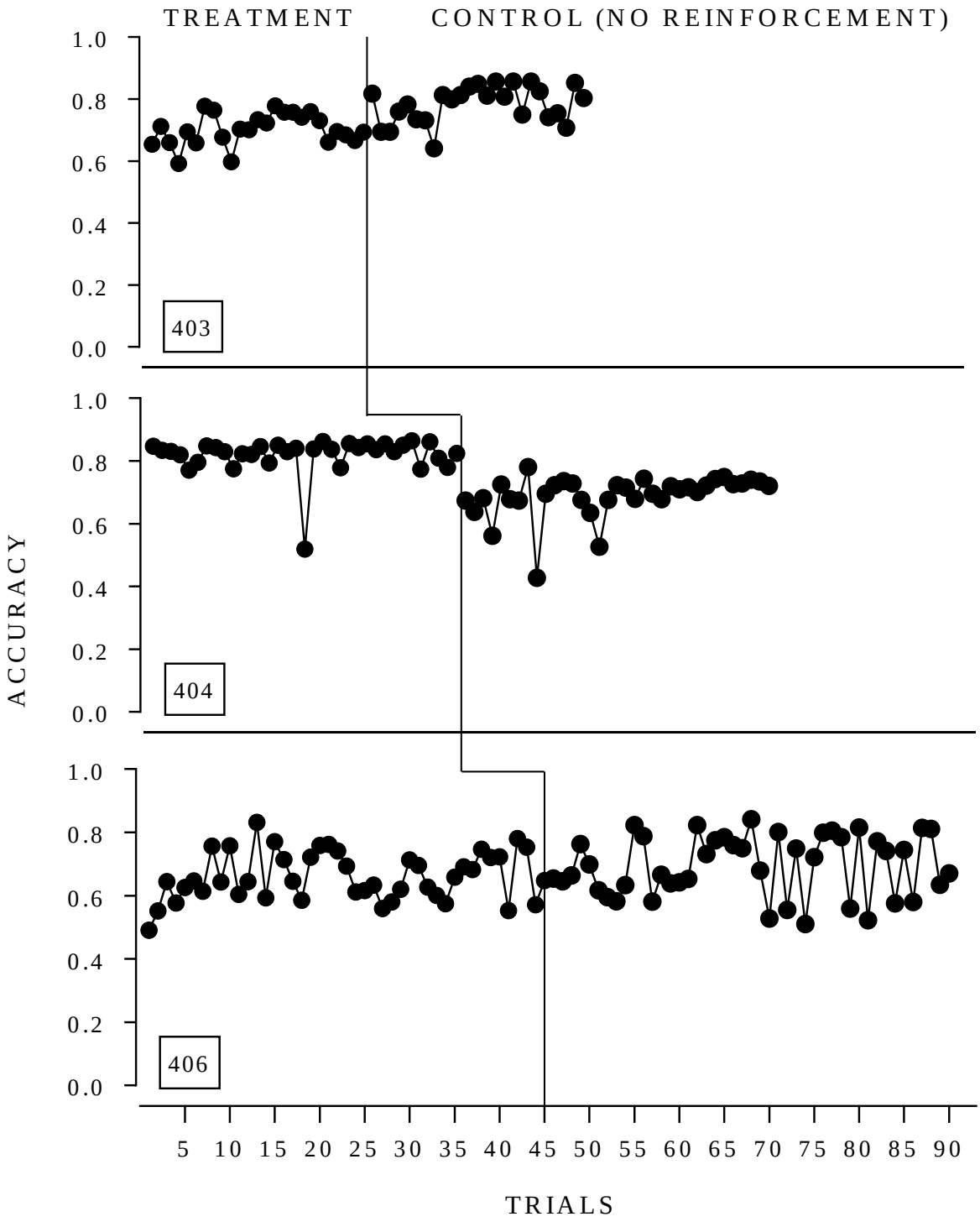

Figure 18. Multiple baseline follow-up study.

Appendix A

Language History Questionnaire[3]
(Version 2.0, 2012)

Language History Questionnaire (LHQ 2.0)

Participant ID: 0

1. Age:

2. Sex: Male Female

3. Education: Education...

4. Have you ever studied or learned a second language in terms of listening, speaking, reading, or writing? Yes No

Reset Next

Language History Questionnaire (LHQ 2.0)

5. Indicate your native language(s) and any other languages you have studied or learned, the age at which you started using each language in terms of listening, speaking, reading, and writing, and the total number of years you have spent using each language.

Language:	Listening:	Speaking:	Reading:	Writing:	Years of use:
Language...					year(s)
Language...					year(s)
Language...					year(s)
Language...					year(s)

6a. Country of residence: Country... 6b. Country of origin: Country...

6c. If 6a and 6b are different, then when did you first move to the country where you currently live?

7. If you have lived or traveled in countries other than your country of residence or country of origin for three or more months, then indicate the name of the country, your length of stay, the language you used, and the frequency of your use of the language for each country.

Country:	Length of stay:	Language:	Frequency of use:
Country...	month(s)	Language...	Rate...
Country...	month(s)	Language...	Rate...
Country...	month(s)	Language...	Rate...
Country...	month(s)	Language...	Rate...

8. Indicate the age at which you started using each of the languages you have studied or learned in the following environments.

Language:	At home:	With friends:	At school:	At work:	Language software:	Online games:
Language...						
Language...						
Language...						
Language...						

[3] Adapted from the online version of LHQ 2.0 at http://blclab.org/language-history-questionnaire. Reproduced with permission from Dr. Ping Li.

9. Indicate the language used by your teachers for instruction at each educational level. If the instructional language switched during any educational level, then also indicate the "Switched to" language.

	Language:	(Switched to:)
Elementary school:	Language...	Language...
Middle school:	Language...	Language...
High school:	Language...	Language...
College/university:	Language...	Language...

10. Rate your language learning skill. In other words, how good do you feel you are at learning new languages, relative to your friends or other people you know? Rate...

11. Rate your current ability in terms of listening, speaking, reading, and writing in each of the languages you have studied or learned.

Language:	Listening:	Speaking:	Reading:	Writing:
Language...	Rate...	Rate...	Rate...	Rate...
Language...	Rate...	Rate...	Rate...	Rate...
Language...	Rate...	Rate...	Rate...	Rate...
Language...	Rate...	Rate...	Rate...	Rate...

12. If you have taken any standardized language proficiency tests (e.g., TOEFL), then indicate the name of the test, the language assessed, and the score you received for each. If you do not remember the exact score, then indicate an "Approximate score" instead.

Test:	Language:	Score:	(Approximate score:)
	Language...		
	Language...		
	Language...		
	Language...		

13. Rate the strength of your foreign accent for each of the languages you have studied or learned.

Language:	Accent:
Language...	Rate...
Language...	Rate...
Language...	Rate...
Language...	Rate...

14. Estimate how many hours per day you spend engaged in the following activities in each of the languages you have studied or learned.

Language:	Watching television:	Listening to radio:	Reading for fun:	Reading for school/work:	Writing emails to friends:	Writing for school/work:
Language...						
Language...						
Language...						
Language...						

15. Estimate how many hours per day you spend speaking with the following groups of people in each of the languages you have studied or learned.

Language:	Family members:	Friends:	Classmates:	Coworkers:
Language...				
Language...				
Language...				
Language...				

16a. Do you mix words or sentences from different languages when you speak? (●) Yes ( ) No

16b. If you answered "Yes" to 16a, then indicate the languages that you mix and estimate the frequency of mixing in normal conversation with the following groups of people.

	Language 1:	Language 2:	Frequency of mixing:
Family members:	Language...	Language...	Rate...
Friends:	Language...	Language...	Rate...
Classmates:	Language...	Language...	Rate...
Coworkers:	Language...	Language...	Rate...

17. In which language do you communicate best or feel most comfortable in terms of listening, speaking, reading, and writing in each of the following environments?

	Listening:	Speaking:	Reading:	Writing:
At home:	Language...	Language...	Language...	Language...
With friends:	Language...	Language...	Language...	Language...
At school:	Language...	Language...	Language...	Language...
At work:	Language...	Language...	Language...	Language...

18. How often do you use each of the languages you have studied or learned for the following activities?

Language:	Thinking:	Talking to yourself:	Expressing emotion:	Dreaming:	Arithmetic:	Remembering numbers:
Language...	Rate...	Rate...	Rate...	Rate...	Rate...	Rate...
Language...	Rate...	Rate...	Rate...	Rate...	Rate...	Rate...
Language...	Rate...	Rate...	Rate...	Rate...	Rate...	Rate...
Language...	Rate...	Rate...	Rate...	Rate...	Rate...	Rate...

19. What percentage of your friends speaks each of the languages you have studied or learned?

Language:	Percentage:
Language...	%
Language...	%
Language...	%
Language...	%

20a. Do you feel that you are bicultural or multicultural? (●) Yes ( ) No

20b. If you answered "Yes" to 20a, then which cultures/languages do you identify with more strongly? Rate the strength of your connection in the following categories for each culture/language.

Culture/language:	Way of life:	Food:	Music:	Art:	Cities/towns:	Sports teams:
Language...	Rate...	Rate...	Rate...	Rate...	Rate...	Rate...
Language...	Rate...	Rate...	Rate...	Rate...	Rate...	Rate...
Language...	Rate...	Rate...	Rate...	Rate...	Rate...	Rate...
Language...	Rate...	Rate...	Rate...	Rate...	Rate...	Rate...

21. Use the comment box below to indicate any additional answers to any of the questions above that you feel better describe your language background or usage.

22. Use the comment box below to indicate anything that you feel is interesting or important about your language background or usage and was not covered by the questions above.

Reset Submit

Appendix B

List of Stimuli and Tones

Word	Pronunciation	First Tone	Middle Tone	Last Tone
一个人	yi ge ren	1	4	2
很高兴	hen gao xing	3	1	4
唐老鸭	tang lao ya	2	3	1
两条路	liang tiao lu	3	2	4
三十五	san shi wu	1	2	3
羽绒衣	yu rong yi	3	2	1
金字塔	jin zi ta	1	4	3
买钢琴	mai gang qin	4	1	2
博物馆	bo wu guan	2	4	3
大理石	da li shi	4	3	2

Appendix C

Counterbalancing Word Order

Group	Pre-baseline	NCR1	CR1	NCR2	CR2
1	一个人 (yi ge ren)	唐老鸭 (tang lao ya)	三十五 (san shi wu)	金字塔 (jin zi ta)	博物馆 (bo wu guan)
	很高兴 (hen gao xing)	**两条路** (liang tiao lu)	**羽**绒衣 (yu rong yi)	买钢琴 (mai gang qin)	**大理石** (da li shi)
2	买钢琴 (mai gang qin)	很高兴 (hen gao xing)	大理石 (da li shi)	两条路 (liang tiao lu)	金字塔 (jin zi ta)
	三十五 (san shi wu)	**博物**馆 (bo wu guan)	**唐老**鸭 (tang lao ya)	**一个人** (yi ge ren)	**羽**绒衣 (yu rong yi)
3	大理石 (da li shi)	金字塔 (jin zi ta)	一个人 (yi ge ren)	很高兴 (hen gao xing)	买钢琴 (mai gang qin)
	羽绒衣 (yu rong yi)	**三十五** (san shi wu)	**两条路** (liang tiao lu)	**博物**馆 (bo wu guan)	**唐老**鸭 (tang lao ya)
4	唐老鸭 (tang lao ya)	一个人 (yi ge ren)	买钢琴 (mai gang qin)	羽绒衣 (yu rong yi)	很高兴 (hen gao xing)
	博物馆 (bo wu guan)	**大理石** (da li shi)	**金字塔** (jin zi ta)	**三十五** (san shi wu)	**两条路** (liang tiao lu)
5	两条路 (liang tiao lu)	羽绒衣 (yu rong yi)	博物馆 (bo wu guan)	大理石 (da li shi)	三十五 (san shi wu)
	金字塔 (jin zi ta)	买钢琴 (mai gang qin)	**很高**兴 (hen gao xing)	**唐老**鸭 (tang lao ya)	**一个人** (yi ge ren)

Appendix D

Supplementary participant graphs

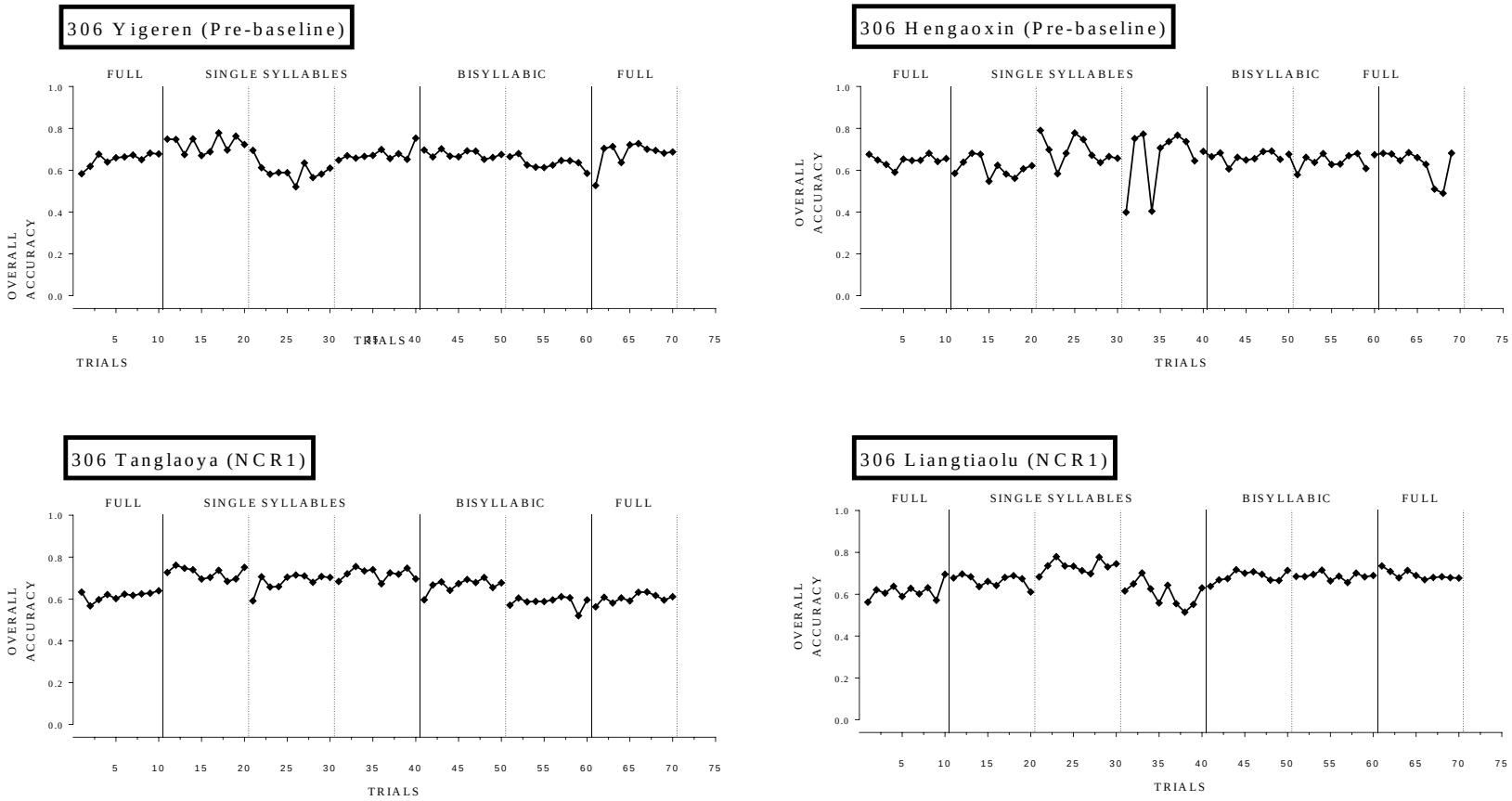

Figure D1. Performance of 306 across all phases of each word (Pre-baseline and NCR1)

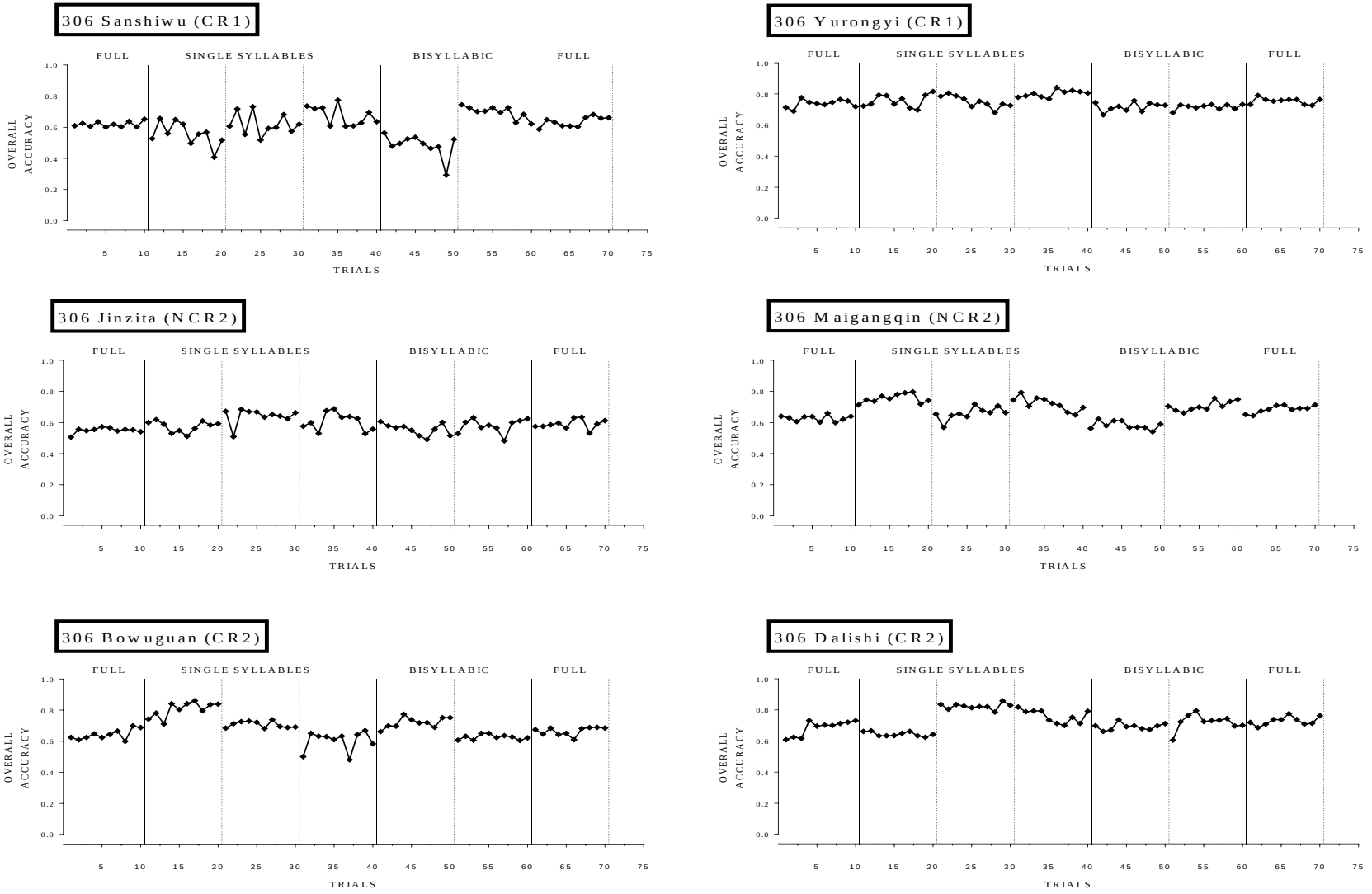

Figure D2. Performance of 306 across all phases of each word (CR1 to CR2)

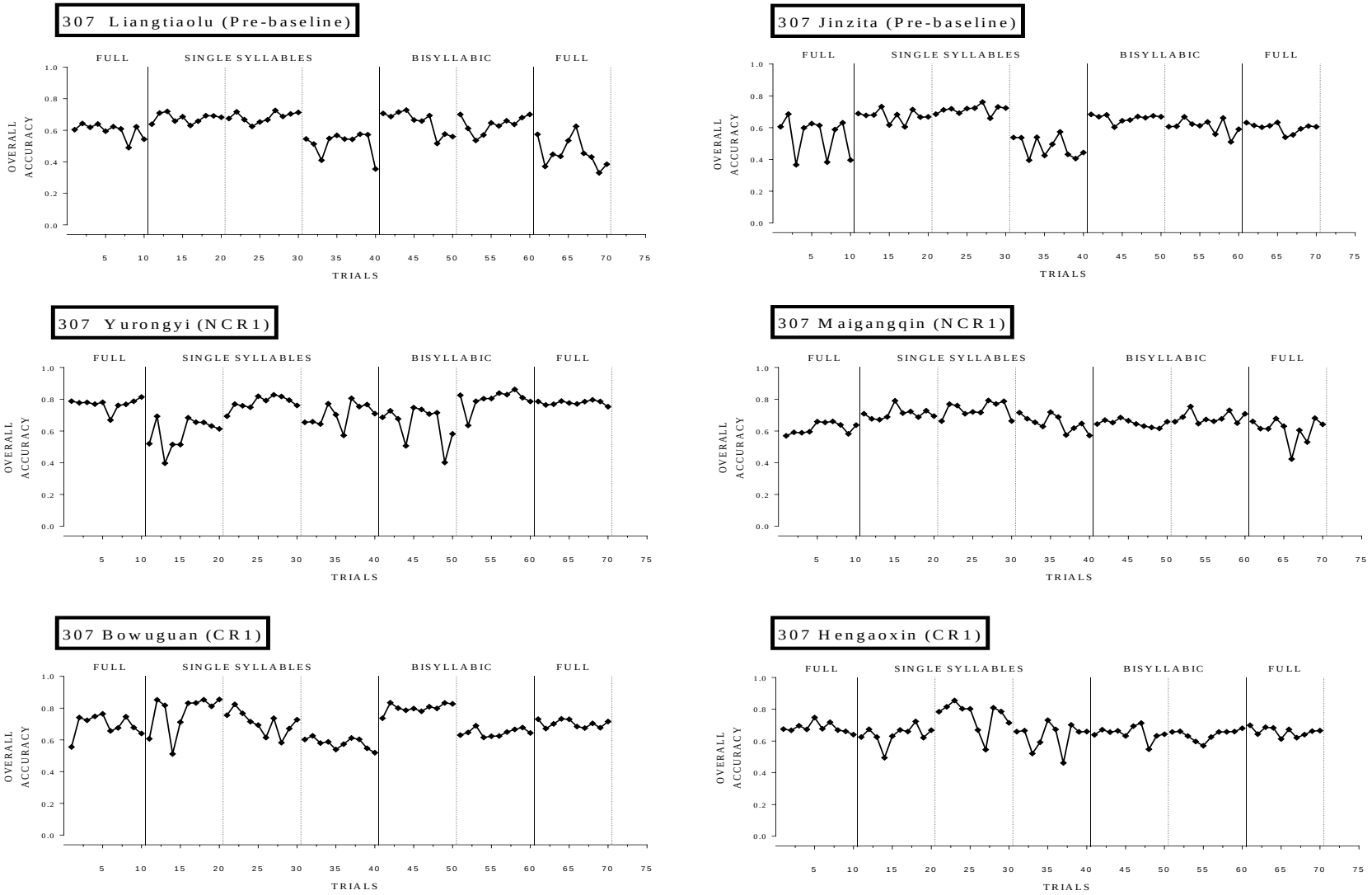

Figure D3. Performance of 307 across all phases of each word (Pre-baseline to CR1)

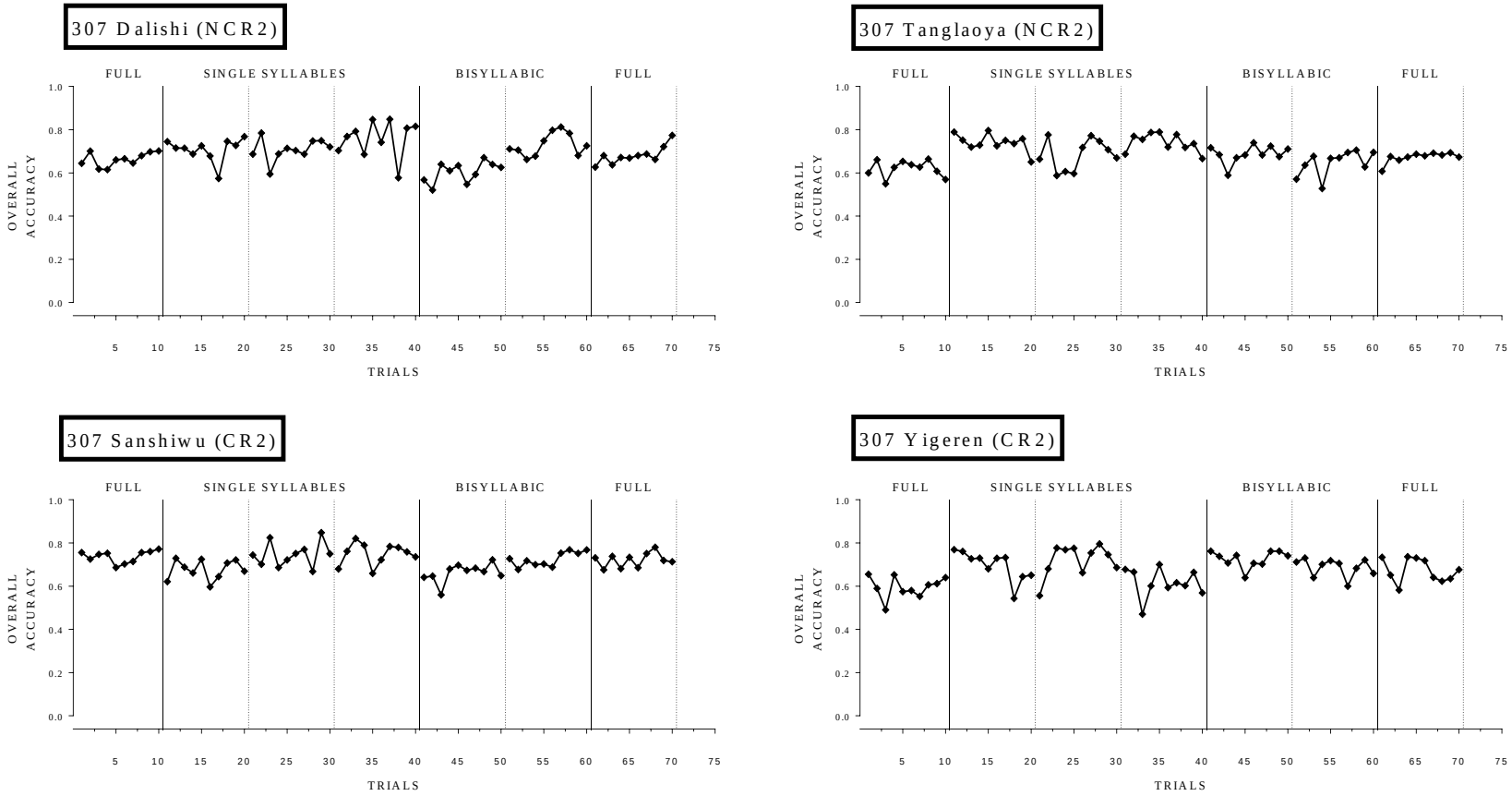

Figure D4. Performance of 307 across all phases of each word (NCR2 and CR2)

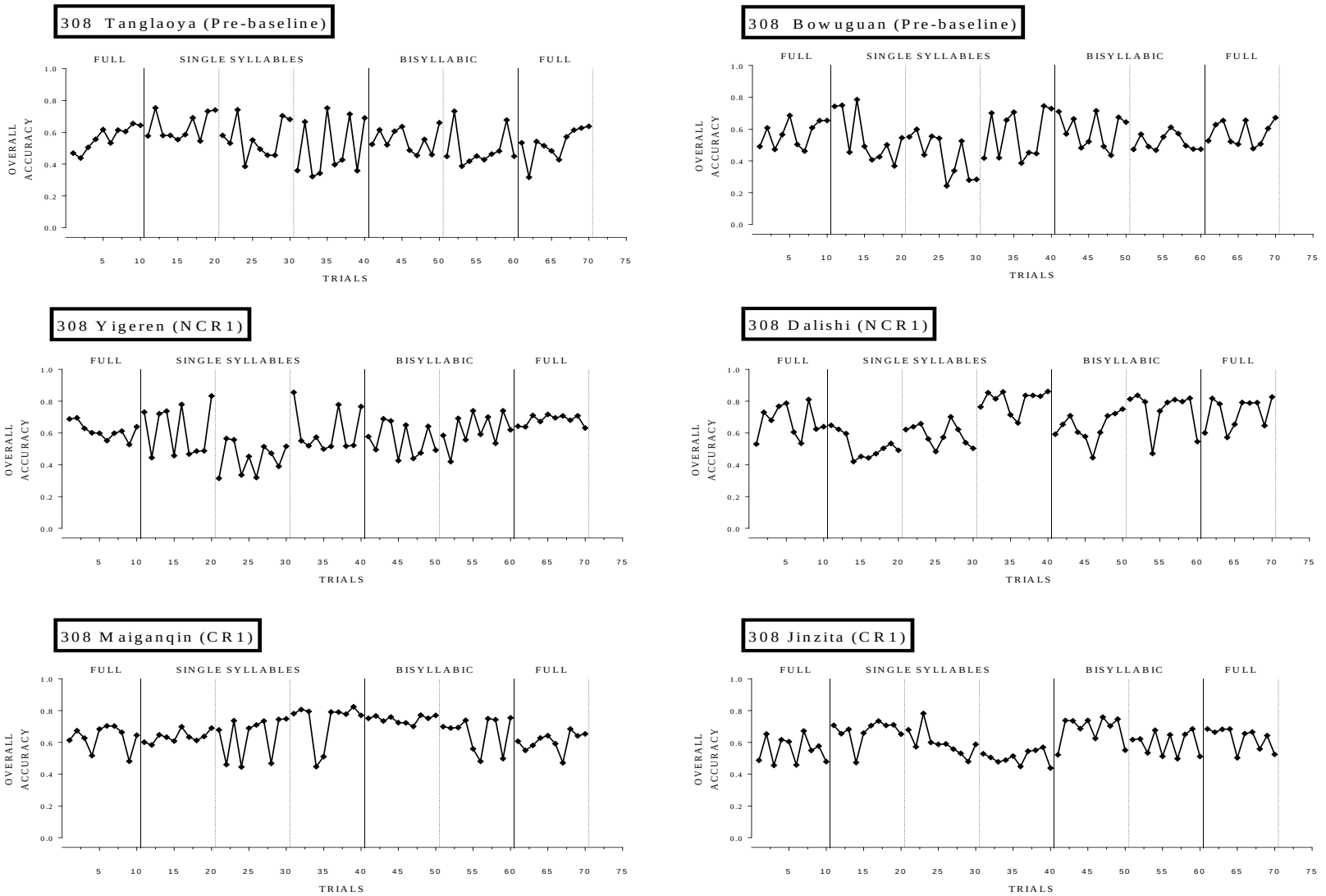

Figure D4. Performance of 308 across all phases of each word (Pre-baseline to CR1)

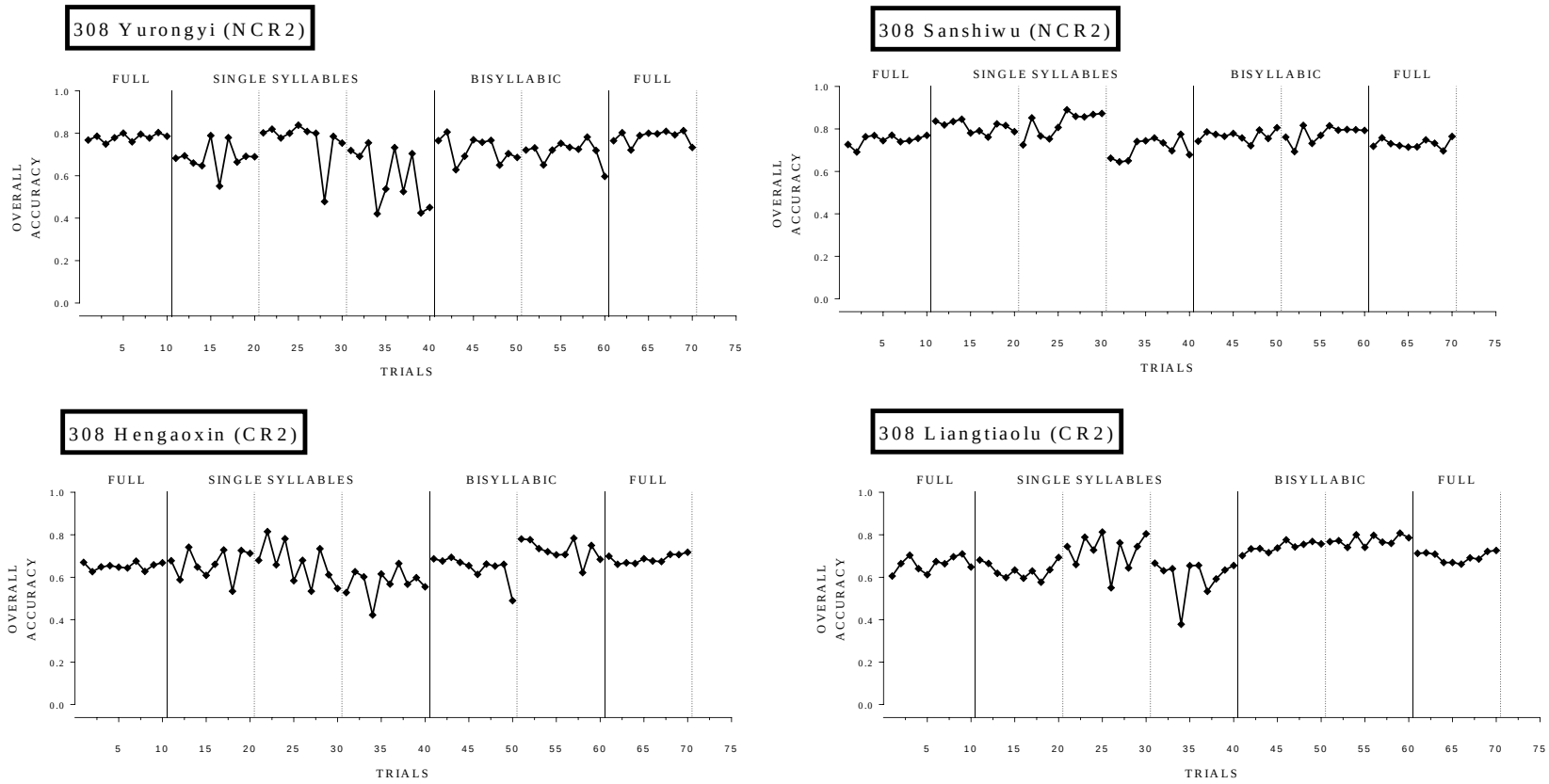

Figure D5. Performance of 308 across all phases of each word (NCR2 and CR2)

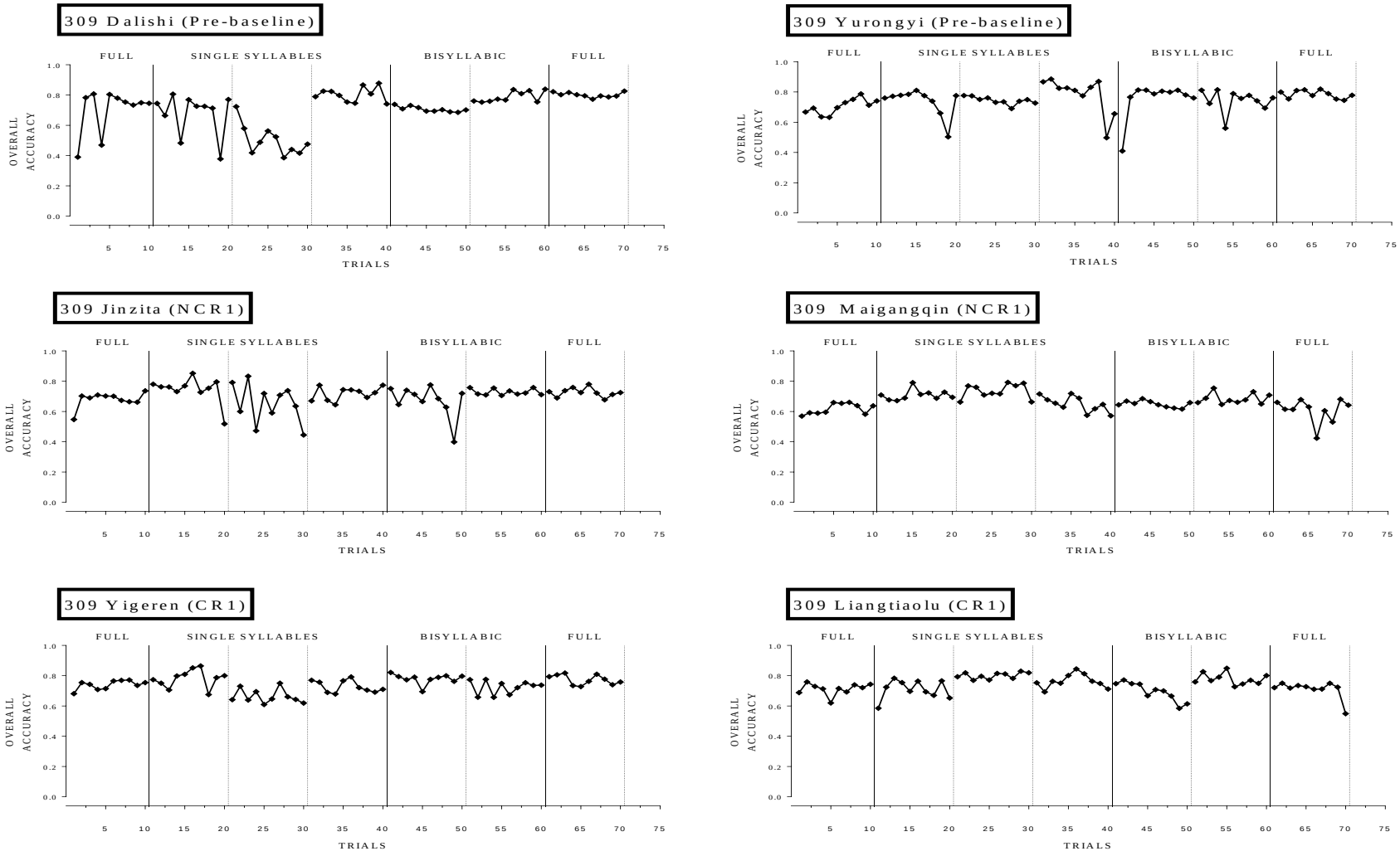

Figure D6. Performance of 309 across all phases of each word (Pre-baseline to CR1)

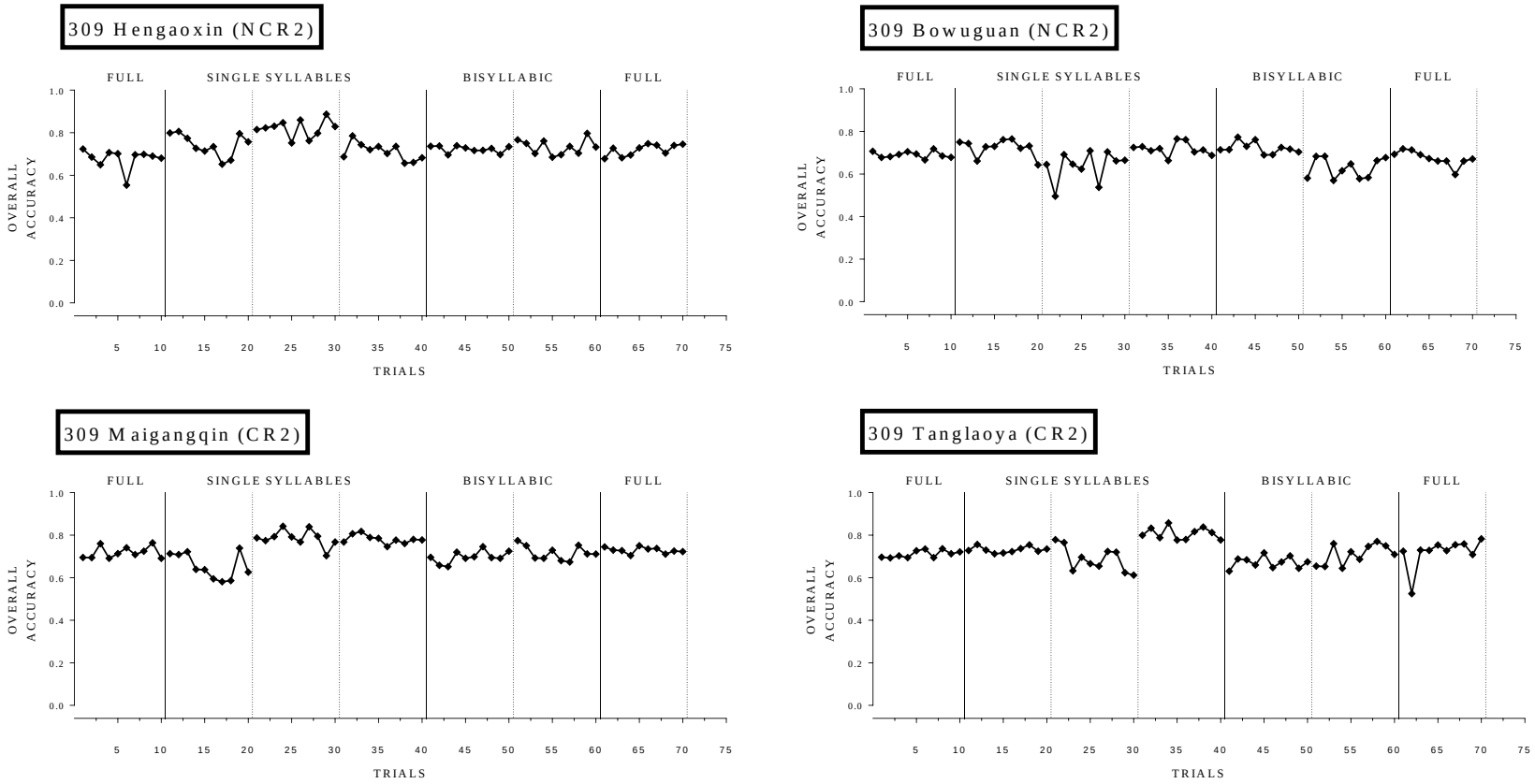

Figure D7. Performance of 309 across all phases of each word (NCR2 and CR2)

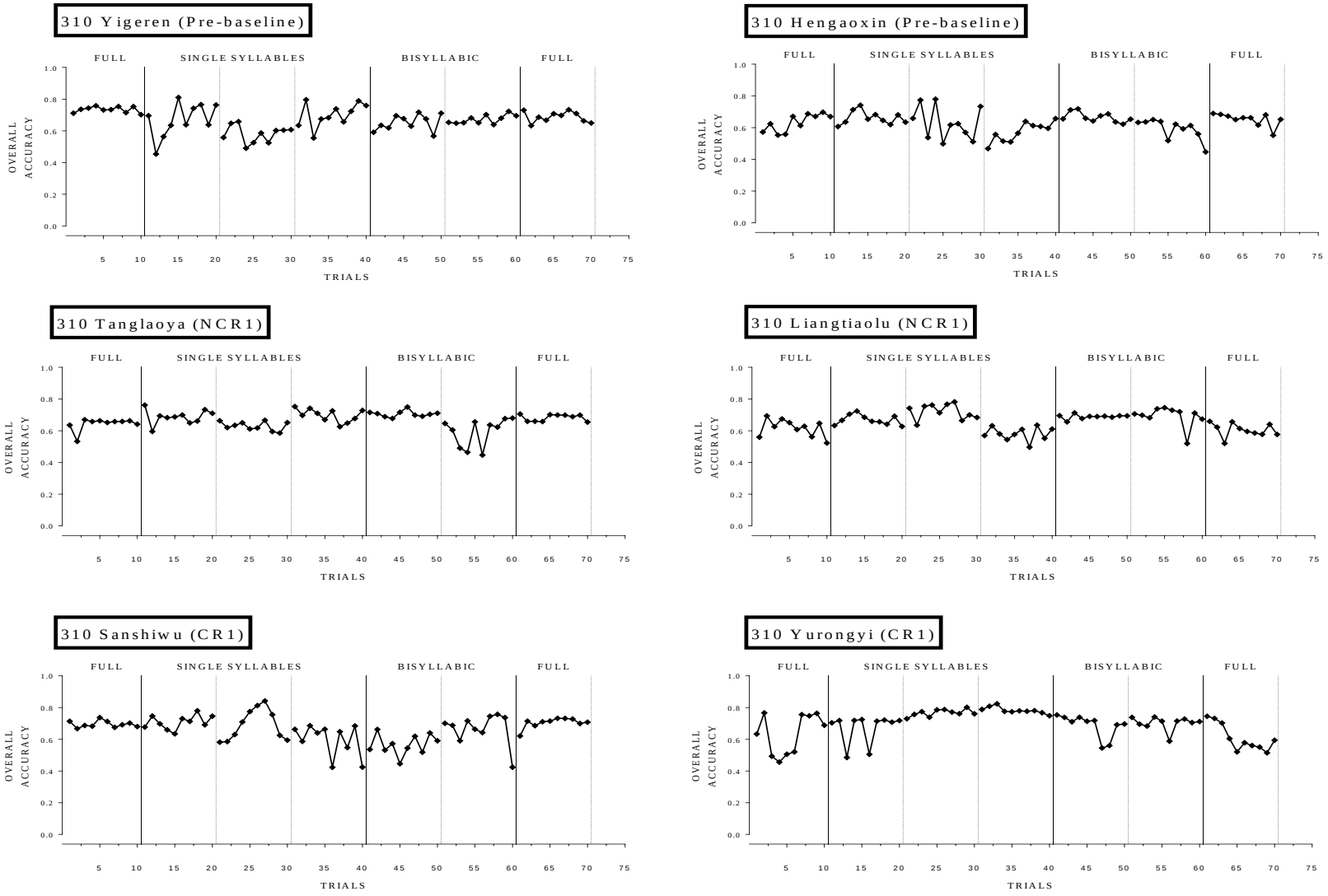

Figure D8. Performance of 310 across all phases of each word (Pre-baseline to CR1)

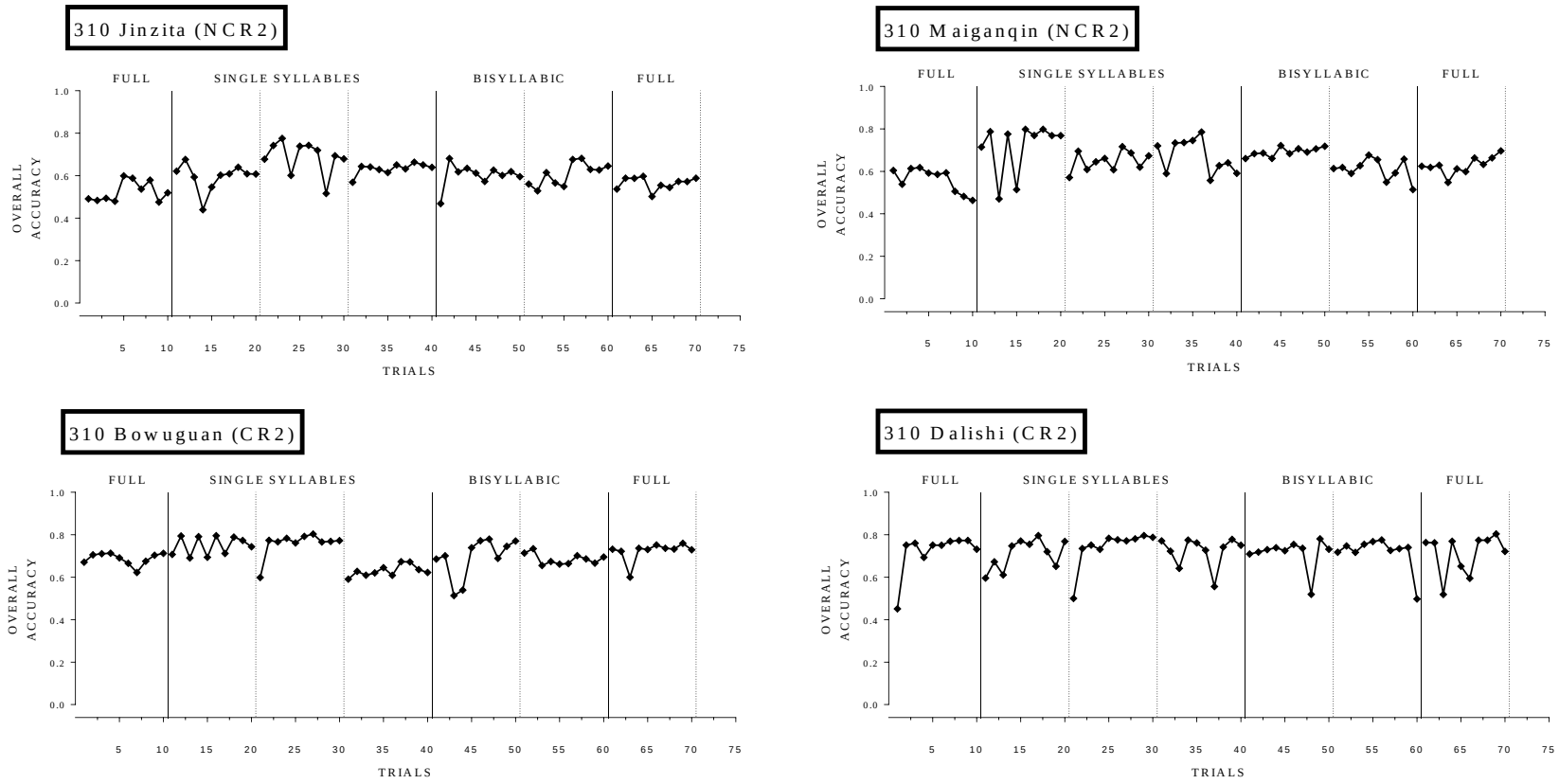

Figure D9. Performance of 310 across all phases of each word (NCR2 and CR2)

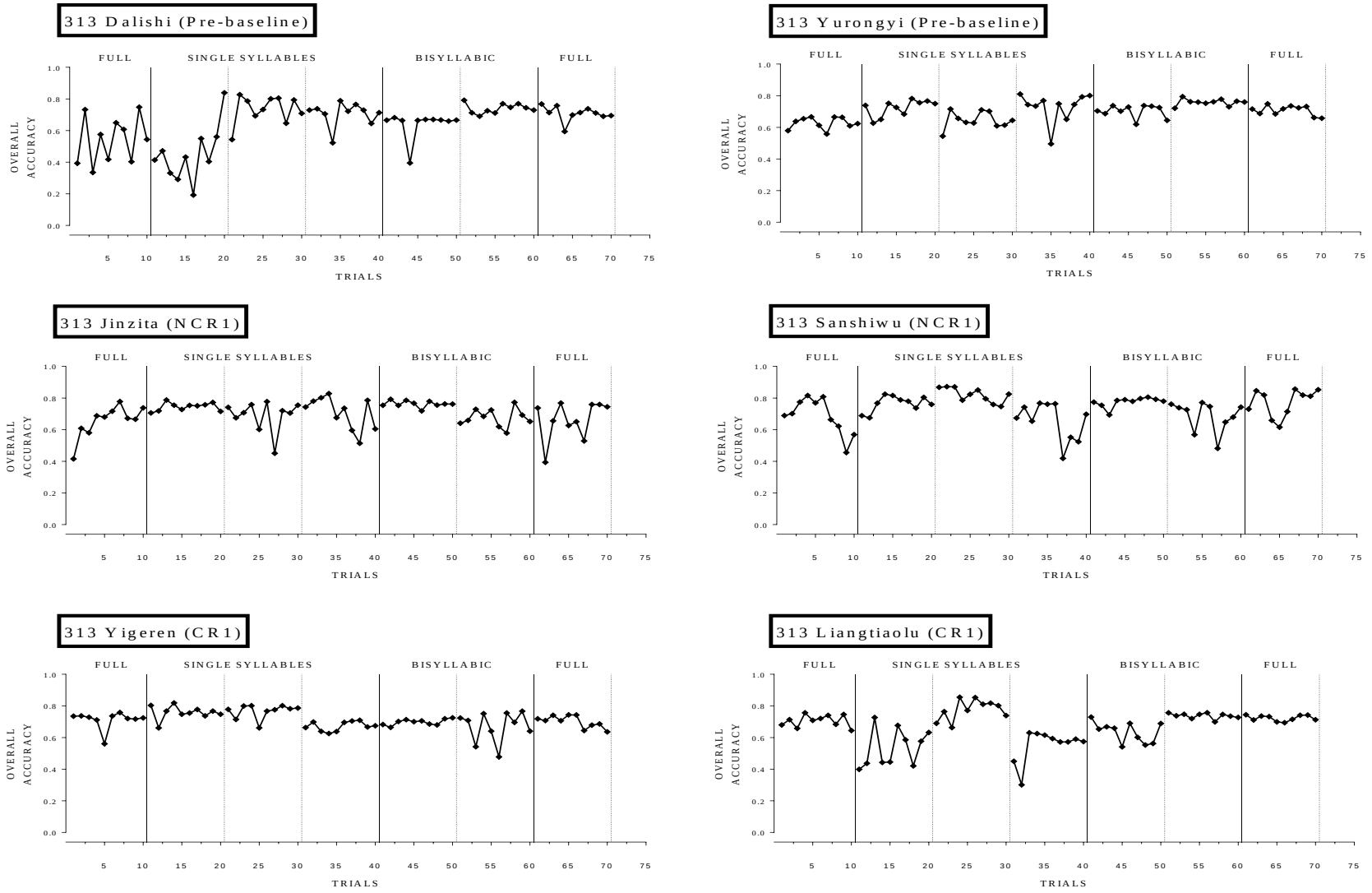

Figure D10. Performance of 313 across all phases of each word (Pre-baseline to CR1)

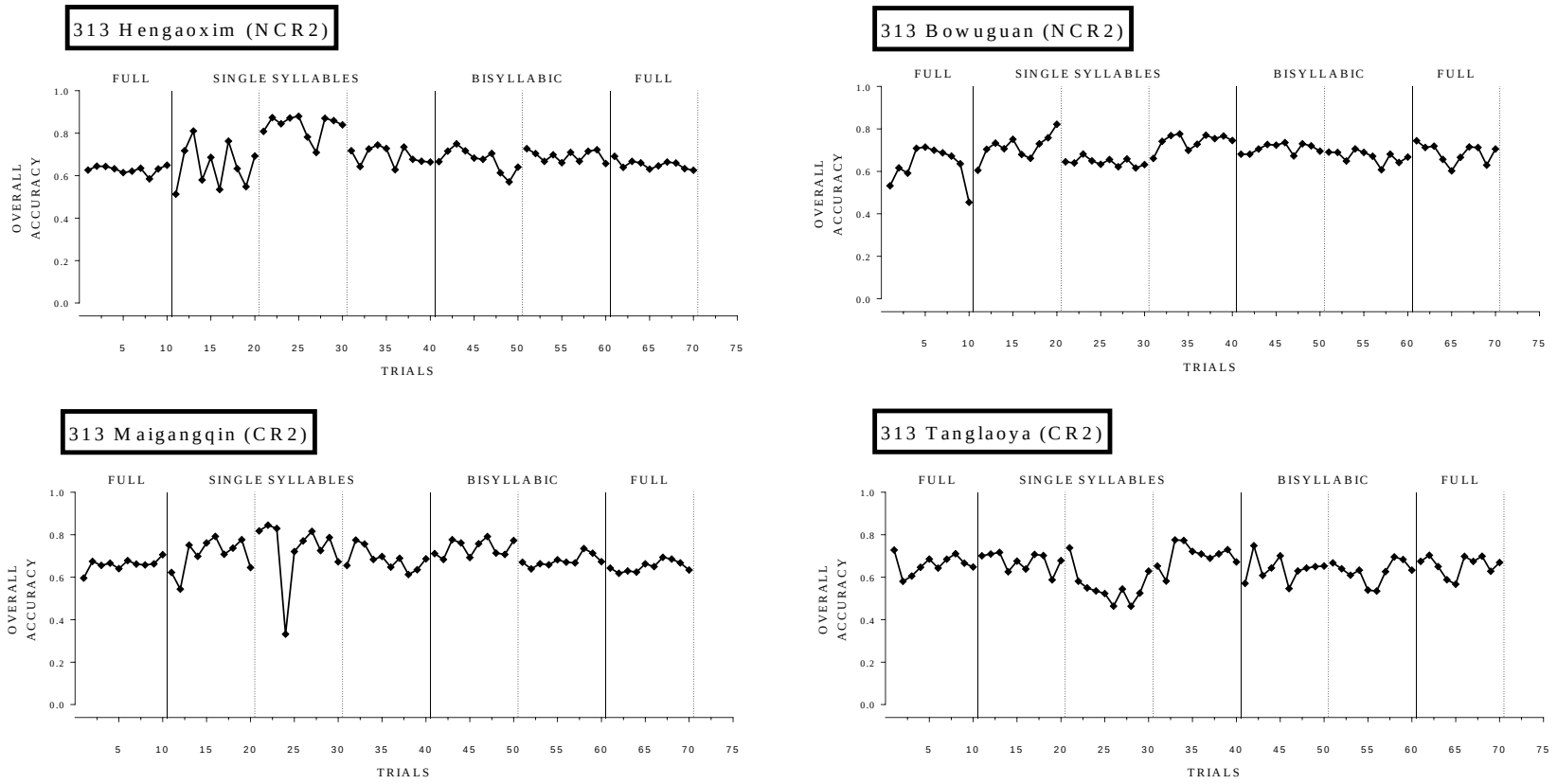

Figure D11. Performance of 313 across all phases of each word (NCR2 and CR2)

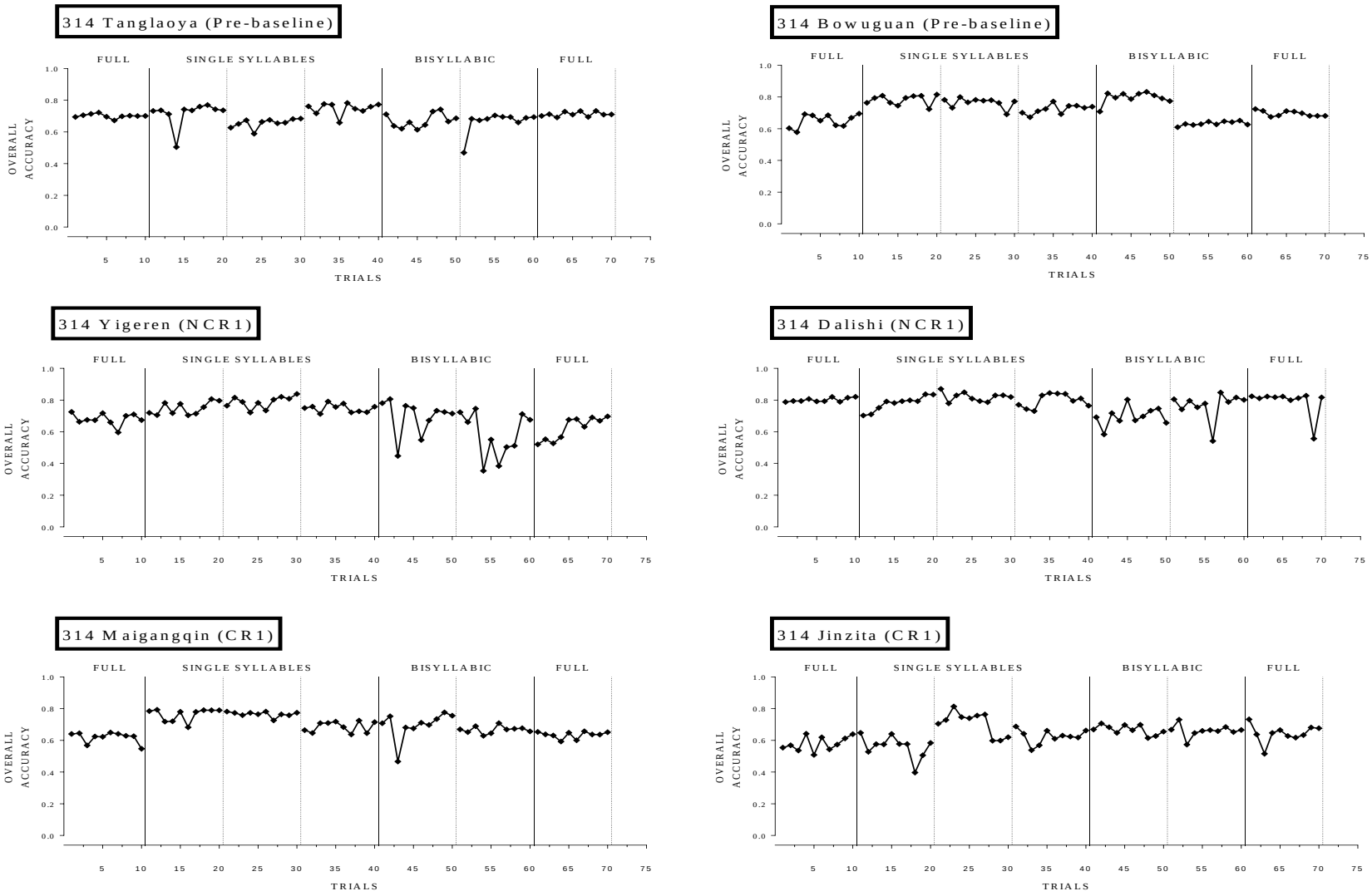

Figure D12. Performance of 314 across all phases of each word (Pre-baseline to CR1)

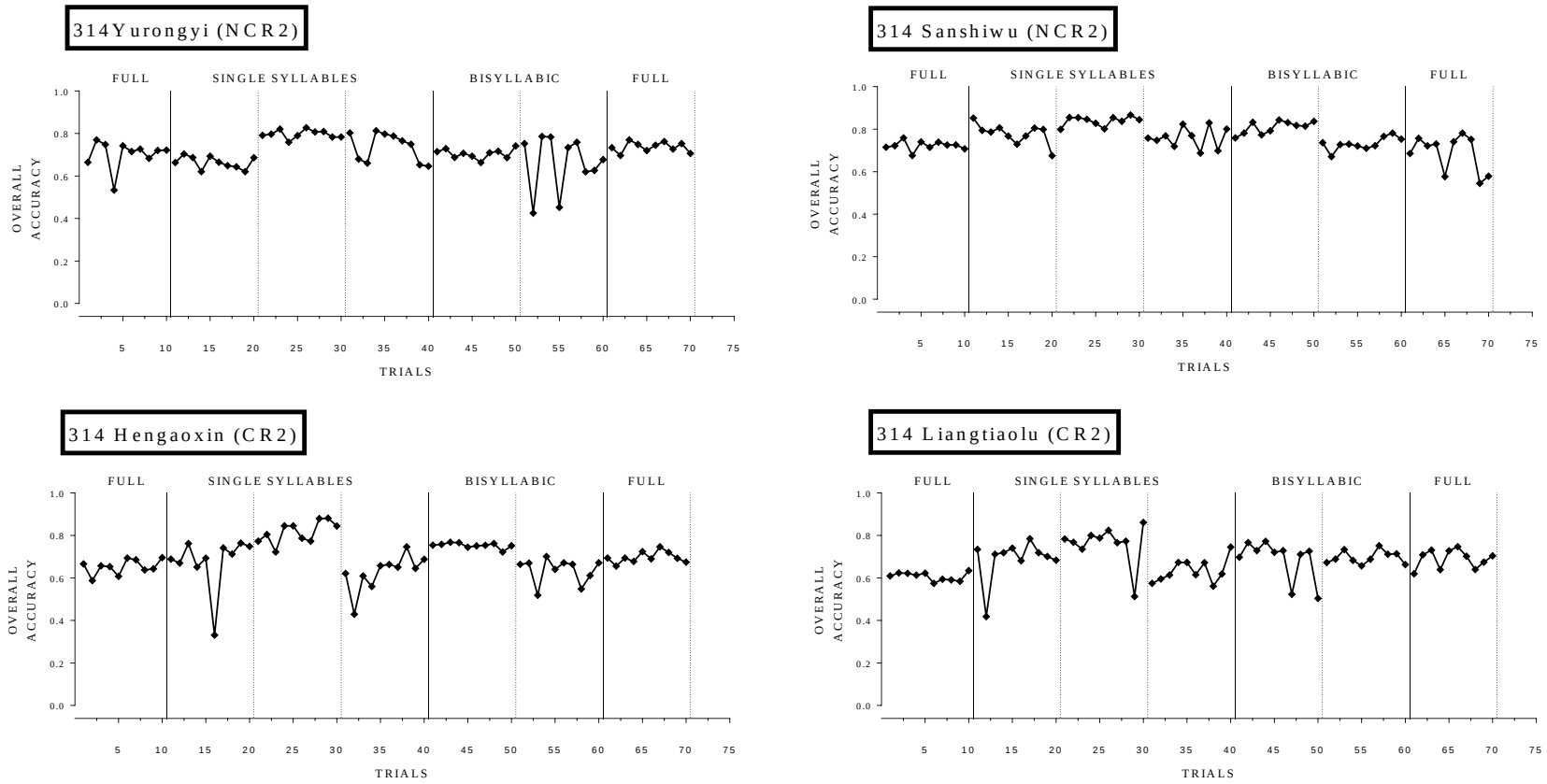

Figure D13. Performance of 314 across all phases of each word (NCR2 and CR2)

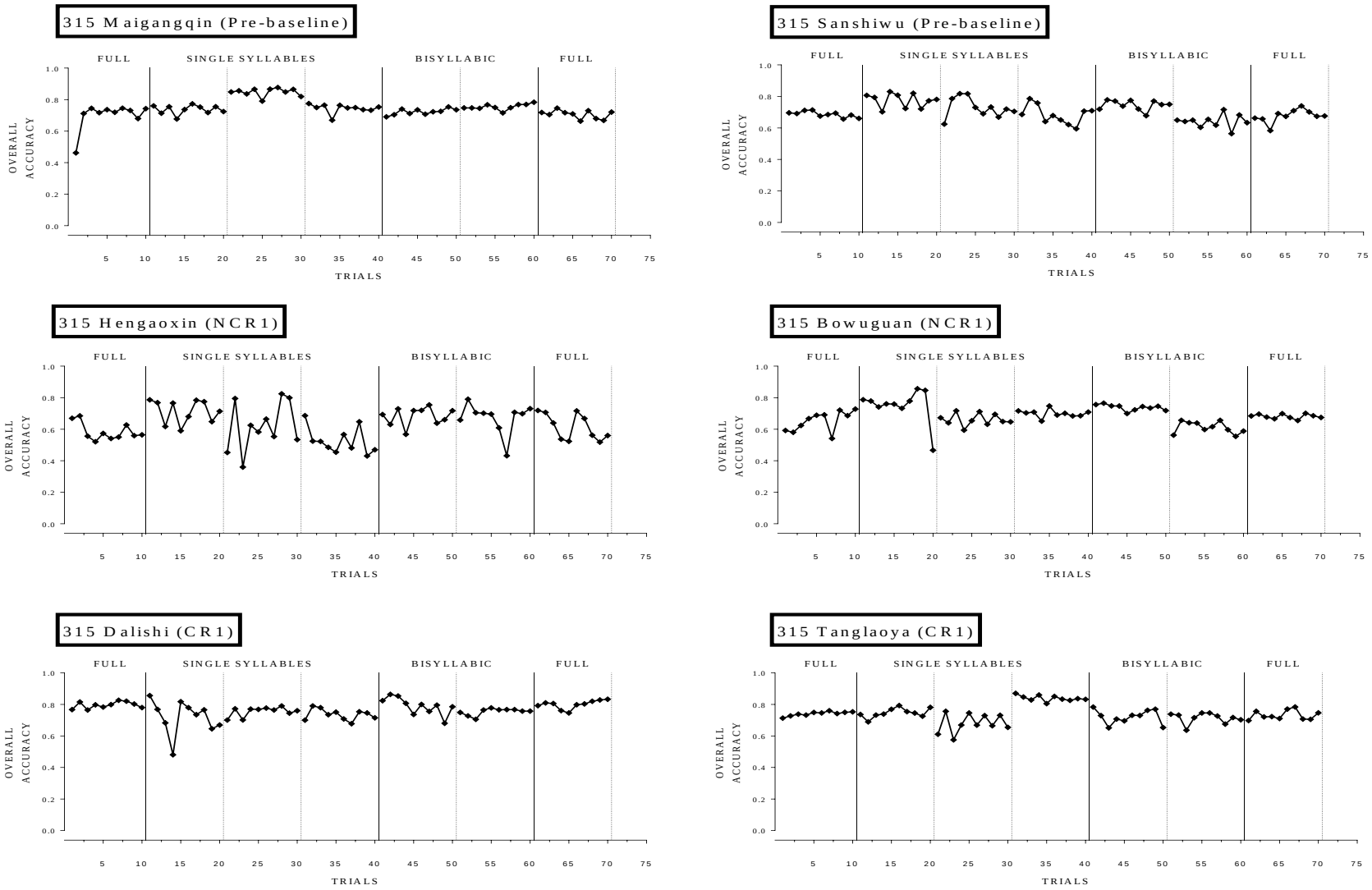

Figure D14. Performance of 315 across all phases of each word (Pre-baseline to CR1)

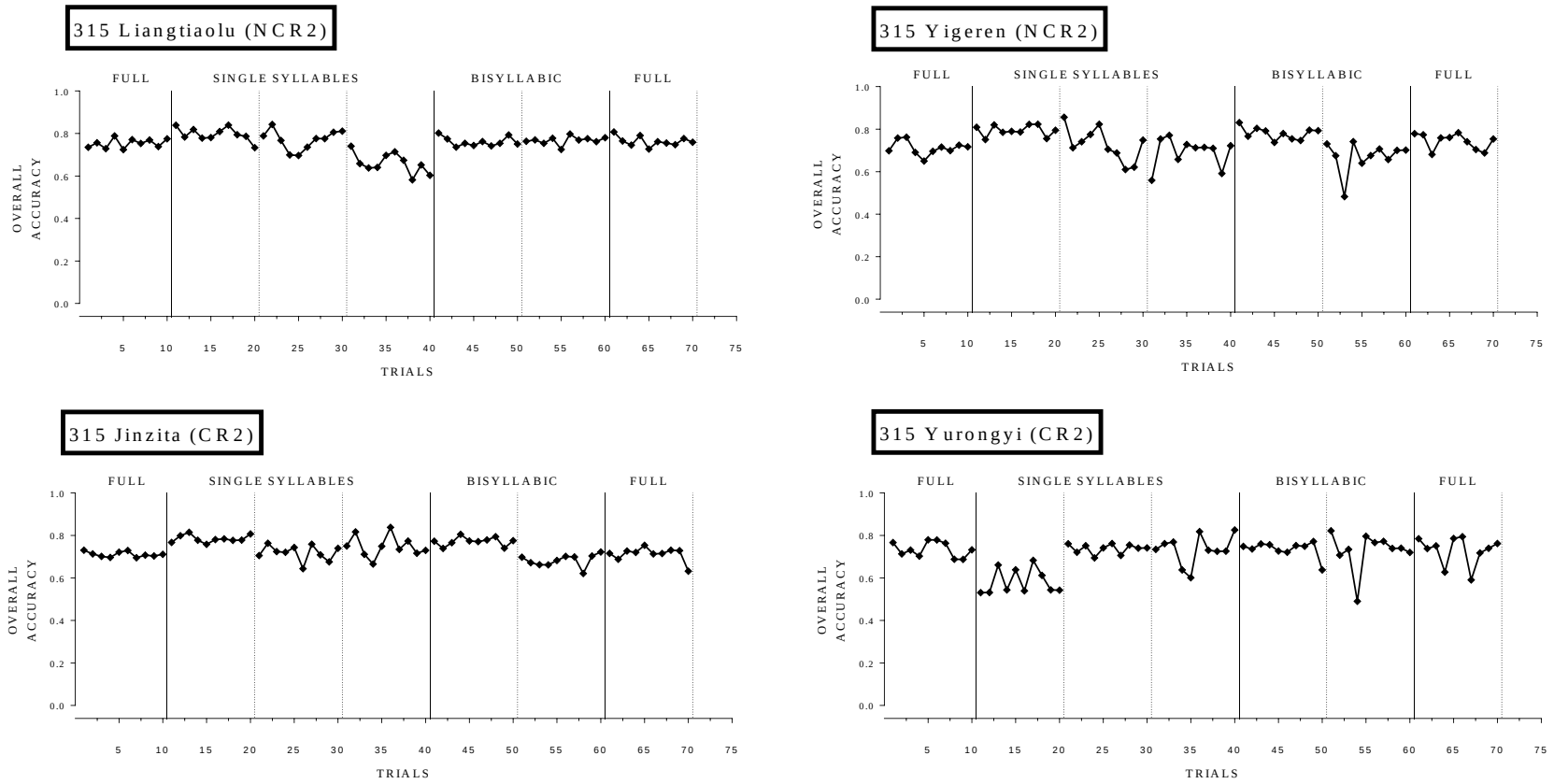

Figure D15. Performance of 315 across all phases of each word (NCR2 and CR2)