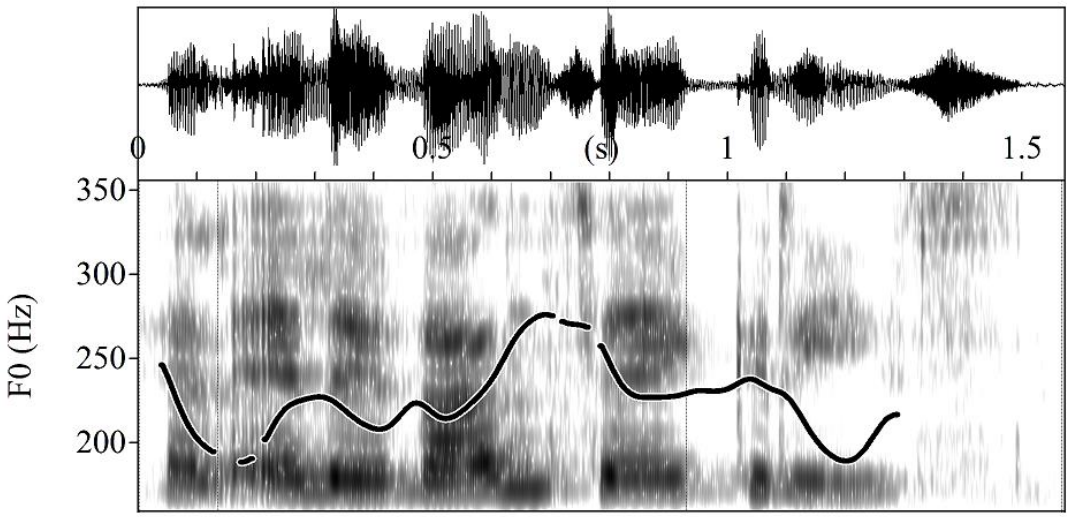
Gestures in Learning Foreign Language Prosody:

The Importance of Considering Task and Learner Characteristics

By Lisette van der Heijden

El elefante gris



Dr. M.W. Hoetjes

Dr. L.J. van Maastricht

MA Thesis – Linguistic and Communication Sciences

Radboud University

April 9, 2021

Abstract

Purpose: Previous findings on the gestural benefit in learning foreign language (FL) prosody are sparse and inconclusive (e.g., Gluhareva & Prieto, 2017; Morett & Chang, 2015; Yuan, Gonzalez-Fuente, Baills, & Prieto, 2018). A possible explanation might be that prior work varied in gesture type (i.e., beat or metaphoric) and physical involvement level (i.e., producing or perceiving gestures) used during FL prosody training. Moreover, learner characteristics, such as working memory (WM) capacity and musical aptitude, might also be relevant in this context (e.g., Özer & Göksun, 2020). Therefore, the present study aimed to disentangle the influence of different gesture types, different physical involvement levels, and learner characteristics WM capacity and musical aptitude in determining the effectiveness of FL lexical stress training.

Method: In the current experiment, Dutch learners of Spanish were instructed to read aloud short, easy to parse, Spanish phrases that differ in lexical stress from their Dutch cognates (e.g., 'pi**RÁ**mides' in Spanish, and 'pira**MI**de' in Dutch) on a pre-test (i.e., T1) and an immediate (i.e., T2) and delayed (i.e., T3) post-test. In between T1 and T2, subjects received lexical stress training in one of five training conditions which varied in gesture type and physical involvement level: *audio-visual* (AV), *AV-beat-perception*, *AV-beat-production*, *AV-metaphoric-perception*, *AV-metaphoric-production*. Additionally, between the two post-tests subjects performed a WM and musical aptitude task. In the acoustic analysis subjects' productions were coded as either on-target or not.

Results: The present study found that, irrespective of training condition, subjects significantly improved on their FL lexical stress productions from T1 to T2 and T3. Although differences between training conditions were non-significant, the present study reported several significant three-way interactions between WM capacity or musical aptitude on the one hand and testing time and training condition on the other hand. Hence, the effectiveness of gesture

type and physical involvement level in FL lexical stress training was significantly influenced by WM capacity and musical aptitude.

Conclusions: Present findings underline the importance of considering task and learner characteristics in determining the gestural benefit in learning FL prosody.

Content

1. Introduction

Previous studies have demonstrated the integrated relation between speech and gesture in communication (e.g., Kendon, 1972; McNeill, 1985), as well as the beneficial role that gestures can play in first language (L1) acquisition (e.g., Goldin-Meadow & Butcher, 2003). However, theories about the use of gestures in foreign language (FL) learning are less conclusive. Although several studies have demonstrated a gestural benefit in FL vocabulary learning (e.g., Kelly, McDevitt, & Esch, 2009; Tellier, 2008; Quin-Allen, 1995), very little is known about the effectiveness of gestures in other aspects of FL learning. Since the gestural benefit in FL vocabulary learning might be explained by the close semantic relation between speech and gesture (McNeill, 1992), one may wonder, whether gestures can also improve FL learning when the relation between speech and gesture is not semantically based. For instance, in learning FL prosody, which comprises the acquisition of phrasal and lexical stress, intonation, and rhythm (Rietveld & Van Heuven, 2009).

Exploring the effect of gestures in learning FL prosody is interesting for two main reasons. One, in FL learning, prosody is very important in how native-like the speech of a non-native speaker is perceived (e.g., Derwing & Munro, 1997; Munro & Derwing, 1995, 1999), and, since achieving native-like pronunciation is still considered the norm (Derwing & Munro, 2005), non-native speakers who make prosodic errors seem to suffer from several negative consequences (e.g., Clark & Paran, 2007; Hendriks, Van Meurs, & Hogervorst, 2016). Hence, if gestures would help in learning FL prosody this could be incorporated in new learning methods and improve acquisition. Two, beat gestures – which commonly visualise speech rhythm – are considered to be temporally aligned with prosodic prominence in speech (Loehr, 2012; Pouw & Dixon, 2019; Wagner, Malisz, & Kopp, 2014). Thus, in natural speech, there appears to be a direct relation, though not a semantic relation, between beat gestures and prosodic prominence.

In previous research some positive trends for the use of gesture in learning FL prosody have been reported (Baills, Suárez-González, González-Fuente, & Prieto, 2019; Hannah et al., 2017), but findings are barely significant (Gluhareva & Prieto, 2017; Llanes-Coromina, Prieto, & Rohrer, 2018), and some studies reported no gestural benefit at all (Eng, Hannah, Leong, & Wang, 2013; Morett & Chang, 2015). A potential explanation for these varying findings is that factors, such as gesture type (i.e., beat or metaphoric), physical involvement level (i.e., producing or perceiving gestures), and learner characteristics, might be important in determining the effectiveness of gestures in FL prosody training.

First, it might be that different types of gestures have diverging effects on learning FL prosody. Beat gestures, for instance, have a natural relation with prosodic prominence in speech (e.g., Loehr, 2012), whereas, metaphoric gestures might be better equipped to visualise specific characteristics of FL prosodic contrasts (e.g., the rising and falling of Mandarin tones (Kelly, Bailey, & Hirata, 2017)). Second, the level of physical involvement in training might also affect learning outcomes. Producing gestures involves a more embodied representation, and an additional modality, compared to perceiving gestures, and in line with the theories of embodied cognition and multimodality, it could be expected that being asked to produce gestures results in a greater learning benefit compared to perceiving gestures (e.g., Kendon, 2004; McNeill, 1992; Wilson, 2002). Third, specific learner characteristics might be important in determining the effect of gestures in learning FL prosody as well. Recently, Özer and Göksun (2020) indicated that the benefit learners might have from perceiving and producing gestures during language learning is dependent on their cognitive abilities, such as their working memory (WM) capacities. Moreover, musical aptitude, which is the ability to hear patterns in sets of sounds, might be another important learner characteristic in determining the effect of training with gestures in learning FL prosody, as it appears to be closely related to prosodic learning (Kraus & Chandrasekaran, 2010).

Although factors such as gesture type, physical involvement level, and learner characteristics appear to be important in determining the gestural benefit in learning FL prosody, it is yet unknown how they together might influence the effectiveness of FL prosody training. Accordingly, the current paper attempts to disentangle the influence of gesture type, physical involvement level, and the learner characteristics musical aptitude and WM capacity in determining the effectiveness of FL lexical stress training. Moreover, to date, studies have not yet explored whether time between training session and testing moment is important in defining the gestural benefit in learning FL prosody. By including a second (delayed) post-test the current study aims to measure the long-term effect of training with gestures.

The present study was conducted in the form of an experimental study in which Dutch subjects were trained to produce Spanish lexical stress. Dutch learners of Spanish generally struggle with Spanish lexical stress, especially in Dutch-Spanish cognates that are highly similar except for the position of the stressed syllable (e.g., 'pira**MI**des' in Dutch, but 'pi**RÁ**mides' in Spanish). In the experiment, subjects were either trained with beat gestures, with metaphoric gestures, or without gestures. Moreover, the training conditions varied in physical involvement level as subjects were told to either produce or perceive the gestures during training. Additionally, productions were measured at three different time points (i.e., pre-test, immediate post-test, and delayed post-test). Finally, WM capacity and musical aptitude were measured using a backwards digit span task (from the Automated Working Memory Assessment (AWMA) test battery (Alloway, 2007)) and a music perception task (using subtests from the Profile of Music Perception Skills (PROMS) test battery (Law & Zentner, 2012)).

The current paper aims to contribute to the growing area of research on gesture-speech integration by exploring which factors should be considered for gestures to be beneficial in learning FL prosody. Thereby, it attempts to add new insights to the debate whether language

learners could benefit from gestures when the relation between speech and gesture is not semantically based.

2. Theoretical background

2.1 Gestures: Definitions and categorisation

There are multiple definitions of gesture. According to McNeill (1992) gestures are the movements of the hands and arms that people make when they talk, and, according to Kendon (2004), a visible action can be defined as a gesture "when it is used as an utterance or as part of an utterance" (p.7). Gestures are also categorised in several ways, but one that is commonly used is McNeill's (1992) scheme that categorises gestures with regard to gesture type, which resulted in four main gesture types: iconic, metaphoric, beat, and deictic gestures. It should be noted that these gesture types are not mutually exclusive and should be seen as dimensions in that one gesture can contain aspects of several gesture types.

In McNeill's classification, iconic and metaphoric gestures are considered imagistic in that they both closely represent the semantic content of speech. *Iconic gestures* represent concrete events or objects. For instance, a speaker can make an iconic gesture for "drinking", by shaping her hands in the form of a cup and bringing it to her mouth. *Metaphoric gestures* similarly resemble what is said in speech but can represent more abstract concepts such as "imagination" or "dream". Moreover, in several studies, gestures visualising non-semantic aspects of speech, such as acoustic properties, are also defined as metaphoric (e.g., Kelly et al., 2017; Yuan, Gonzalez-Fuente, Baills, & Prieto, 2018). These gestures, for instance, visualise the rising and falling of Mandarin tones (Kelly et al., 2017).

Beat and deictic gestures are generally considered non-imagistic, meaning that they are not related to the semantic content of speech. *Deictic, or pointing gestures*, are prototypically performed with the pointing finger and can be used whenever someone tries to locate something. It should be noted that pointing gestures do not always have to point to concrete entities and do not necessarily have to be produced with the index finger (e.g., Kita, 2003). *Beat gestures* are considered to represent the rhythm of speech as small up and down

movements produced with the hands or fingers follow speech rhythm. Speakers, for instance, can use beat gestures to highlight important topics (i.e., prosodic prominence) in their speech.

2.2 Theories on gesture-speech integration

Over the past decades, researchers have become increasingly interested in how different gesture types are related to the co-occurring speech, which resulted in many theories about the relation between speech and gesture (see Wagner et al., 2014, for an overview). McNeill (1985; 1992) argued that gesture and speech must form one single integrated system in communication as they: almost always co-occur, are phonologically aligned, represent the same semantic and pragmatic concepts, co-develop in children, and break down together in aphasia. Moreover, in his *growth point theory*, McNeill (2008) argued that in an utterance the visuo-spatial image of the gesture and the linear-segmented hierarchical structure of speech cannot be separated and therefore form a single integrated system.

Following McNeill, Kita (2000) proposed the *information packaging hypothesis* which claims that – since gesture and speech are integrated – producing gestures can help speakers to structure and package visuo-spatial information into units appropriate for the linear-segmented format of speech. Therefore, speech and gesture do not only seem to be integrated in communication, but producing gestures – which closely represent visual information – can benefit speakers in structuring information to use in their co-occurring speech, as speech often has a more difficult and indirect relation to the same visuo-spatial information (see Kita & Özyürek, 2003). This close relation between gesture and speech was demonstrated in children acquiring their L1, as young children seem to use gestures when they are not yet able to express themselves in speech, and later replace these gestures by spoken words (e.g., Goldin-Meadow & Butcher, 2003; Iverson & Goldin-Meadow, 2005).

Since the information packaging hypothesis suggests that gestures might help when speech is compromised, it raises the question whether gestures can, for instance, be useful in FL learning as speech is then by definition challenging. This assumption is supported by the theories of embodied cognition and multimodality. *Embodied cognition* claims that cognition and cognitive processes are strongly influenced by the physical body and how the body moves and interacts with the world (Wilson, 2002). It suggests that our representations of a concept, object, or event involve perceptual, somatosensory, and motoric re-experiencing of the relevant event in one's self (Niedenthal, 2007). Therefore, embodied cognition argues that using the physical body benefits memory and learning. *Multimodality*, on the other hand, argues that language is a multimodal system in which the motor, visual, and speech modality are integrated to convey meaning (Kendon, 1980, 2004; McNeill, 1985, 1992, 2008; Zwaan, 2004) and proposes that using multiple modalities benefits language learning. Together, both theories seem to argue that perceiving gestures (i.e., adding the visual modality), and especially producing gestures (i.e., adding the motor modality or physical body), benefits FL learning.

2.3 Gestures in learning FL vocabulary

Many studies have explored whether gestures could benefit FL vocabulary learning, and found that iconic gestures, representing the semantic content of the novel words, benefitted word learning (e.g., Kelly et al., 2009; Tellier, 2008; Quin-Allen, 1995). Kelly et al. (2009), for instance, found that, in English native speakers learning Japanese words, congruent iconic gestures (i.e., representing the semantic meaning of the novel words) benefitted word learning, whereas incongruent iconic gestures (i.e., representing another meaning than the novel word) hurt word learning. Several studies argued that, in line with the theories of multimodality and embodied cognition, the gestural benefit in FL vocabulary learning should

be greater when learners are instructed to produce gestures compared to when learners are instructed to merely perceive gestures (Tellier, 2008; Quin-Allen, 1995), but to date findings are inconclusive. Morett (2018), for instance, found that producing gestures resulted in better memorisation of novel words compared to perceiving gestures, whereas Sweller, Shinooka-Phelan, and Austin (2020) reported a similar benefit for producing and perceiving gestures in FL vocabulary learning. Some researchers proposed that novel words learned with iconic gestures would also be more resistant to memory decay (Macedonia & Klimesch, 2014; Macedonia, Müller, & Friederici, 2011). Macedonia and Klimesch (2014), for instance, found that when learners were instructed to produce gestures in training, they were able to name significantly more words at each time point, including fourteen weeks and fourteen months after training, compared to learners who were trained without gestures.

2.4 Gestures in learning FL phonetics

Although many studies have demonstrated a benefit for gestures in FL vocabulary learning, very little is known about the possible benefits of gestures in other aspects of FL learning. A possible explanation for the gestural benefit in FL vocabulary learning might be the close semantic relation between speech and gesture (McNeill, 1992). For instance, the verb *typing* can be easily represented by an iconic gesture that shows the act of typing by placing both hands horizontally in front of the body and moving the fingers vertically up and down. Since, in other aspects of language, the relation between speech and gesture might be less immediately apparent, one may wonder, whether gestures would still improve FL learning. However, as theories of embodied cognition and multimodality claim that using multiple modalities and the physical body benefits learning in general, gestures are expected to benefit FL learning beyond vocabulary acquisition.

One aspect in which the relation between gesture and speech might be less straightforward is phonetic learning, which includes both segmental and suprasegmental learning, as it constitutes the acquisition of speech sounds. Segmental learning can be defined as acquiring a language's vowels and consonants (i.e., learning the small sounds within a word) (Rietveld & Van Heuven, 2009). Several studies have explored whether gestures could benefit FL segmental learning and reported mixed findings which might be explained by the difference between perception and production studies. Studies found that in learning to *perceive* Japanese vowel-length contrasts[1] or singleton vs. geminate distinctions[2], FL learners did not benefit from producing or perceiving (beat or metaphoric) gestures during a short training session (Hirata & Kelly, 2010; Hirata, Kelly, Huang, & Manansala, 2014; Kelly et al., 2017; Kelly, Hirata, Manansala, & Huang, 2014; Kelly & Lee; 2012; Li, Baills, & Prieto, 2020).

On the other hand, Li et al. (2020) found that metaphoric gestures did significantly improve the *production* of Japanese vowel-length contrasts in FL learners. Similarly, Hoetjes and Van Maastricht (2020) examined whether perceiving pointing (i.e., pointing to lips to focus attention on articulatory movements) or iconic gestures (i.e., representing articulatory movements necessary to produce target phoneme) in a short training session improved the *production* of the Spanish phonemes /u/ and /θ/ in Dutch-speaking learners. They found that, in general, gestures seemed to benefit FL phoneme acquisition, but that the combination of type of gesture and type of phoneme mattered. So, whereas training with gestures does not seem to aid the *perception* of FL segments it might benefit the *production* of FL segments. Another possible explanation for the limited gestural benefit in FL segmental learning might

[1] In Japanese the length of the vowel – long or short – changes the meaning of the word (Hirata & Kelly, 2010).

[2] In Japanese the duration of a consonant can change the meaning of a word. A geminate is a relatively long consonant, whereas a singleton is short (Kelly & Lee, 2012).

be that gestures do not easily map onto such small units within a word (i.e., segments), which can cause confusion in learners about what the gestures represent (Kelly et al., 2017).

2.5 Gestures in learning FL prosody

Another aspect of phonetic learning is suprasegmental, or prosodic, learning. Prosody can be defined as all sound properties that accompany vowels and consonants, but which are not limited to single sounds and often extend over syllables, words, or phrases (Rietveld & Van Heuven, 2009). Prosodic aspects of language, for instance, include phrasal and lexical stress, intonation, and rhythm. Prosody has several important functions. Phrasal prosody, for instance, can function to structure information or express emotion, and, at a lexical level, prosodic emphasis can serve to differentiate between words. The English word *present*, for example, has two interpretations (i.e., the verb *to pre**SENT*** and the noun ***PRE**sent*) depending on the syllable that receives lexical stress.

There are two main arguments to explore the effectiveness of gestures in learning FL prosody. One, prosody appears to be highly important in how native-like the speech of a non-native speaker is perceived (e.g., Derwing & Munro, 1997; Munro & Derwing, 1995, 1999). Van Maastricht, Krahmer, and Swerts (2016), for instance, found that native Dutch speakers were able to distinguish native from non-native Dutch speakers based on prosodic cues alone. Moreover, as achieving native-like pronunciation is considered the norm (Derwing & Munro, 2005), non-native speakers who make prosodic errors seem to suffer several negative consequences, such as being taken less seriously (Clark & Paran, 2007), perceived as less likeable (Hendriks et al., 2016), and as less credible (Lev-Ari & Keysar, 2010) than native speakers. Even though several studies found that training learners on FL prosody, either in combination with segments or not, significantly improved native listeners' perception of the learners' comprehensibility (Derwing, Munro, & Wiebe, 1998; Derwing & Rossiter, 2003;

Gordon, Darcy, & Ewert, 2013), accentedness (Behrman, 2014; Derwing et al., 1998), and fluency (Derwing & Rossiter, 2003), prosody is often overlooked in FL teaching methods. Therefore, finding an effective teaching method, which might be incorporating gestures, can help FL speakers to improve their prosodic abilities and avoid negative consequences.

Two, beat gestures are considered to be temporally aligned with prosodic prominence in speech (Loehr, 2012; Pouw & Dixon, 2019; Wagner et al., 2014), which indicates that beat gestures are commonly produced at the same time as the most prominent part of the co-occurring speech stream. Hence, there appears to be a clear relation, though not a semantic relation, between beat gestures and speech prosody. Moreover, in both the perception and production of speech, beat gestures seem to influence prosodic prominence. In speech perception, beat gestures can influence what part of the co-occurring speech is perceived as most prominent (Krahmer & Swerts, 2017), and in speech production, beat gestures can lead the co-occurring word to receive prosodic prominence in speech – even if this word did not receive prosodic prominence at first (Krahmer & Swerts, 2017), which suggests that the acoustic realisation of speech can be changed by simply producing beat gestures (Bosker & Peeters, 2021). So, as beat gestures are related to prosodic prominence in speech and can influence what part of the co-occurring speech receives prosodic prominence, it might be that beat gestures are especially helpful in learning FL prosody.

Several studies have started to explore the effectiveness of gestures in learning FL prosody, but findings are inconclusive, and only a limited number of studies has found a significant benefit for training with gestures (Baills et al., 2019; Hannah et al., 2017; Yuan et al., 2018). A potential explanation for these varying findings is that factors, such as gesture type, physical involvement level, and learner characteristics, might be important in determining the effectiveness of gestures FL prosody training. Therefore, relevant findings will be discussed in relation to these three factors.

2.5.1 Gesture type

In studies exploring the effectiveness of gestures in FL prosody training two types of gestures have been used, being: beat and metaphoric gestures. Gluhareva and Prieto (2017) and Llanes-Coromina et al. (2018), for instance, both explored the effect of training with beat gestures in Catalan-Spanish speakers learning English phrasal stress[3]. Gluhareva and Prieto (2017) reported a trend for learners to be considered as less accented after training with beat gestures compared to after training without gestures. However, only in the difficult items the difference in improvement between learners trained with or without gestures was found to be significant. Llanes-Coromina et al. (2018) found a similar (non-significant) trend suggesting that learners who were instructed to produce beat gestures during training were given higher ratings on their accentedness, comprehensibility, and fluency while reading a FL text compared to learners who did not produce any gestures during training. Therefore, although beat gestures are closely related to prosodic prominence in speech (e.g., Loehr, 2012), the non-significant findings suggest that the benefit of beat gestures in learning FL phrasal stress might be limited.

Other studies explored the effectiveness of metaphoric gestures in learning FL prosody as these gestures can visualise specific characteristics of the prosodic contrasts (e.g., the rising and falling of Mandarin tones (Kelly et al., 2017)). These studies seem to suggest that the effectiveness of metaphoric gestures might depend on the FL prosodic contrast being

[3] There is a clear difference in linguistic rhythm patterns between English (i.e., syllable-timed) and Catalan-Spanish (i.e., stress-timed). English, for instance, is considered to have stronger pre-boundary lengthening effects and stronger head marking compared to Catalan-Spanish (Prieto et al., 2012).

learned. In learning to perceive Mandarin tones[4], several studies reported contrasting findings on the effectiveness of metaphoric gestures (representing tone pitch contours). Findings by both Eng et al. (2013) and Morett and Chang (2015) suggest that perceiving metaphoric gestures did not significantly improve Mandarin tone perception in FL learners. Eng et al. (2013), for instance, instructed Catalan-Spanish learners to identify Mandarin tones presented in either an audio-facial, an audio-gesture, or an audio-facial-gesture modality, and found that learners had similar improvements in all three modalities. On the other hand, Baills et al. (2019) and Hannah et al. (2017) both found that training with metaphoric gestures significantly improved Mandarin tone perception in FL learners compared to training without gestures. Hannah et al. (2017) found that, especially when Mandarin tones were embedded in noise, Catalan-Spanish learners significantly improved from the use of metaphoric gestures. These findings suggest that metaphoric gestures might especially benefit learning Mandarin tone perception under certain circumstances, for instance, if sounds are embedded in noise.

In acquiring other types of prosodic contrast metaphoric gestures might be more effective. Yuan et al. (2018), for instance, explored whether speakers of a tonal language (i.e., Mandarin) could benefit from perceiving metaphoric gestures in learning an intonational language[5] (i.e., Spanish) and found that learners performed significantly better on a discourse completion task after a short training session with metaphoric gestures compared to after training without gestures. Moreover, Kelly et al. (2017) studied the effect of perceiving congruent or incongruent metaphoric gestures in English speakers learning to perceive

[4] Unlike intonational languages (e.g., English and Catalan-Spanish), in tonal languages (i.e., Mandarin) lexical tone contrasts are used to distinguish in semantic meaning between otherwise segmentally identical words (Baills et al., 2019).

[5] In learning an intonational language (e.g., English), speakers of tonal languages (i.e., Mandarin) have the tendency to produce a high tone on the target nuclear accent when it should be a low tone (Yuan et al., 2018).

Japanese phrasal stress[6], and found that incongruent gestures hurt the perception of Japanese phrasal stress, whereas congruent gestures benefitted perception. So, whereas the effectiveness of gestures is unclear in learning Mandarin tones, in learning intonational languages and Japanese phrasal stress, metaphoric gestures seem to improve perception, which suggests that the type of FL prosodic contrast might be important in defining the effect of metaphoric gestures.

To date, only one study has explored the difference in effectiveness between perceiving beat or metaphoric gestures in learning FL prosody. In a precursor of the current study, Van Maastricht, Hoetjes, and Van Drie (2019) trained native Dutch speakers – without any experience with Spanish – to produce Spanish lexical stress[7]. Learners performed a pre- and post-test in between which they received training in one of three experimental conditions: audio-visual, audio-visual with metaphoric gestures, or audio-visual with beat gestures. The metaphoric gestures represented the duration of the stressed syllables and were produced by placing both hands together – in front of the body – and moving them apart at the moment when lexical stress should be placed. The beat gestures were made by making a chopping movement with one hand during the stressed syllable. They found no significant differences in improvement between training conditions but reported a trend in which subjects improved most in the audio-visual-metaphoric condition, followed by the audio-visual-beat condition. Least improvement was found in the audio-visual condition. So, in comparing the effect of beat and metaphoric gestures in FL learning FL prosody, metaphoric gestures seemed to be most effective. However, more research is needed as to date only Van Maastricht et al. (2019)

[6] Unlike in English declarative and question intonations, in Japanese phrasal stress throughout the sentence is identical and only the intonation of the final syllable changes (i.e., falls or rises) (Kelly et al., 2017).

[7] Dutch learners generally struggle with Spanish lexical stress, especially in Dutch-Spanish cognates that are highly similar except for the position of the stressed syllable (Van Maastricht et al., 2019).

investigated the possible differences in effectiveness between different gesture types in learning FL prosody.

Although, in learning FL prosody, more positive findings on the effectiveness of metaphoric gestures compared to beat gestures have been reported, the varying findings in earlier studies suggest that several underlying factors might be important in determining the effectiveness of both beat and metaphoric gestures in learning FL prosody. A first factor that should be considered is the type of FL prosodic contrast being learned. For instance, whereas merely positive findings were reported on the effectiveness of metaphoric gestures in learning FL phrasal stress (Kelly et al., 2017; Yuan et al., 2018), findings concerning the benefit of metaphoric gestures in learning Mandarin tones were inconclusive (e.g., Baills et al., 2019; Eng et al., 2013). Another factor that might be relevant is whether learners' improvements on FL prosody are measured with a production or perception task. Li et al. (2020), for instance, demonstrated that training with metaphoric gestures significantly improved the *production*, but not the *perception*, of FL vowel-length contrasts. Therefore, it might be that the effectiveness of both beat and metaphoric gestures in FL prosody training varies depending on whether learners are asked to produce or perceive the FL prosodic contrast.

A final factor that should be considered is the type of measurement used to analyse learners' FL productions. Whereas most studies used accentedness ratings, instead of acoustic analyses, to determine whether learners improved on their FL productions, accentedness ratings might not be the best suited measurement, as it is often difficult for raters to solely focus on prosody in determining how accented speech is. Moreover, several studies that used accentedness ratings did not find a significant benefit for training with (beat) gestures (e.g., Gluhareva & Prieto, 2017; Kushch et al., 2018), whereas Kelly et al. (2017), who used acoustic analyses, did find a significant benefit for training with (metaphoric) gestures.

Hence, it might be that type of measurement influenced earlier findings on the effectiveness of both beat and metaphoric gestures in learning FL prosody.

2.5.2 Physical involvement level

Whereas previous studies explored the effectiveness of either perceiving or producing gestures in FL prosody training, only a limited number of studies investigated whether physical involvement level (i.e., perceiving or producing gestures) influences the effectiveness of training with gestures. Kushch et al. (2018) explored the difference between perceiving and producing beat gestures in Catalan-Spanish speakers learning to produce English phrasal stress, and showed that learners who were instructed to produce beat gestures during training were considered to have improved significantly more on accentedness compared to subjects who were instructed to merely perceive beat gestures. Similarly, Baills (2016), found that, in Catalan-Spanish speakers, producing metaphoric gestures in training resulted in significantly better Mandarin tone identification compared to perceiving metaphoric gestures. In contrast, Baills et al. (2019) reported no significant differences between perceiving and producing metaphoric gestures, even though they used a highly similar experimental design as Baills (2016).

Previous findings seem to suggest that, in line with multimodality and embodied cognition theories, producing gestures, which adds an extra modality and creates a more embodied representation (e.g., Kendon, 2004; McNeill, 1992; Wilson, 2002), is more effective in FL prosody training compared to perceiving gestures, but findings are inconclusive. Moreover, the number of studies that have examined the influence of physical involvement level in FL prosody training is limited.

2.5.3 Learner characteristics

The variability in previous findings might also be explained by individual differences between learners. Generally, gesture studies focus on group-comparisons, disregarding all individual variation, but, according to Özer and Göksun (2020), depending on their cognitive abilities, people highly differ in how much they can benefit from producing or perceiving gestures during comprehension and learning. Two specific learner characteristics, learners' WM capacity and musical aptitude, might be important in explaining the effect gestures could have in learning FL prosody.

WM capacity

WM involves the temporary storage and manipulation of information and, although, it is considered to be important for language processing (Baddeley, 2003), the role WM has in multimodal language processing is still unknown. Several conflicting theories have been proposed. According to Paivio's (1991) dual coding theory, it might be easier to retain information in memory when it is presented via multiple routes, for instance via speech and gesture, as these multiple routes lower the WM load. This was demonstrated in several studies who found that using co-speech gestures lowered speakers' WM load, making it easier to retain new information (e.g., Cook, Mitchell, & Goldin-Meadow, 2008; Cook, Yip, & Goldin-Meadow, 2012). Hence, as co-speech gestures lower the WM load, learners with a low WM capacity might benefit more from using gestures in learning a FL compared to learners with a high WM capacity.

On the other hand, Mayer and Moreno (2003) argue that activating multiple routes, in this case speech and gesture, can overload the WM system. This was also demonstrated in several studies who found that a combination of gesture and speech overloaded the system and negatively affected memory retention (e.g., Hirata & Kelly, 2010; Kelly et al., 2017).

Kelly et al. (2017) suggested that the level of abstractedness of the relation between speech and gesture might be important. Therefore, it might be that when the relation between speech and gesture is clear (e.g., in vocabulary learning), both learners with high and low WM capacity could benefit from training with gestures (e.g., dual coding theory), whereas, when the relation between speech and gesture is less immediately apparent (e.g., in segmental learning), only learners with high WM capacity benefit from training with gestures, as gestures overload the WM system in learners with a lower WM capacity. Moreover, it should be noted that WM also seems to be a measure of language aptitude in that learners with greater WM capacity are thought to be better equipped to learn a FL (Christiner & Reitener, 2013, 2015, 2016; Wen & Skehan, 2011; Robinson, 2005).

Earlier findings concerning the influence of WM capacity on learning FL prosody and concerning the influence of WM capacity on the effectiveness of training with gestures are sparse and inconclusive. Li et al. (2020), for instance, found no significant influence of WM capacity on learner's production and perception of FL vowel-length contrasts, whereas Zhang, Baills, and Prieto (2020) found that learners with a higher WM capacity produced more on-target FL lexical stress compared to learners with a lower WM capacity. Moreover, to date, only Li et al. (2020) explored the influence of WM capacity on the effectiveness of gestures in FL prosody training and found no predictive function for WM capacity. Hence, more research is needed.

Musical aptitude

Another relevant learner characteristic to consider is musical aptitude. Musical aptitude can be defined as the ability to see patterns in sets of sounds, or as the potential for learning music (Law & Zentner, 2012). It seems to be related to learning prosody as the basic acoustic features of musical sounds (pitch, timing and timbre) play a key role in both speech and music

(Kraus & Chandrasekaran, 2010). Hence, these shared features might suggest that learners with high musical aptitude are better equipped to learn FL prosody. Moreover, according to Christiner and Reitener (2018) there might be an interaction between WM capacity and musical aptitude as they found that learners with a high musical aptitude generally showed an enhanced WM capacity.

Previous findings concerning the influence of musical aptitude on learning FL prosody and concerning the influence of musical aptitude on the effectiveness of training with gestures are limited. Yuan et al. (2018) found that learners with a high musical aptitude – measured as melody and pitch perception – performed better on producing FL intonational phrases compared to learners with moderate and low musical ability. Similarly, Li et al. (2020) found that a higher rhythm aptitude, but not pitch aptitude, positively influenced FL Japanese vowel-length perception, though not production. In contrast, Zhang et al. (2020) found no significant influence of musical aptitude on FL lexical stress production. Moreover, Yuan et al. (2018) found that learners with moderate musical abilities benefitted more from instruction with metaphoric gestures compared to learners with high or low musical abilities. According to the authors, the finding that learners with high musical abilities benefitted less from training with gestures might be explained by ceiling effects which resulted in less available room to demonstrate improvement. In contrast, Li et al. (2020) found that musical aptitude did not significantly predict the benefit learners experienced from gestures in FL prosody training.

3. Research questions & hypotheses

Although it seems that gestures can be effective when the relation between speech and gesture is not semantically based, such as in learning FL prosody, studies have reported many conflicting results. The varying findings might be explained by factors such as learner characteristics and studies' methodological choices. Although several studies have started to explore the influence of some of these factors (e.g., Li et al., 2020; Yuan et al., 2018), no study to date has disentangled the influence of multiple influential factors on the effectiveness of FL prosody training. Therefore, the current study aims to disentangle how (1) type of gesture, (2) physical involvement level, and (3) learner characteristics (i.e., WM capacity and musical aptitude) influence the effectiveness of FL lexical stress training by Dutch learners of Spanish. Lexical stress was selected because Dutch learners of Spanish often have much difficulties in acquiring Spanish lexical stress, especially in Dutch-Spanish cognates that are highly similar except for the position of the stressed syllable (e.g., pi**RÁ**mides in Spanish, but pira**MI**de in Dutch). Moreover, to date, studies on the use of gestures in learning FL lexical stress are sparse. Additionally, the current study aims to explore whether there are any long-term effects concerning the use of gestures in learning FL prosody. Although some long-term effects for gestures have been reported in FL vocabulary learning (e.g., Macedonia & Klimesch, 2014; Macedonia et al., 2011), these effects have not yet been researched in learning FL prosody.

The present study aims to determine what the effects are of different gesture types, different physical involvement levels, and the learner characteristics WM capacity and musical aptitude on learning FL lexical stress. Additionally, the present study will make a first try in exploring whether a benefit of training with gestures could be visible in the long-term. Therefore, the current project aims to answer the following questions:

1. Does gesture type (i.e., beat or metaphoric) influence the effectiveness of FL lexical stress training?
2. Does level of physical involvement (i.e., producing or perceiving gestures) influence the effectiveness of FL lexical stress training?
3. Can the subject's improvement on FL lexical stress production be explained by the subject's WM capacity or musical aptitude?
4. Are there any long-term effects of training with gestures on producing FL lexical stress?

Concerning these questions, we have several hypotheses:

1. In line with previous studies on the effectiveness of beat versus metaphoric gestures in learning FL prosody (e.g., Van Maastricht et al., 2019), we hypothesise that metaphoric gestures will be more effective compared to beat gestures in FL lexical stress training. Even though beat gestures are closely aligned with prosodic prominence in speech (Loehr, 2012; Pouw & Dixon, 2019; Wagner et al., 2014), we suggest that metaphoric gestures will be better suited to learn FL lexical stress as they can visualise specific characteristics of the prosodic contrast and therefore might be more notable to the subjects.
2. In line with the theories of multimodality and embodied cognition, and with findings by Kushch et al. (2018) and Baills (2016), we hypothesise that producing gestures – and thereby adding an extra modality and a more embodied representation (e.g., Kendon, 2004; McNeill, 1992; Wilson, 2002) to the speech signal will result in more correct FL lexical stress productions compared to merely perceiving gestures.

3.

a. As earlier studies suggest that WM capacity can be seen as a measure of language aptitude (e.g., Christiner & Reitener, 2013, 2015, 2016), we hypothesise that subjects with a higher WM capacity will produce more correct FL lexical stress compared to subjects with a lower WM capacity. Moreover, previous studies propose conflicting theories about the influence of WM capacity on the effectiveness of training with gestures (e.g., Mayer & Moreno, 2003; Paivio, 1991) and suggest that the level of abstractedness between speech and gesture might be important (Kelly et al., 2017). As the relation between speech and gesture is not immediately apparent in learning FL lexical stress, we hypothesise that lexical stress training with gestures will be more effective in subjects with a higher WM capacity. Additionally, we hypothesise that the influence of WM capacity might differentiate depending on gesture type and physical involvement level used in FL lexical stress training. However, as this has not yet been previously explored, we cannot predict what these differences will be.

b. In line with earlier studies who suggest that learning prosody might be related to musical aptitude (e.g., Kraus & Chandrasekaran, 2010), and earlier findings by Yuan et al. (2018), we hypothesise that subjects with a higher musical aptitude will produce more correct FL lexical stress compared to subjects with a lower musical aptitude. Moreover, as previous findings are sparse it is difficult to make predictions about the influence of musical aptitude on the effectiveness of gestures in learning FL lexical stress. But, in line with Yuan et al. (2018), we hypothesise that especially subjects with moderate to high musical abilities will benefit from lexical stress training with gestures.

Additionally, we hypothesise that the influence of musical aptitude might differentiate depending on gesture type and physical involvement level used in FL lexical stress training. However, as this has yet not been previously explored, it is impossible to predict what these differences will be.

4. As studies in FL vocabulary learning (e.g., Macedonia & Klimesch, 2014; Macedonia, et al, 2011) have reported long-term memory enhancing benefits for training with gestures, we hypothesise that in the long-term training with gestures will result in more correct FL lexical stress productions compared to training without gestures.

4. Methodology

4.1 Subjects

A sample of 60 native Dutch speakers was recruited to explore the effectiveness of gestures in learning Spanish lexical stress. The sample contained 45 women (75%) and 15 men (25%). The mean age of subjects in the sample was 23.86 (*SD* = 8.68) years. Men were on average 23.07 (*SD* = 3.86) years old, whereas women had a mean age of 24.13 (*SD* = 9.76) years old, which was a non-significant difference, $F(2.38, 8.87) = .79$, $p = .683$ (calculated in R (version 4.0.3, R Core Team, 2019)). Almost all subjects were currently enrolled as a student in higher education or university or already graduated with a high level of education (90%: 11.67% higher education, 58.33% WO Bachelor, and 20% WO Master). Six subjects (10%) graduated with a lower level of education (1.67%) or did not provide information about educational level (8.33%).

Subjects who were raised multilingually or who had previous experience with the Spanish language were discouraged from participating during recruitment. Subjects were considered to have earlier experience with Spanish if they: were native Spanish speakers, (had) followed a formal Spanish course, had stayed in a Spanish speaking country to learn the language, or had stayed in a Spanish speaking country for more than three months. Unfortunately, nine subjects (15%) indicated that they had followed some Spanish lessons. Four subjects (6.67%) had shortly used an online language learning program, whereas the others (8.33%) had received Spanish lessons in school for a brief period of time. Moreover, most subjects were raised monolingually, but six subjects (10%) indicated that they came from multilingual homes in which two or three languages were used. To keep a maximum number of subjects per experimental condition, all subjects were included in the analysis, but both whether subjects (had) followed Spanish lessons and the number of languages subjects

spoke at home were added as control predictors to the analysis in which both appeared to be irrelevant.

Subjects with auditory or visual impairments which could not be adjusted were similarly advised to not participate in the experiment. However, the final sample contained two subjects who had audio-visual impairments (3.33%) and three subjects who were diagnosed with dyslexia (5%). Again, to keep a maximum number of subjects, all subjects were included in the analysis and two variables, on whether subjects have audio-visual impairments and/or dyslexia, were added as control predictors to the analysis in which both appeared to be irrelevant.

The current experiment contained five training conditions: (1) audio-visual (AV), (2) audio-visual beat gesture perception (AV-B-perc), (3) audio-visual beat gesture production (AV-B-prod), (4) audio-visual metaphoric gesture perception (AV-M-perc), and (5) audio-visual metaphoric gesture production (AV-M-prod), across which subjects were evenly distributed (i.e., between-subject design). Table 1 shows the male/female ratio, mean age, and current educational level of subjects within each of the five training conditions. Chi-square tests revealed that men and women were evenly distributed across conditions, $\chi 2(4) = .889$, $p = .926$ (calculated in R using the *gmodels* package (version 2.18.1, Warnes et al., 2018)). Similarly, current educational level was evenly distributed across training conditions, $\chi 2(12) = 19.50$, $p = .077$. Moreover, a one-way ANOVA revealed that there were no significant differences in subjects' mean age between conditions, $F(35.46, 84.54) = 1.23$, $p = .287$.

Table 1

Male/Female Ratio, Mean Age (in Years), and Current Educational Level of Subjects within Each of the Five Training Conditions

	Gender		**Mean Age (SD)**	**Current Educational Level[a]**			
	Male (N)	*Female (N)*	*(In years)*	*MBO (N)*	*HBO (N)*	*WO-BA(N)*	*WO-MA(N)*
AV *(N = 12)*	3	9	22.50 (2.54)	1	0	5	6
AV-B-perc *(N = 12)*	3	9	25.58 (13.31)	0	3	5	3
AV-B-prod *(N = 12)*	4	8	21.25 (1.21)	0	3	8	1
AV-M-perc *(N = 12)*	2	10	22.50 (5.33)	0	0	9	1
AV-M-prod *(N = 12)*	3	9	27.42 (12.94)	0	1	8	1
Total	15	45	23.86 (8.68)	1	7	35	12

[a]Five subjects did not provide information about educational level and were therefore not included in this table.

Table 2 shows how subjects who had followed Spanish lessons, were raised multilingually, had audio-visual impairments, and/or had dyslexia were divided across training conditions. A one-way ANOVA revealed that the mean number of languages spoken at home did not significantly differ between conditions, $F(4.8, 115.2) = 1.19$, $p = .312$. Moreover, chi-square

tests demonstrated that subjects who had followed some Spanish lessons were evenly distributed across conditions, χ 2(4) = 5.75, p = .218. The same holds for subjects who had an audio-visual impairment, χ 2(4) = 3.10, p = .541 and/or dyslexia, χ 2(4) = 2.11, p = .716. Furthermore, to control for these individual characteristics they were added as control predictors to the analysis in which all appeared to be irrelevant.

Table 2

Division of Subjects who had Followed Spanish Lessons, were Raised Multilingually, had Audio-Visual Impairments, and/or Dyslexia across the Five Conditions

	Spanish Lessons		**Mean Number of Languages (SD)**	**Audio-Visual Impairment**		**Dyslexia**	
	Yes (N)	*No (N)*		*Yes (N)*	*No (N)*	*Yes (N)*	*No (N)*
AV *(N = 12)*	3	9	1.08 (.28)	1	11	0	12
AV-B-perc *(N = 12)*	3	9	1.08 (.28)	0	12	0	12
AV-B-prod *(N = 12)*	4	8	1.15 (.36)	0	12	1	11
AV-M-perc (*N* = 12)	2	10	1.08 (.27)	1	11	1	11
AV-M-prod *(N = 12)*	3	9	1.17 (.55)	0	12	1	11
Total	15	45	1.11 (.37)	2	58	3	57

Subjects were recruited via the subject recruitment system of the Radboud University (SONA) and recruitment posters placed in the university (see Appendix A). The working title of the study, which was visible to subjects, was "Viva España! Crash course Spanish in 60 minutes" ("*Viva España! Stoomcursus Spaans in 60 minuten"* in Dutch). After participating, subjects could earn either one participant-hour or ten euros in Iris cheques. Subjects were instructed that during the experiment they would perform several tasks most of which were aimed at learning the Spanish language. Before starting the experiment, subjects received some general information about the experiment (see Appendix B), signed informed consent (see Appendix C), and were asked whether they had any experience with the Spanish language. Other individual characteristics, such as experience with other languages and level of education were considered in a questionnaire at the end of the experiment.

4.2 Material

To examine the role of gesture type, physical involvement level, and learner characteristics on the effectiveness of FL lexical stress training the current experiment contained several tasks: 1) read-aloud task, 2) lexical stress training, 3) musical aptitude task, 4) WM task, and 5) questionnaire.

4.2.1 Read-aloud task

In the read-aloud task subjects were instructed to read thirty-one Spanish phrases (see Appendix D) at three different time points: (1) at a pre-test (i.e., T1) before any training, (2) at an immediate post-test (i.e., T2), measuring the direct effect after a short training session, and (3) at a delayed post-test (i.e., T3), measuring the 'long-term' effects of training after a thirty to forty minute break in which other tasks were performed. All phrases contained a Spanish-Dutch cognate with either similar (i.e., filler items) or dissimilar (i.e., target items) lexical

stress in Spanish and Dutch. In total, seventeen phrases contained a target item and fourteen phrases included a filler item. The target and filler items displayed the different rules for lexical stress in Spanish. The Spanish language knows three rules for lexical stress: One, if a word ends in a vowel, or with -s or -n, stress is on the penultimate syllable. For example, in ele**FAN**te (*elephant*); Two, if a word ends in a consonant that is not -s or -n, stress is on the final syllable. For example, in profe**SOR** (*professor*); Three, if a word contains a written accent, stress is on the syllable that contains the accent. For example, in pi**RÁ**mides (*pyramids*). Table 3 shows the number of target and filler items belonging to each of the three rules. The current study differentiated between rule number three – words with a written accent, and rules number one and two – words without a written accent. Since a written accent clearly indicates where lexical stress should be placed, rule number three was hypothesised to have a different learning effect compared to rules number one and two. Therefore, in the analysis, productions on rule number one and two together are compared to productions on rule number three.

Table 3

Number of Target and Filler Items Belonging to Each of the Three Rules

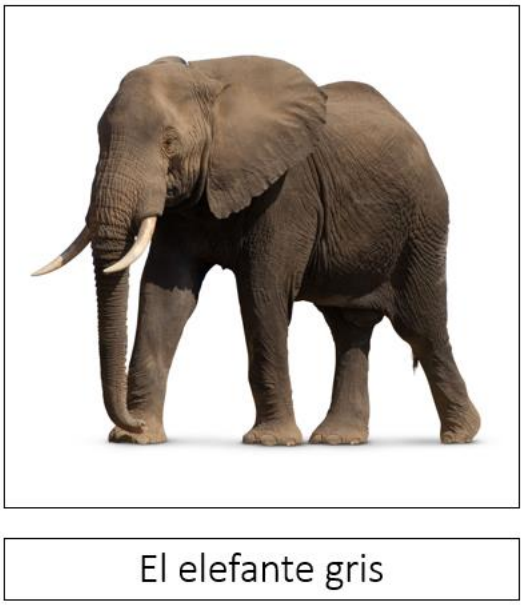
	Rule 1	**Rule 2**	**Rule 3**
Target items	4	5	8
Filler items	3	4	7

The target and filler items were always the second word of a three-word, easy to parse noun phrase (e.g., “El ele**FAN**te gris” (“*The grey elephant*”)). The noun phrases were also accompanied by a picture depicting the meaning of the phrase (see Figure 1). This picture provided some context, which enabled subjects to learn new words, and made the task more

interesting. Additionally, the pictures were used to enable the subjects to focus on the pronunciation of the new words, instead of having to focus on learning the meaning of the new words, as this would have placed additional cognitive demands on semantic processing, making the task more difficult (Zhang et al., 2020). Moreover, the current study used full noun phrases instead of single words to prevent list intonation on the target word. Another argument for using longer stretches of text compared to single words comes from Hirata (2004), who found that auditory sentence training was generalisable to isolated words, whereas auditory word training was not generalisable to the sentence context. Therefore, the use of noun phrases might have made it easier for subjects to transfer their new knowledge.

Figure 1

Example of Phrase Presentation to Subjects



The three different testing times (T1, T2, and T3) contained the same phrases but in different randomised orders to prevent subjects from learning the order of the phrases (see Appendix E). On all testing times, no more than two filler or two target items were presented in a row. At maximum, three items belonging to the same rule were presented before an item

belonging to a different rule was displayed. Moreover, T1, T2, and T3 all had two versions (version A and B). The order in version B mirrored the order in version A.

The read-aloud tasks were presented to participants in PowerPoint, showing each phrase on a separate slide. During the read-aloud tasks, the researcher moved through the slides that included the Spanish phrases, while subjects were told that the phrases occurred automatically. Subjects were allowed to repeat the phrases as often as they wanted and only their last production was analysed. The researcher could intervene when the subjects' productions were non-fluent, for instance, if the productions had too many pauses within and between words. In these cases, subjects were asked to produce the phrase again. Generally, this occurred between one to five times on each of the three testing times. However, it was noted, that especially at T2, directly after training, subjects produced too many pauses within the words. A possible explanation seems to be that subjects were trying to incorporate the newly learned rules but failed to do this fluently. The subjects' speech production during T1, T2, and T3 was recorded with a Wave/MP3 recorder (Roland R-05).

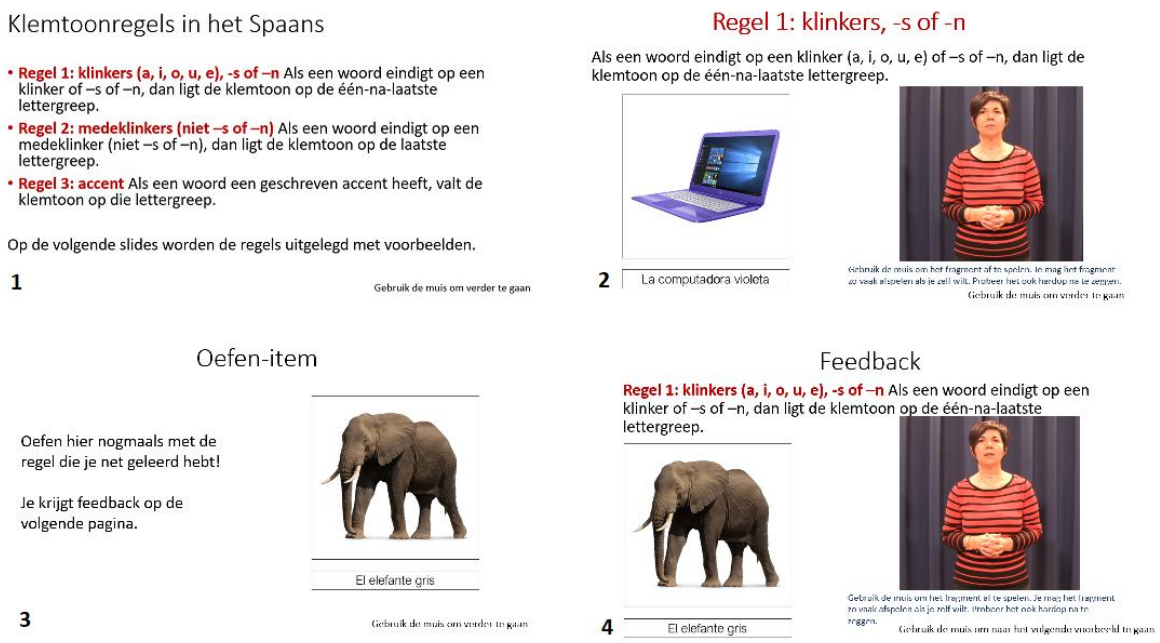
4.2.2 Lexical stress training

Between read-aloud task T1 and T2, subjects received lexical stress training in which they learned the three rules for Spanish lexical stress in one of the five training conditions: 1) AV, (2) AV-B-perc, (3) AV-B-prod, (4) AV-M-perc, and (5) AV-M-prod. The training was conducted in PowerPoint. In all training conditions, subjects were first presented with a written instruction about the three rules for lexical stress in Spanish, followed by several examples illustrating the new rules. For each rule, subjects first viewed the written rule accompanied by an example phrase, in which lexical stress was marked in bold (e.g., "La computa**do**ra vio**le**ta" (*"The purple computer"*)), and a video of a native Spanish speaker producing the example phrase (see Figure 2). After this example (which was not included in

the read-aloud task) subjects were presented with a practice item and were instructed to practice the rule out loud. Next, subjects received indirect feedback on this item with the help of a video in which the same native Spanish speaker produced the correct form. Here, the written rule was also shown once more. The practice items were also included in the read-aloud tasks, but, as subjects had practiced with them, excluded from the analysis. The practice and feedback moments were added to make subjects more familiar with the rules and to verify whether the rules were understood. A similar method with indirect feedback was used in Kelly et al. (2014).

Figure 2

Slides Lexical Stress Training



Note. Explaining the stress rules in Spanish (1), an example (2), a practice item (3), and feedback to the practice item (4) (from left to right, top to bottom) in the AV condition.

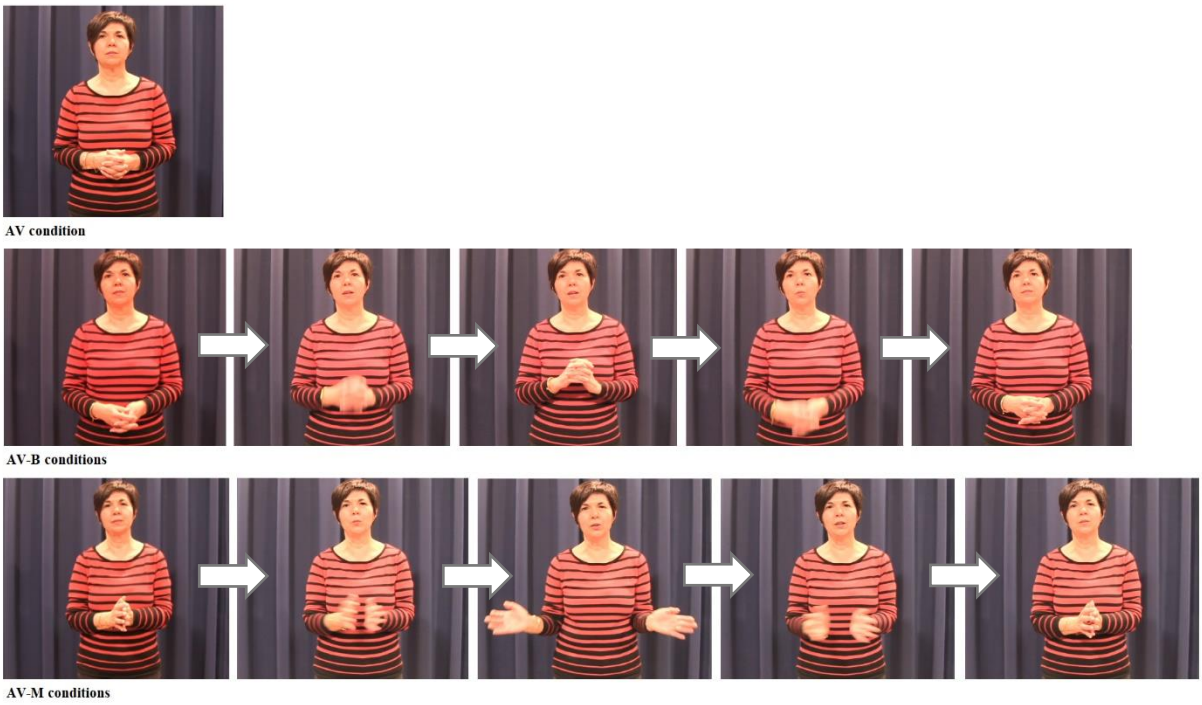
The videos presented during training contained different gesture types and required different physical involvement levels depending on the training condition subjects were in. In

the AV condition, training videos contained a native Spanish speaker producing the examples without any gestures. In the AV-B-perc and AV-B-prod condition, training videos contained the same native Spanish speaker producing the examples while making a beat gesture when lexical stress occurs in the target word. In the current study, a beat gesture was produced by having both hands together and vertically moving them (see Figure 3), by which the stroke of the beat gesture was temporally aligned with the accented syllable in the target word. In the AV-M-perc and AV-M-prod condition, the same native speaker produced a metaphoric gesture when lexical stress occurred in the target word. The metaphoric gesture that was used represented the relatively longer duration of the stressed syllables in comparison to the relatively shorter duration of unstressed syllables. In Spanish, duration is the only cue for lexical stress that does not overlap with phrasal stress and therefore seems to be most suited for teaching lexical stress (Ortega-Llebaria, 2006; Ortega-Llebaria, del Mar Vanrell, & Prieto, 2010). When producing the metaphoric gesture, the speaker held her hands in front of her and moved them apart horizontally while producing the stressed syllable (see Figure 3). The apex of the gesture (i.e., when the hands are furthest apart) was aligned with the accented syllable in the target word. All training videos were recorded with a professional HD camera.

During training, subjects were instructed to imitate the speech of the speaker in the AV and gesture perception conditions (i.e., AV-B-perc and AV-M-perc), whereas in the gesture production conditions subjects were instructed to imitate both the speech and the hand movements of the speaker (i.e., AV-B-prod and AV-M-prod). When subjects forgot to imitate the speaker, subjects were reminded once in training to imitate the speech and/or movements. During the read aloud tasks, subjects were allowed but not prompted to use gestures.

Figure 3

Video stills of training-videos in AV, AV-B, and AV-M conditions



All training videos contained the same female native Spanish speaker. Moreover, a separately recorded audio-file (by the same speaker) was dubbed over all videos to minimise audio differences between training conditions. This could have caused small, but generally invisible, differences between lip movements and audio-signal. In the metaphoric conditions, the speaker had a short pause between producing the subject and adjective in the video-recordings. Because this pause was not present in the audio-file it was added in. As a consequence, these videos were slightly longer than the videos in the other two training conditions.

4.2.3 Musical aptitude task

The musical aptitude task was abstracted from the PROMS test battery (Law & Zentner, 2012). PROMS is designed to measure perceptual musical abilities across several domains:

"tonal (melody, pitch), qualitative (timbre, tuning), temporal (rhythm, rhythm-to-melody, accent, tempo), and dynamic (loudness)" (Law & Zentner, 2012, p.1). The current experiment used a modular version of the task which included short versions of the melody, rhythm, and accent subtests. These aspects were selected because they appear to be most closely related to lexical stress. Previous studies used a modular version of PROMS to measure musical aptitude and its relation to learning FL prosody as well but differentiated in which subtests were used. Li et al. (2020) used the pitch and rhythm subtests, Yuan et al. (2018) the pitch and melody subtests, and Zhang et al. (2020) the rhythm and rhythm-accent subtests.

In the task, subjects heard the same sound fragment two times followed by a comparison sound fragment. The subjects were asked whether this comparison fragment was the same or different from the first two. In each subtest, subjects were instructed to focus on another musical aspect, respectively melody, rhythm, and accent.

4.2.4 WM task

The WM task was a backwards digit span task abstracted from the AWMA test battery (Alloway, 2007). The task was presented on the computer using *OpenSesame* (Mathôt, Schreij, & Theeuwes, 2012). In the WM task, subjects heard digit spans and were asked to repeat these spans backwards, thus beginning with the last digit they heard and ending with the first. The task started with three practice items (two of two digits and one of three digits) followed by six blocks of digit spans increasing in length (from two to seven digits). If subjects produced four (out of six) spans within one block correctly, they automatically went to the next block. However, if a subject made errors in more than three spans within one block the task ended. Subjects were scored on how many digit spans they repeated correctly. Originally, the task was developed for children, but it has also been used in studies assessing WM capacity in adults (Dunning & Holmes, 2014). In the present experiment, only one

subject reached ceiling level scores, suggesting that the task was indeed suited to differentiate between WM abilities in adults.

4.2.5 Questionnaire

At the end of the experiment, after subjects performed the read-aloud task for the third and final time, the researcher conducted a short questionnaire with the subjects (see Appendix F). The questionnaire informed the researcher about the subject's age, current educational level, language background, experience with Spanish, and musical experience. Questions about language background were added to see whether subjects who spoke multiple languages were better equipped to learn FL prosody but appeared to be irrelevant in the analysis. Subjects were asked about their experience with Spanish at the beginning of the experiment, but also in the questionnaire to get a more detailed picture (e.g., to see whether someone had stayed in a Spanish speaking country for a long period of time). Furthermore, several questions about subjects' musical experience were asked, which included questions about the number of instruments (incl. singing) someone played, the amount of musical training someone has had, and the amount of time someone has been part of a band, musical group, or choir. Moreover, subjects were asked whether they had ever received instructions from a musical conductor, who might have used gestures to show rhythm. Finally, subjects in the gesture perception conditions (i.e., A-B-perc and AV-M-perc) were asked whether they had seen the gestures in during lexical stress training and, if so, what gestures they had seen. These questions were used as a control to see whether subjects had paid attention to the gestures but appeared to be irrelevant in the analysis. This question was not asked in the gesture production conditions (i.e., A-B-prod and AV-M-prod) as subjects were explicitly told to observe and reproduce the gestures in the read-aloud training task. A similar post-experiment question was used in Kelly et al. (2017).

4.3 Procedure

Before starting data collection, a pilot experiment was conducted. The few inaccuracies and errors that became apparent during piloting were edited out before testing begun. Subjects were tested in one of the labs at Radboud University. Each subject was tested individually, and subjects were unaware of the goal of the experiment. During the experiment, subjects were seated across the table from the researcher but were unable to see the researcher as a computer screen blocked their view (see Figure 4). All experimental tasks, apart from the questionnaire, ran on a laptop which was connected to the computer screen behind which the subjects were seated (see Figure 4). In the lexical stress training and musical aptitude task, subjects were instructed to use the mouse and choose their pace themselves. The researcher controlled the pace in the other experimental tasks and was present in the lab to instruct subjects when necessary. In the lexical stress training, musical aptitude task, and WM task subjects wore headphones. The entire experimental session was audio- and video-recorded. The video-recordings were used to control for subjects' tendencies to use (natural) gestures, despite the experimental condition they were in, as in Hirata et al. (2014) (Sony Handycam – type HDR-CX210). In total, the experiment took about fifty to sixty minutes to complete (see Appendix G for stepwise protocol).

Figure 4

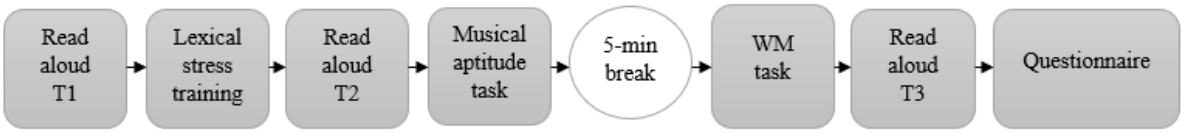
Sideview (Left) and Bird's view (Right) of Experimental Set-up

Note. In both pictures, the researcher was seated on the left and the subject on the right side of the table.

Before starting the experiment, subjects were instructed to read information about the experiment and ask questions if necessary. Subjects also needed to sign for informed consent. Thereafter, subjects were instructed that the experiment contained several tasks and lasted about sixty minutes with a short break in between. The experiment started with read-aloud task T1, lexical stress training, and read-aloud task T2. These tasks were followed by the musical aptitude task. Thereafter, subjects had a short five-minute break during which they received a candy bar and had a short conversation with the researcher. This conversation was as similar as possible with every subject (mostly about their current study or work). After the break, subjects started with the WM task followed by read-aloud task T3. The final task was the questionnaire (see Figure 5). Before each task subjects received both verbal and written instructions. Subjects were allowed to ask questions before the tasks started, but only questions related to task instructions were answered.

Figure 5

Order of the Tasks in the Experiment



4.3.1 Read-aloud task

Before starting the read-aloud task at T1, T2, and T3, subjects were instructed to read the noun phrases presented on the screen in their best Spanish. They were also told that the picture on the screen matched the meaning of the sentence, but that they did not have to learn this. Before T1, subjects were also informed that after reading the noun phrases they would receive some training and then would have to read the same phrases (in a different order) again (i.e., T2). Additionally, at T1, after receiving verbal instructions, one practice item was presented to enable the researcher to check whether the subject had understood the task and whether their production was fluent (i.e., without pauses within or between the words). Thereafter, subjects were able to read the instructions one more time. When the subject had no more questions the task started. Before T2 and T3, verbal instructions were much shorter, but subjects were reminded to think about what they had learned during training. Moreover, subjects still received written instructions and were allowed to ask questions before starting the task. The read-aloud tasks took about five minutes each to complete.

4.3.2 Lexical stress training

Lexical stress training was presented directly after read-aloud task T1. Before training started, subjects were told that they would be trained on the Spanish lexical stress rules. Subjects were also instructed to go through the slides themselves by using the mouse. Importantly, subjects received training in one of the five pre-defined training conditions: AV, AV-B-Perc, AV-B-

Prod, AV-M-Perc, or AV-M-Prod. The subjects could watch the training videos as often as they wanted, which was usually between three to six times. Subjects could also spend as much time as they wanted on the training, but generally took about five to ten minutes to complete it. Following training, subjects immediately viewed instructions for read-aloud task T2.

4.3.3 Musical aptitude task

After read-aloud task T2, subjects performed the musical aptitude task. Before starting the task, subjects were instructed to go through the task by using the mouse. At the start of each subtest, subjects received written instructions about the aspect they needed to focus on. This instruction was followed by two practice trials in which subjects received feedback by viewing the correct answers. After these practice trials, the actual task started. In total, the task lasted about ten minutes and afterwards subjects viewed the number of answers they had given correctly on each of the subtests.

4.3.4 WM task

After a short break (after the musical aptitude task), subjects performed the WM task. Before the task started, subjects were told that they would hear several series of digits and that their task was to say these digits backwards. For example, if they heard 5 – 9, they would have to say 9 – 5. The subjects were also instructed that the task would start simple with two digits, but that the number of digits would soon increase. The task took about ten minutes to complete.

4.3.5 Questionnaire

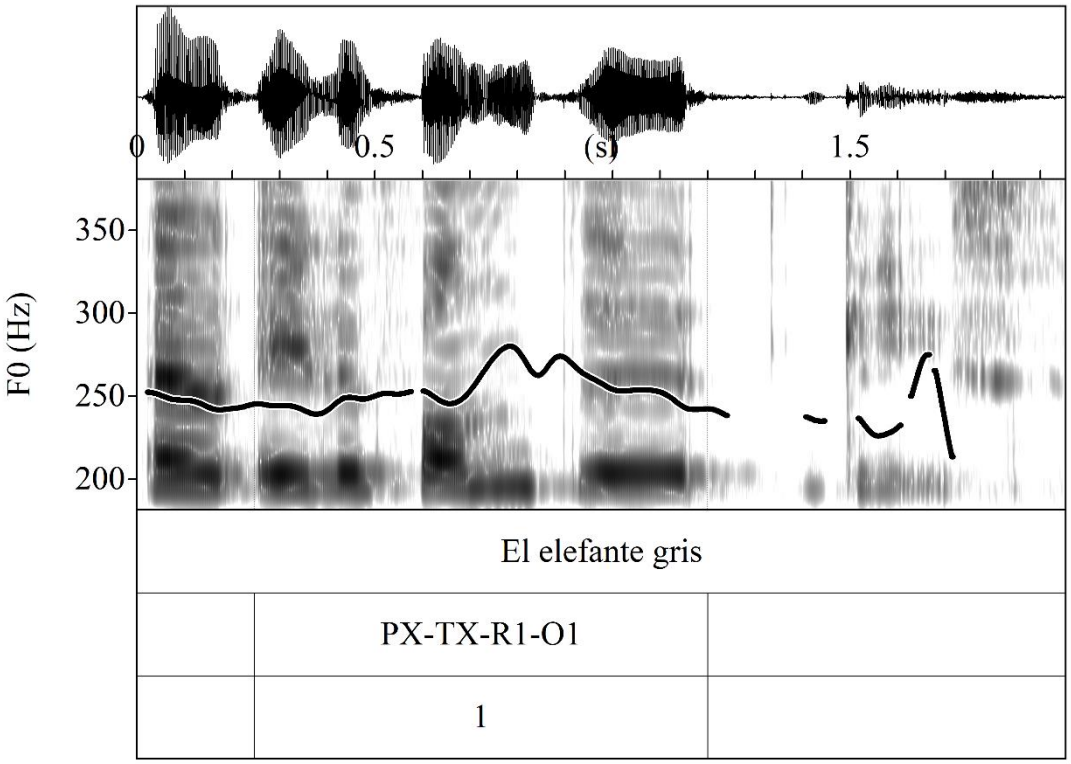
Following read-aloud task T3 (which was conducted after the WM task), subjects completed a questionnaire together with the researcher. By conducting the questionnaire together, the researcher was able to ask for elaboration on some answers given by the subjects. Conducting the questionnaire took about five to ten minutes and marked the end of the experiment.

4.4 Coding read-aloud data

Subjects' speech productions at T1, T2, and T3 were coded and annotated in *Praat* (Broersma & Weenink, 2018). First, all speech productions were coded at phrase level and all target items were subsequently coded at word level by visually and audibly inspecting the sound fragments (see Figure 6). The used codes included information about subject-number (P1, P2, etc.), time of testing (T1, T2, or T3), stress rule (R1, R2, or R3), and item number (E1, E2, etc.) in that order. For instance, P1_T1_R1_E1 refers to the production of subject number one, at the first testing time, of an item belonging to rule number one, and that item being item number one (see Appendix D for all codes). Second, subjects' productions on the target items were randomly annotated on whether they followed correct Spanish lexical stress (1) or not (0) at word level (see Figure 6). In order to do so, several different *Praat* scripts were used: (1) a script that extracted all target words from the larger sound files and made separate .wav files of them (Crosswhite, n.d.), (2) a script that randomly presented these .wav files and created textgrids for each of them making it possible to annotate the files (Phonetics Laboratory, n.d.), and (3) a script that extracted information from these textgrids and created one large datafile (Lennes, 2003) (see Appendix H for scripts).

Figure 6

Example Coding Speech Productions in Praat



To check for inter-rater reliability (calculated in R (version 2.18.1, Warnes et al., 2018) using *irr* package (version 0.84.1, Gamer, 2019)), ten percent of the data, which included items belonging to all rules and at all testing times, was annotated by two independent raters. At first, inter-rater reliability indicated only moderate agreement between the two raters, $\kappa = .55$, $p < .001$, which was considered too low for this experiment. Therefore, the items on which both raters differentiated were annotated again and when in doubt discussed between them. This resulted in an almost perfect inter-rater reliability, $\kappa = .987$, $p < .001$. Therefore, the other ninety percent of the data was annotated by one of the raters and items that were unclear were marked and later discussed with the other rater to reach agreement.

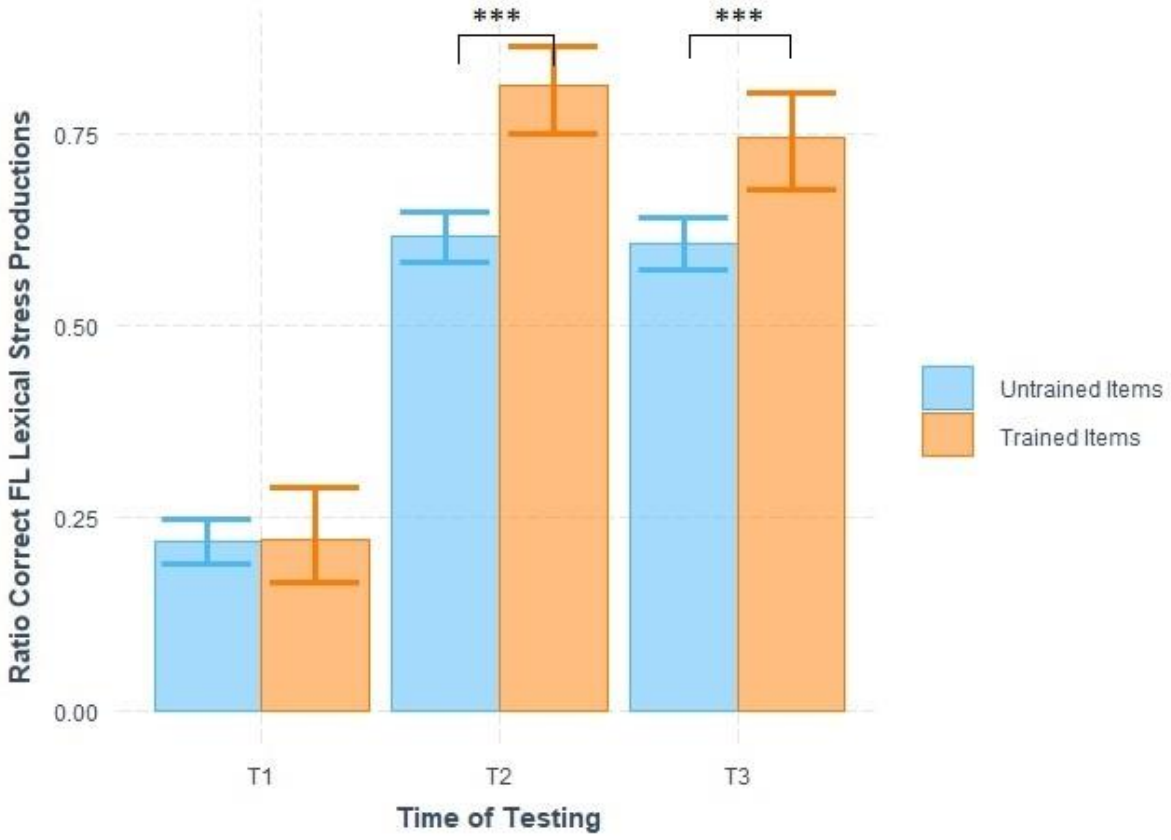
4.5 Analysis

To explore how different factors, such as gesture type, physical involvement level, and learner characteristics, influence the effectiveness of FL lexical stress training a multilevel logistic regression analysis was conducted in R and R studio (version 1.3.1093, RStudio Team, 2020) using the *lme4* package (version 1.1-26, Bates et al., 2020) (see Appendix I for scripts). The data was also plotted in R using the *ggplot* (version 3.3.3, Wickham et al., 2020), *gmodels* (version 2.18.1, Warnes et al., 2018), and *interactions* (version 1.1.3, Long, 2020) packages. The outcome variable was whether Spanish lexical stress productions were correct or incorrect on the read-aloud tasks.

Since the read-aloud tasks included three items that were practiced during lexical stress training, it was explored whether subjects' performances significantly differed on these items compared to the items that were not in training. A simple logistic regression analysis revealed that before training subjects' performances did not differentiate between the trained and untrained items, whereas after training (at T2 and T3) subjects performed significantly better on the trained items compared to the untrained items, respectively $b = .98$, $SE = .29$, $p < .001$ and $b = .62$, $SE = .27$, $p < .001$ (see Figure 7). Therefore, the practice items were excluded from the analysis. Moreover, as nine subjects did not meet the pre-defined requirements, it was explored whether these subjects performed significantly differently on the read-aloud tasks compared to subjects who did meet requirements. Chi-square tests indicated that performance was not significantly different in subjects who did not meet the pre-defined experimental criteria compared to subjects who did meet them, $\chi 2(1) = .76$, $p = .393$. Therefore, all subjects were kept in the analysis.

Figure 7

Ratio of Correct FL Lexical Stress Productions by Trained vs. Untrained Items Separated by Time of Testing



Note. Error bars represent 95% confidence intervals. ***: $p < .001$.

The main predictors that were included in the analysis were: *time of testing* (i.e., T1, T2, or T3), *training condition* (i.e., AV, AV-B-perc, AV-B-prod, AV-M-perc, AV-M-prod), *stress rule* (i.e., whether the target word included a written accent or not), *melody score* ($M = 5.09$, $SD = 1.76$), *rhythm score* ($M = 4.58$, $SD = 1.31$), *accent score* ($M = 5.14$, $SD = 1.31$), and *WM score* ($M = 4.43$, $SD = 1.03$) (see Table 4).

Table 4

Means and Standard Deviations of Subjects' Melody, Rhythm, Accent, and WM Scores

	N	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
Melody score	60	1.5	9	5.09	1.76
Rhythm score	60	1.5	7	4.58	1.31
Accent score	60	1.5	8	5.14	1.31
WM score	60	2	7	4.43	1.03

The predictors *melody score*, *rhythm score*, and *accent score*, were all extracted from the musical aptitude task. This task gave scores for each of its three subtests (i.e., melody (0 – 10), rhythm (0 – 8), and accent (0 – 10) and a total musical aptitude score (0 – 28)). Cronbach's alpha was calculated to see whether each subtest should be entered separately in the analysis or whether the total score could be used (calculated in R using the *psych* package (version 2.0.12, Revelle, 2020)) and revealed that the three subtests together had good reliability, $\alpha = .71$, indicating that they all contributed to measuring musical aptitude. Moreover, reliability did not improve by deleting any of the subtests. However, correlations of each of the subtests with the other two only indicated moderate relations, for melody $r = .52$, for rhythm $r = .51$, and for accent $r = .56$. Therefore, separate scores for each of the subtests were used in the analysis. Furthermore, the predictor *WM score* was extracted from the WM task in which scores were given based on the final block subjects had reached. Subjects received a score between 0 (no blocks completed) and 7 (all blocks completed). The other three predictors, *time of testing*, *training condition*, and *stress rule*, were all extracted from the read-aloud tasks.

Data from the questionnaire was used to control for several other subject characteristics. From these control predictors, *gender*, *Spanish lessons*, *music lessons*, *music group*, *audio-visual impairments*, *dyslexia*, *current educational level*, and *registered gestures* (i.e., whether or not the subjects noticed the gestures in the gesture perception conditions) were considered to be fixed factors (see Table 5), as they only had two values[8] (i.e., yes or no), whereas *age* (*M* = 23.86, *SD* = 8.68) and *number of languages at home* (*M* = 1.11, *SD* = 0.37) were considered to be random (see Table 6). These predictors were thought to be control predictors as they are not related to the main research questions but were added to the regression model to account for the variation within these predictors. Moreover, *item* was included as a random factor that was nested under the fixed factor *stress rule*.

Table 5

Frequencies and Percentages of Control Predictors

		Frequency	**Percent (%)**
Gender	*Male*	45	75
	Female	15	25
Spanish lessons	*Yes*	9	15
	No	51	85
Music group	*Yes*	5	8.33
	No	55	91.67
Music lessons	*Yes*	35	58.33
	No	25	41.67
Audio-visual impairments	*Yes*	2	3.33

[8] According to Grace-Martin (2019), a random factor should always include more than two values as if it has only two values the variance cannot be calculated. In these cases a factor should be fixed.

		Frequency	Percent (%)
	No	58	96.67
Dyslexia	*Yes*	3	5
	No	57	95
Current educational level	*No student*	8	13.33
	HBO	6	10
	WO Bachelor	36	60
	WO Master	10	16.67
Registered gestures	*Yes*	9	37.5
	No	15	62.5

Note. Calculated in R, using the *psych* package (version 2.0.12, Revelle, 2020).

Table 6

Means and Standard Deviations of Control Predictors

	N	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
Age	60	17	67	23.86	8.68
Number of languages at home	60	1	3	1.11	0.37

Note. Calculated in R, using the *psych* package (version 2.0.12, Revelle, 2020).

The first baseline model included all eleven control predictors. However, as two of the three random factors explained almost no variance (i.e., *age* and *number of languages at home*) the model was overfitted. Therefore, a new baseline model excluding these random factors was created. This model revealed that, from the fixed control factors, only whether a

subject had been part of a music group was a significant predictor, b = .78, SE = .31, p = .012. Since the model was complex, and to keep the model interpretable, only the significant predictor *music group* was added as a fixed factor to the final model. Hence, the final baseline model includes the fixed factor *music group* and the nested random factor *item* (see Appendix J, Baseline Model). Next, a stepwise forward procedure was used to add hypothesised main effects and interactions to this baseline model (see results).

Concerning the assumptions of multilevel logistic regression, multicollinearity was reported for the predictors *stress rule* and *item* as their VIF values were above 10 and tolerance statistics were below 0.2 (calculated in R using the *rms* package (version 6.2-0, Harrell, 2021)). Therefore, the assumption of multicollinearity appears to be violated in the dataset. Moreover, the significant effect found between the continuous predictor *WM score* and the logit of the outcome variable suggested that the assumption of linearity was not met for this predictor, b = 1.25, SE = .62, p = .043. The other continuous predictors (i.e., *melody score*, *rhythm score*, and *accent score*) demonstrated no problems with linearity. Furthermore, the data reveals that the residual deviance (2415.8) was smaller than the residual degrees of freedom (2463), which suggests that there are no signs of overdispersion in the data. Finally, multilevel regression can handle dependent errors and missing data. In short, in this sample the assumptions of linearity and multicollinearity were violated. Therefore, the continuous predictors were centred, and the data will be interpreted with caution.

5. Results

5.1 Building the regression model

To explore which factors influence the effectiveness of gestures in learning FL prosody a multilevel logistic regression was performed. The analysis started with a baseline model including *item* as a random factor, which was nested under the fixed factor *stress rule*, and the fixed factor *music group*. In the baseline model the estimated variance associated with the nested random intercept was 1.57 (*SD* = 1.25) and the residual deviance was 2946.7 (see Appendix J, Baseline Model).

Initially, all hypothesised main predictors (i.e., *time of testing*, *training condition*, *melody score*, *rhythm score*, *accent score*, *WM score*, and *stress rule*) were added to the baseline model. This model revealed significant main effects for *time of testing*, *training condition*, *melody score*, *WM score*, and *stress rule* and reduced the likelihood ratio with 265.4, against 11 degrees of freedom which was highly significant, $p < .001$. Predictors that were non-significant, and explained little variance, were removed from the model (i.e., *rhythm score* and *accent score*), which, in comparison to the model with all main predictors, resulted in a neglectable increase in likelihood ratio (1.8), against 2 degrees of freedom which was not significant, $p = .252$. Therefore, the final model excluded these non-significant predictors, and was a significant improvement ($p < .001$) on the baseline model as it reduced the likelihood ratio with 264, against 9 degrees of freedom (see Appendix J, Model 1).

Next, all predicted two-way interactions were added to Model 1. These were the interactions between:

1. *time of testing* and *training condition*;
2. *time of testing* and *WM score*;
3. *time of testing* and *melody score*;
4. *time of testing* and *rhythm score*;

5. *time of testing* and *accent score;*
6. *time of testing* and *stress rule*.

Adding these interactions to the model reduced the likelihood ratio with 26.4, against 20 degrees of freedom which was highly significant, $p < .001$. However, only one interaction term appeared to be a significant predictor (i.e., *time of testing* * *stress rule*) and the main predictors *training condition* and *melody score* were no longer significant. The non-significant interaction terms were stepwise removed from the model, starting with the interactions explaining the least variance. After removing the interactions between *time of testing* and *WM score*, *time of testing* and *melody score*, and *time of testing* and *accent score*, the interactions between *time of testing* and *rhythm score* and between *time of testing* and *training condition* became significant, and were therefore kept in the model. Compared to the model with all predicted three-way interactions, removing the three non-significant interaction terms resulted in a small increase in likelihood ratio (2.3), against 8 degrees of freedom which was non-significant, $p = .789$. Therefore, the final model excluded these non-significant interaction terms, and was a significant improvement compared to Model 1 ($p < .001$) as it reduced the likelihood ratio with 24.1, against 12 degrees of freedom (see Appendix J, Model 2).

Next, all predicted three-way interactions were added to Model 2. These were the interactions between:

1. *time of testing*, *training condition*, and *stress rule;*
2. *time of testing*, *training condition*, and *WM score;*
3. *time of testing*, *training condition*, and *melody score;*
4. *time of testing*, *training condition*, and *rhythm score;*
5. *time of testing*, *training condition*, and *accent score;*
6. *time of testing*, *stress rule*, and *WM score;*

7. *time of testing*, *stress rule*, and *melody score;*
8. *time of testing*, *stress rule*, and *rhythm score;*
9. *time of testing*, *stress rule*, and *accent score.*

Adding these three-way interactions to the model reduced the likelihood ratio with 86.3, against 80 degrees of freedom which was highly significant, $p < .001$. However, not all three-way interactions were significant (i.e., *time of testing * training condition * stress rule*, *time of testing * WM score * stress rule*, and *time of testing * accent score * stress rule*). Moreover, the two-way interactions *time of testing * training condition* and *time of testing * melody score* were no longer significant. To minimize model complexity, the non-significant interaction terms were excluded from the model which, in comparison to the model including all predicted three-way interactions, increased the likelihood ratio with 7, against 18 degrees of freedom which was non-significant, $p = .727$. Therefore, the final model excluded the non-significant three-way interactions, and was a significant improvement compared to Model 2 ($p < .001$) as it reduced the likelihood ratio with 79.3, against 62 degrees of freedom (see Appendix J, Model 3).

Finally, all predicted four-way interactions were added to Model 3. These included the interactions between:

1. *time of testing*, *training condition*, *WM score*, and *melody score;*
2. *time of testing*, *training condition*, *WM score*, and *rhythm score;*
3. *time of testing*, *training condition*, *WM score*, and *accent score;*
4. *time of testing*, *training condition*, *stress rule*, and *WM score;*
5. *time of testing*, *training condition*, *stress rule*, and *melody score;*
6. *time of testing*, *training condition*, *stress rule*, and *rhythm score;*
7. *time of testing*, *training condition*, *stress rule*, and *accent score*

Adding these four-way interactions to the model reduced the likelihood ratio with 128.7, against 99 degrees of freedom which was highly significant, $p < .001$. However, not all four-way interactions were significant (i.e., *time of testing * training condition * stress rule * rhythm score*). Moreover, the three-way interaction between *time of testing*, *stress rule*, and *accent score* was no longer significant. Additionally, in comparison to Model 2 and 3, the main effect of *training condition* became significant again. As four-way interactions are difficult to interpret, the non-significant interaction term was excluded from the model, which, in comparison to the model with all hypothesised four-way interactions, increased the likelihood ratio with 6.84, against 12 degrees of freedom which was a non-significant difference, $p = .322$. Therefore, the final model[9] excluded the non-significant four-way interaction, and was a significant improvement compared to Model 3 ($p < .001$) as the likelihood ratio was reduced with 121.9, against 87 degrees of freedom (see Appendix J, Model 4).

5.2 Interpreting the model

There was a significant learning effect as *time of testing* significantly predicted how likely subjects were to produce correct FL lexical stress (see Table 7). Subjects were significantly more likely to produce correct FL lexical stress at T2 (61.60% correct) compared to at T1

[9] ~ TimeofTesting + Condition + Melody + WM + StressRule + TimeofTesting:Condition
+ TimeofTesting:Melody + TimeofTesting:StressRule + TimeofTesting:Condition:WM
+ TimeofTesting:Condition:Melody + TimeofTesting:Condition:Rhythm + TimeofTesting:Condition:Accent
+ TimeofTesting:StressRule:Melody + TimeofTesting:StressRule:Rhythm
+ TimeofTesting:Condition:WM:Melody + TimeofTesting:Condition:WM:Rhythm
+ TimeofTesting:Condition:WM:Accent + TimeofTesting:Condition:StressRule:WM
+ TimeofTesting:Condition:StressRule:Melody + TimeofTesting:Condition:StressRule:Accent
+ Musicgroup + (1 | StressRule:Item)

(21.85% correct), $b = 2.52$, $SE = .76$, $p < .001$ (see Figure 8). Similarly, subjects were significantly more likely to produce correct FL lexical stress at T3 (60.7% correct) compared to at T1, $b = 3.35$, $SE = .75$, $p < .001$. Bonferroni corrected contrast analyses revealed that there were no significant differences in subjects' performances between T2 and T3, $b = -.02$, $SE = .27$, $p = 1.00$ (see Figure 8). Therefore, subjects, across training conditions, seem to improve on their FL lexical stress productions after they have received training.

Table 7

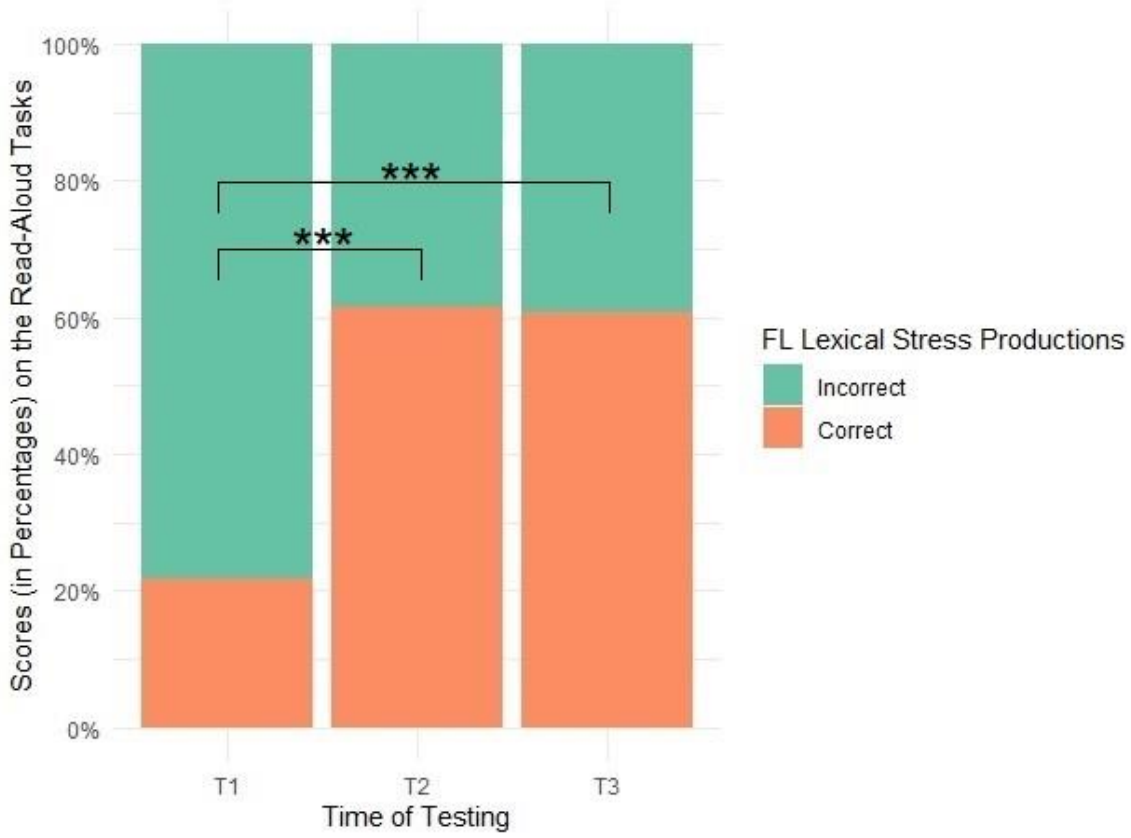
Estimated Effects and Coefficients for Main Effects of Time of Testing, Training Condition, and Stress Rule

Predictor	***b* estimate**	**Std. Error**	***z* value**	***p* value**
T2	2.52	.76	-1.45	<.001
T3	3.35	.75	3.34	<.001
AV-B-prod	-1.95	.94	-2.09	.037
Stress Rule	-3.88	1.08	-3.58	<.001

Note. Table only includes significant main effects from the regression model.

Figure 8

Accuracy Score (in Percentages) on FL Lexical Stress Productions by Time of Testing



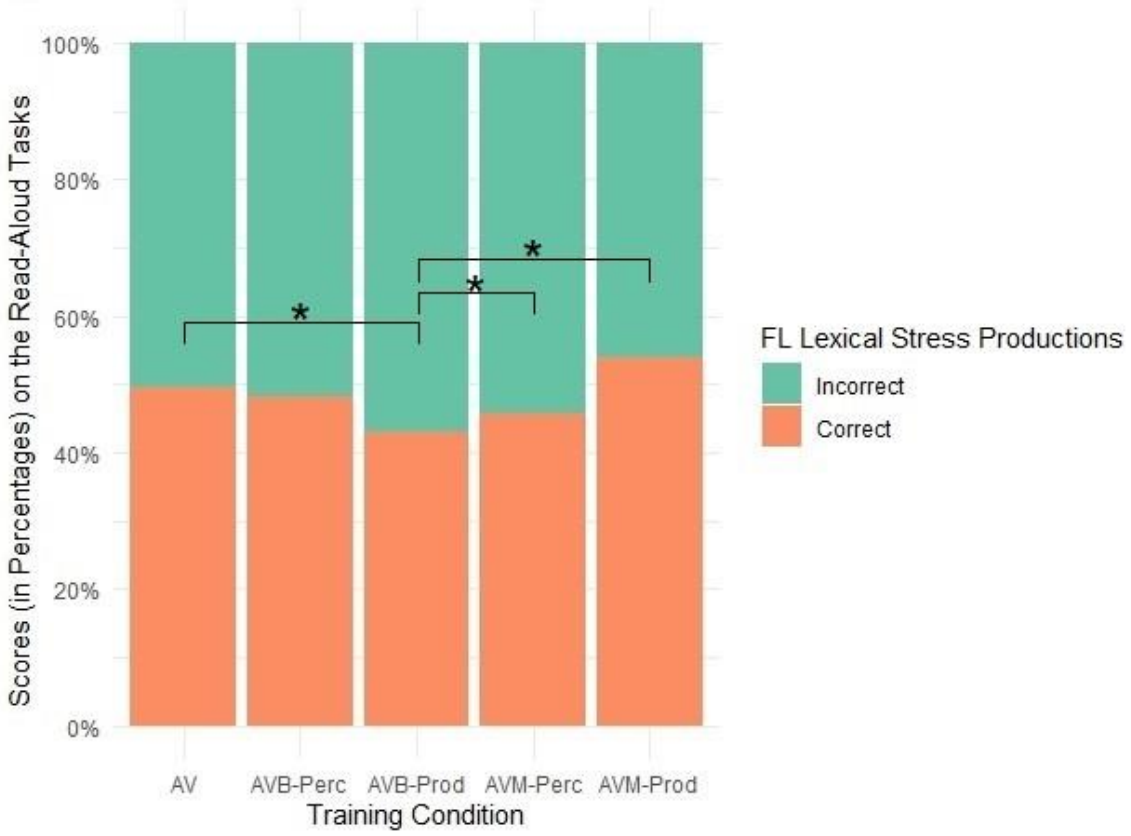
Note. ***: $p < .001$.

Next, there was a significant main effect of *training condition* (see Table 7), which indicates that, irrespective of testing times, subjects were significantly less likely to produce correct FL lexical stress in the AV-B-prod condition (42.89% correct) compared to the AV condition (49.60% correct), $b = -1.95$, $SE = .94$, $p = .037$ (see Figure 9). Bonferroni corrected contrast analyses revealed that subjects in the AV-B-prod condition were also significantly less likely to produce correct FL lexical stress compared to the AV-M-perc (45.78% correct) and AV-M-prod (53.92% correct) condition, respectively $b = -1.34$, $SE = .41$, $p = .011$ and $b = -1.52$, $SE = .46$, $p = .010$. It should be noted that, this main effect of *training condition* reflects performances at all three testing times, and since subjects had not yet received

training at T1, these findings do not provide conclusive information about the effectiveness of the different training conditions. Moreover, although subjects in the AV-B-prod condition were significantly less likely to produce correct FL lexical stress compared to subjects in the other training conditions, figure 9 shows that the differences between training conditions were small.

Figure 9

Accuracy Score (in Percentages) on FL Lexical Stress Productions per Training Condition Collapsing over Time of Testing



Note. *: $p < .05$.

Thirdly, the regression analyses revealed that *stress rule* significantly predicted performance on FL lexical stress production (see Table 7), as subjects were significantly more

likely to produce correct FL lexical stress in words with a written accent (64.32%) compared to in words without a written accent (35.68%), b = -3.88, SE = 1.08, p < .001. This main effect was quantified by a significant interaction effect between *stress rule* and *time of testing* (see Table 8), indicating that the differences in performance on FL lexical stress production between words with and without a written accent were significantly different at T1 compared to at T2, b = 2.46, SE = .76, p = .001, and T3, b = 2.63, SE = .76, p < .001.

Table 8

Estimated Effects and Coefficients for the Interaction between Time of Testing and Stress Rule

Predictor	***b* estimate**	**Std. Error**	***z* value**	***p* value**
*T2 * Stress Rule*	2.46	.76	3.24	.001
*T3 * Stress Rule*	2.63	.76	3.46	<.001

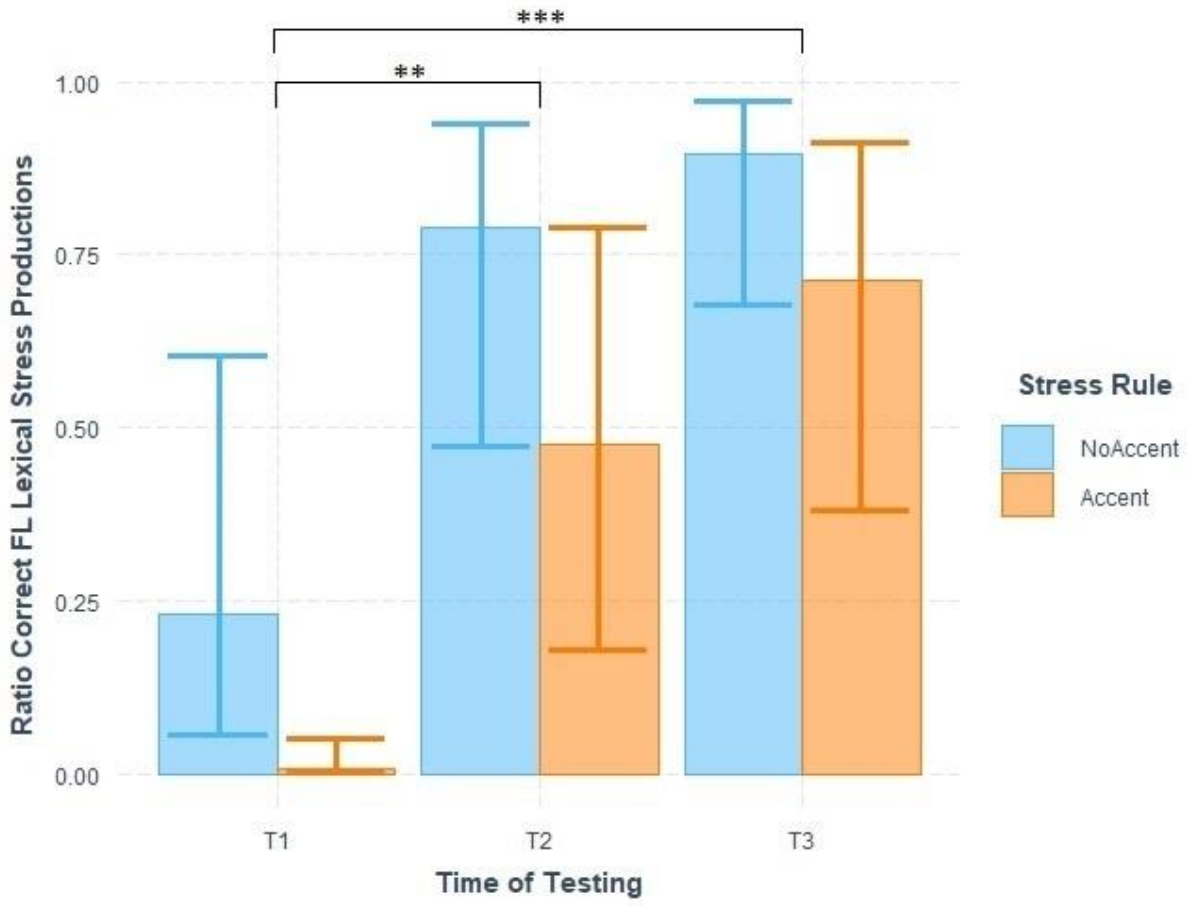
Note. Table only includes significant interactions from the regression model.

Bonferroni corrected contrast analyses revealed that only at T1, and not at T2 and T3, FL lexical stress was significantly less often produced correctly in words with a written accent compared to in words without a written accent, b = 3.88, SE = 1.08, p < .001 (see Figure 10). Hence, the differences in performance on FL lexical stress production caused by the different stress rules disappeared after lexical stress training. Moreover, this effect was similar in all training conditions as the interaction between *time of testing*, *training condition*, and *stress rule* was non-significant[10].

[10] It is unknown whether the two-way interaction between *stress rule* and *training condition* was significant as this interaction was not included in the regression model.

Figure 10

Ratio of Correct FL Lexical Stress Productions by Time of Testing



Note. "No Accent" refers to words without a written accent, and "Accent" refers to words with a written accent. Error bars represent 95% confidence intervals. ***: $p < .001$, **: $p < .01$.

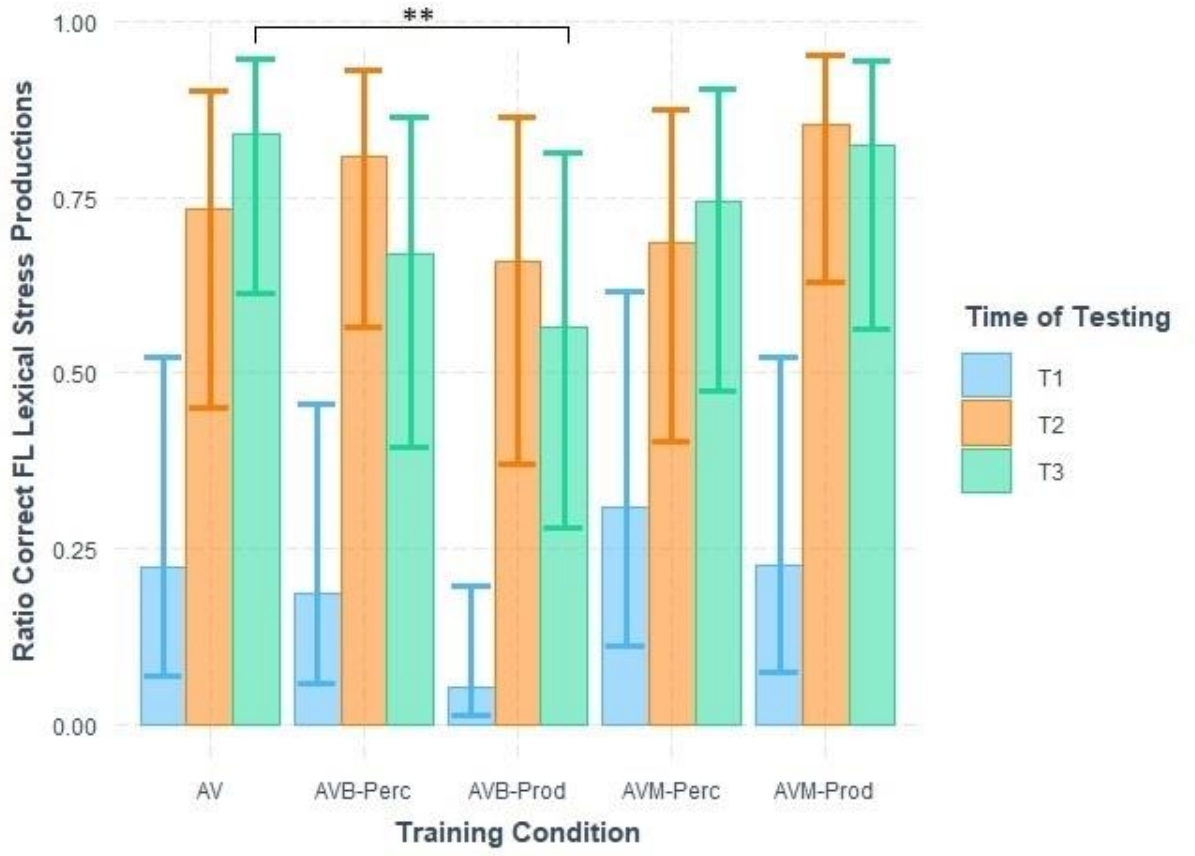
5.2.1 The influence of gesture type and physical involvement level

The interaction between *time of testing* and *training condition* was non-significant, indicating that the differences in performance on FL lexical stress production between T1, T2, and T3 cannot be explained by the different training conditions. More specifically, there were no significant differences in improvement from T1 to T2, from T2 to T3, and from T1 to T3 between training with or without gestures, between having beat or metaphoric gestures in

training (i.e., gesture type), and between producing or perceiving gestures in training (i.e., physical involvement level) (see Figure 11). Bonferroni corrected contrast analyses revealed that the main effect of *time of testing* holds true, in that subjects in all training conditions significantly improved on their FL lexical stress productions from T1 to T2, and from T1 to T3, whereas the difference in performance between T2 and T3 was always non-significant.

Figure 11

Ratio of Correct FL Lexical Stress Productions per Training Condition Separated by Time of Testing



Note. Error bars represent 95% confidence intervals. **: $p < .01$.

Contrast analyses indicated that, at T3, subjects in the AV condition were significantly more likely to produce correct FL lexical stress compared to subjects in the AV-B-prod condition, *b*

= 2.22, *SE* = .63, *p* = .005. However, this finding does not necessarily indicate that producing beat gestures in training was less effective in learning to produce FL lexical stress compared to having no gestures in training, as overall performance (i.e., at T1, T2, and T3) of subjects in the AV-B-prod condition was significantly worse compared to overall performance in all other training conditions (see main effect *training condition*; see Figure 11). Although almost no significant differences between training conditions can be reported, a visual inspection of the data revealed some interesting trends which will be considered in the discussion.

5.2.2 The influence of learner characteristics

While a visual inspection of the data suggested that higher WM and musical aptitude scores predicted slightly more correct FL lexical stress productions, main effects of *WM*, *melody*, *rhythm*, and *accent scores* were all non-significant, indicating that subjects' WM capacity and musical aptitude did not significantly influence FL lexical stress productions (see Table 9). The two-way interactions between *time of testing* and *WM*, *melody*, *rhythm*, or *accent scores* were also non-significant, indicating that, across training conditions, subjects' improvements on producing FL lexical stress were independent of their WM capacities or musical aptitudes.

Table 9

Mean (Standard Deviation) WM, Melody, Rhythm, and Accent Scores Split between Incorrect and Correct FL Lexical Stress Productions Collapsed over Time of Testing

	FL Lexical Stress Productions	
	Incorrect *Mean (SD)*	Correct *Mean (SD)*
WM score	4.34 (1.03)	4.52 (1.02)
Melody score	4.97 (1.74)	5.24 (1.79)
Rhythm score	4.51 (1.32)	4.67 (1.29)
Accent score	5.08 (1.30)	5.22 (1.31)

There was a significant three-way interaction between *time of testing*, *stress rule*, and *melody score* (see Table 10) and between *time of testing*, *stress rule*, and *rhythm score* (see Table 10), suggesting that the relation between *time of testing* and *melody score* and between *time of testing* and *rhythm score* and their influence on producing correct FL lexical stress varied between words with and without a written accent.

Table 10

Estimated Effects and Coefficients for the Interactions between Time of Testing, Stress Rule and Melody or Rhythm Score

Predictor	***b* estimate**	**Std. Error**	***z* value**	***p* value**
*T2 * Melody score * Stress rule*	1.50	.58	2.59	.010
*T3 * Rhythm score * Stress rule*	-.66	.24	-2.83	.005

Note. Table only includes significant interactions from the regression model.

In the interaction with melody aptitude, at T2, as opposed to at T1 and T3, melody scores influenced subjects' productions on FL lexical stress significantly differently depending on stress rule, $b = 1.50$, $SE = .58$, $p = .010$. Namely, at T2, in words without a written accent, a lower melody scores predicted better performance, and in words with a written accent, higher melody scores predicted better performance, whereas at T1 and T3, in both words with and without a written accent, higher melody scores predicted more correct FL lexical stress productions (see Figure 12).

Figure 12

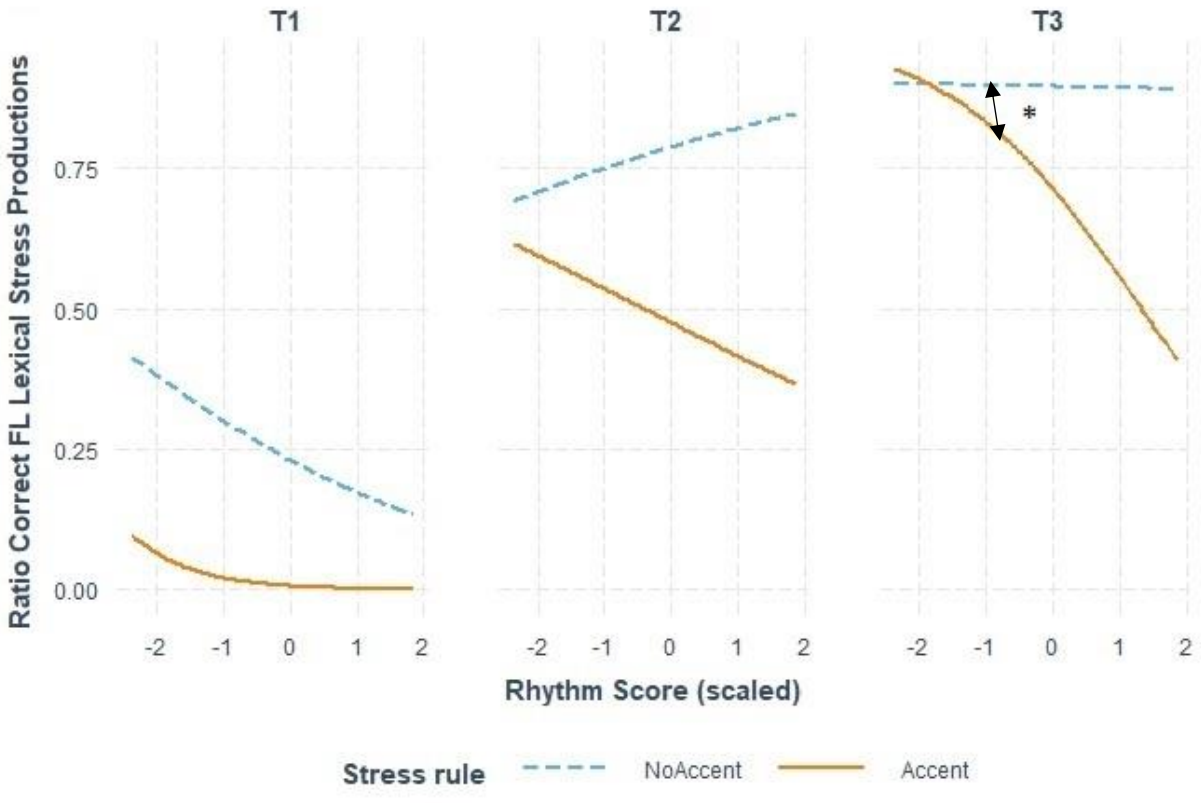
The Influence of the Relation Between Time of Testing and Melody Score on the Ratio of Correct FL Lexical Stress Productions Separated by Stress Rule

Note. Higher melody scores (scaled) indicate higher melody aptitude. "No Accent" refers to words without a written accent, and "Accent" refers to words with a written accent. *: $p < .05$.

In the interaction with rhythm aptitude, at T3, as opposed to at T1 and T2, rhythm scores influenced subjects' FL lexical stress productions significantly differently depending on stress rule, b = -66, SE = .24, p = .005. Namely, at T3, in words with a written accent, lower rhythm scores predicted better performance, and in words without a written accent, rhythm scores did not seem to have a predictive function (see Figure 13), whereas at T1 and T2, there was no significant difference in correct FL lexical stress productions between words with and without a written accent.

Figure 13

The Influence of the Relation Between Time of Testing and Rhythm Score on the Ratio of Correct FL Lexical Stress Productions Separated by Stress Rule



Note. Higher rhythm scores (scaled) indicate higher rhythm aptitude. "No Accent" refers to words without a written accent, and "Accent" refers to words with a written accent. *: $p < .05$.

The three-way interactions between *time of testing* and *training condition* on the one hand and *WM* or *musical aptitude scores* on the other hand were also significant, suggesting that WM capacity and musical aptitude influenced the effectiveness of the different training conditions over time. First, there was a significant three-way interaction between *time of testing*, *training condition*, and *WM score* (see Table 11), which suggests that the relation between training condition and WM capacity varied across testing times.

Table 11

Estimated Effects and Coefficients for Interactions between Time of Testing, Training Condition, and WM, Melody, Rhythm, or Accent score

Predictor	***b* estimate**	**Std. Error**	***z* value**	***p* value**
*T1 * AV-B-prod * WM score*	-2.60	1.20	-2.16	.031
*T1 * AV-M-prod * WM score*	2.54	1.25	2.03	.042
*T2 * AV-B-prod * Melody score*	4.18	1.33	3.14	.002
*T3 * AV-B-prod * Melody score*	3.94	1.38	2.86	.004
*T2 * AV-B-prod * Rhythm score*	1.13	.36	3.15	.002
*T3 * AV-B-prod * Rhythm score*	2.00	.43	4.61	< .001
*T2 * AV-M-prod * Rhythm score*	1.24	.48	2.57	.001
*T3 * AV-M-prod * Rhythm score*	1.52	.54	2.83	.005
*T3 * AV-B-prod * Accent score*	-1.98	.69	-2.87	.004
*T1 * AV-M-prod * Accent score*	2.02	.91	2.22	.026

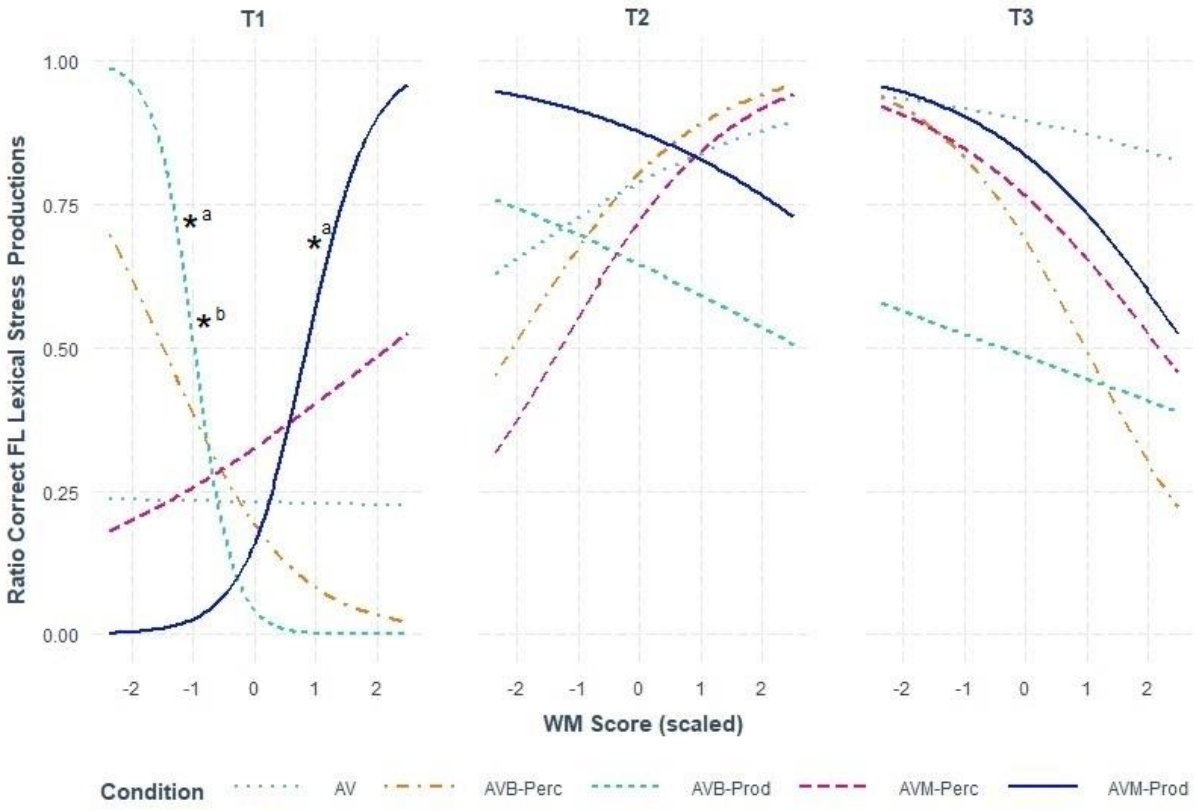
Note. Table only includes significant three-way interactions from the regression model.

At T1, in the AV-B-prod and AV-M-prod condition, the influence of WM capacity on producing correct FL lexical stress was significantly different from its influence at other times

of testing, respectively b = -2.60, SE = 1.20, p = .031 and b = 2.54, SE = 1.25, p = .042. Bonferroni corrected contrast analyses indicated that, in the AV-B-prod condition, the influence of WM capacity on producing correct FL lexical stress was significantly different at T1 compared to at T2, b = -3.35, SE = 1.07, p = .005, and at T3, b = -3.32, SE = 1.09, p = .007. Namely, whereas at T1 subjects with lower WM scores performed better on FL lexical stress production than subjects with higher WM scores, at T2 and T3 this difference was much less clear, though in the same direction (see Figure 14). Concerning the significant interaction in the AV-M-prod condition, as can be seen in figure 14, whereas at T1 subjects with higher WM scores performed better on FL lexical stress production than subjects with lower WM scores, at T2 and T3 this difference was much less clear and in the opposite direction.

Figure 14

The Relation Between Training Condition and WM Score on the Ratio of Correct FL Lexical Stress Productions Separated by Time of Testing



Note. Higher WM scores (scaled) indicate better WM capacity. *: $p < .05$.

[a]Predictions in the AV-M-prod and AV-B-prod condition at T1 were significantly different than at T2 and T3. [b]At T1, predictions in the AV-B-prod condition were significantly different than in AV, AV-M-perc, and AV-M-prod condition.

Contrast analyses were used to explore how the relation between *time of testing* and *WM score* varied across *training conditions*, and revealed that, at T1, the relation between WM capacity and performance on FL lexical stress production was significantly different in the AV-B-prod condition compared to the AV, $b = 3.73$, $SE = 1.20$, $p = .019$, AV-M-perc, $b = -4.37$, $SE = 1.10$, $p < .001$, and AV-M-prod condition, $b = -5.51$, $SE = 1.43$, $p = .001$. Namely,

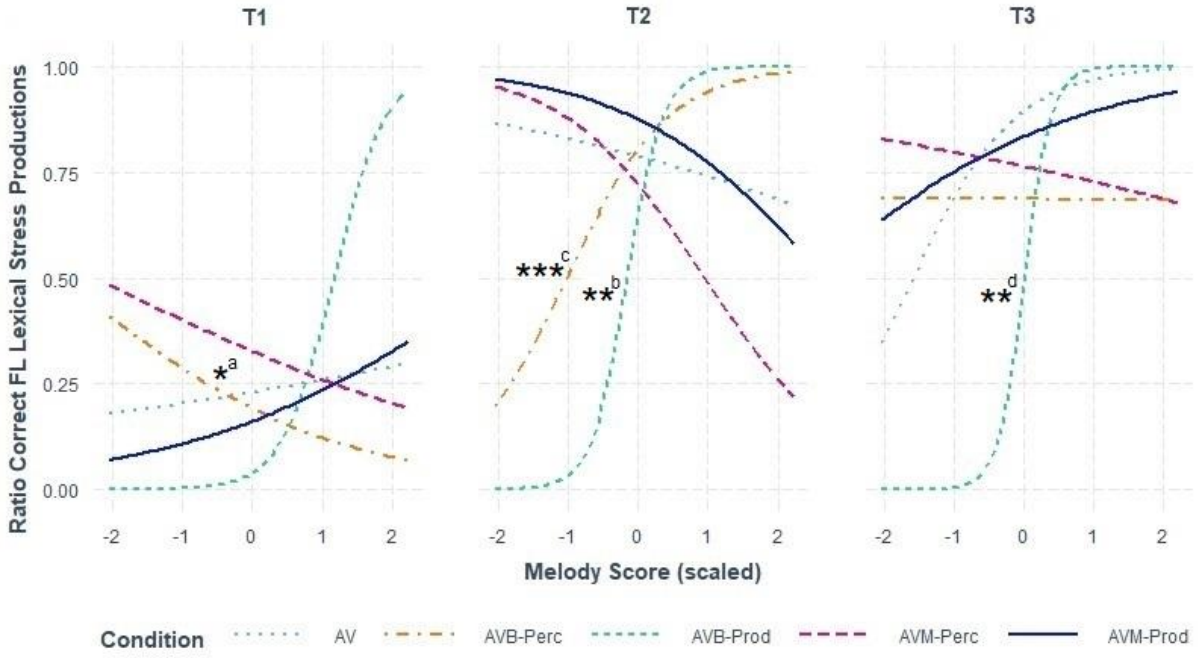
whereas at T1, in the AV-B-prod condition, lower WM scores predicted better performance on FL lexical stress production (which was also visible in the AV-B-perc condition), in the AV-M-perc and AV-M-prod condition, higher WM scores seemed to predict better performance (see Figure 14). Moreover, in the AV condition, WM scores did not predict performance. At T2 and T3, no significant differences between training conditions concerning the influence of WM capacity on producing correct FL lexical stress were found, though some interesting trends were observed which will be considered in the discussion. Since at T1 subjects had not yet received lexical stress training, present findings do not allow for any conclusions about the effectiveness of the different training conditions.

Second, there was a significant three-way interaction between *time of testing*, *training condition*, and *melody score* (see Table 11), which suggests that the relation between testing time and melody aptitude varied across training conditions. At T2 and T3, the influence of melody aptitude on producing correct FL lexical stress was significantly different in the AV-B-prod condition compared to in the other training conditions, respectively, $b = 4.18$, $SE = 1.33$, $p = .002$ and $b = 3.94$, $SE = 1.38$, $p = .004$. Bonferroni corrected contrast analyses revealed that, at T2, melody aptitude influenced FL lexical stress production differently in the AV-B-prod condition than in the AV-M-perc, $b = 4.35$, $SE = 1.01$, $p < .001$, and AV-M-prod condition, $b = 3.42$, $SE = 1.12$, $p = .023$. Namely, whereas in the AV-B-prod condition, higher melody scores predicted better performance, the opposite was true for the AV-M-perc and AV-M-prod condition (see Figure 15). Contrast analyses also indicated that, at T2, the influence of melody aptitude on FL lexical stress production was significantly different in the AV-B-perc condition compared to the AV-M-perc condition, $b = 1.95$, $SE = .48$, $p < .001$. Namely, in the AV-B-perc condition, as in the AV-B-prod condition, higher melody scores predicted better performance, whereas the opposite was true for the AV-M-perc condition (see Figure 15). At T3, the influence of melody aptitude on FL lexical stress production in the AV-

B-prod condition was significantly different from that in the AV-M-perc, b = 5.14, SE = 1.12, p < .001, and AV-B-perc condition, b = -4.35, SE = 1.10, p < .001. Namely, whereas in the AV-B-prod condition, higher melody scores predicted better performance, in the AV-B-perc and AV-M-perc condition melody scores did not have a predictive function (see Figure 15).

Figure 15

The Influence of the Relation Between Training Condition and Melody Score on the Ratio Correct FL Lexical Stress Productions Separated by Time of Testing



Note. Higher melody scores (scaled) indicate higher melody aptitude. ***: p < .001, **: p < .01, *: p < .05.

[a]Predictions in the AV-B-perc condition at T1 were significantly different than at T2. [b]At T2, predictions in the AV-B-prod condition were significantly different than in the AV-M-perc and AV-M-prod condition, [c]and predictions in the AV-B-perc condition were significantly different than in the AV-M-perc condition. [d]At T3, predictions in the AV-B-prod condition were significantly different than in the AV-B-perc and AV-M-perc condition.

Bonferroni corrected contrast analyses were used to explore how the relation between *training conditions* and m*elody score* varied across *time of testing*, and revealed that in the AV-B-perc condition the influence of melody aptitude on producing correct FL lexical stress was significantly different at T1 compared to at T2, b = -1.44, SE = .58, p = .036. Namely, whereas at T1, the influence of melody aptitude on FL lexical stress production was limited, but lower melody scores seemed to predict better performance, at T2, there was a clear relationship between melody aptitude and FL lexical stress production, indicating that higher melody scores predicted better performance (see Figure 15).

In short, whereas there were, as expected, no significant differences between training conditions concerning the influence of melody aptitude on producing correct FL lexical stress at T1, there were some significant differences between training conditions at T2 and T3. At T2, in the beat gesture conditions (i.e., AV-B-prod and AV-B-perc), higher melody scores predicted better performance, whereas in the metaphoric gesture conditions (i.e., AV-M-prod and AV-M-perc), lower melody scores predicted better performance. At T3, only in the AV-B-prod condition, higher melody scores predicted better performance, whereas in all other training conditions there was no clear relation. Therefore, these findings suggest that different training conditions were differently influenced by melody aptitude in predicting performance on FL lexical stress production.

Third, there was a significant three-way interaction between *time of testing*, *training condition*, and *rhythm score* (see Table 11), which suggests that the relation between time of testing and rhythm aptitude varied across training conditions. At T2 and T3, the influence of rhythm aptitude on producing correct FL lexical stress was significantly different in the AV-B-prod condition compared to in the other training conditions, respectively b = 1.13, SE = .36, p = .002, p = .002 and b = 1.99, SE = .43, p < .001. As can be seen in figure 16, at T2, in the AV-B-prod condition, higher rhythm scores predicted better performance (which was also

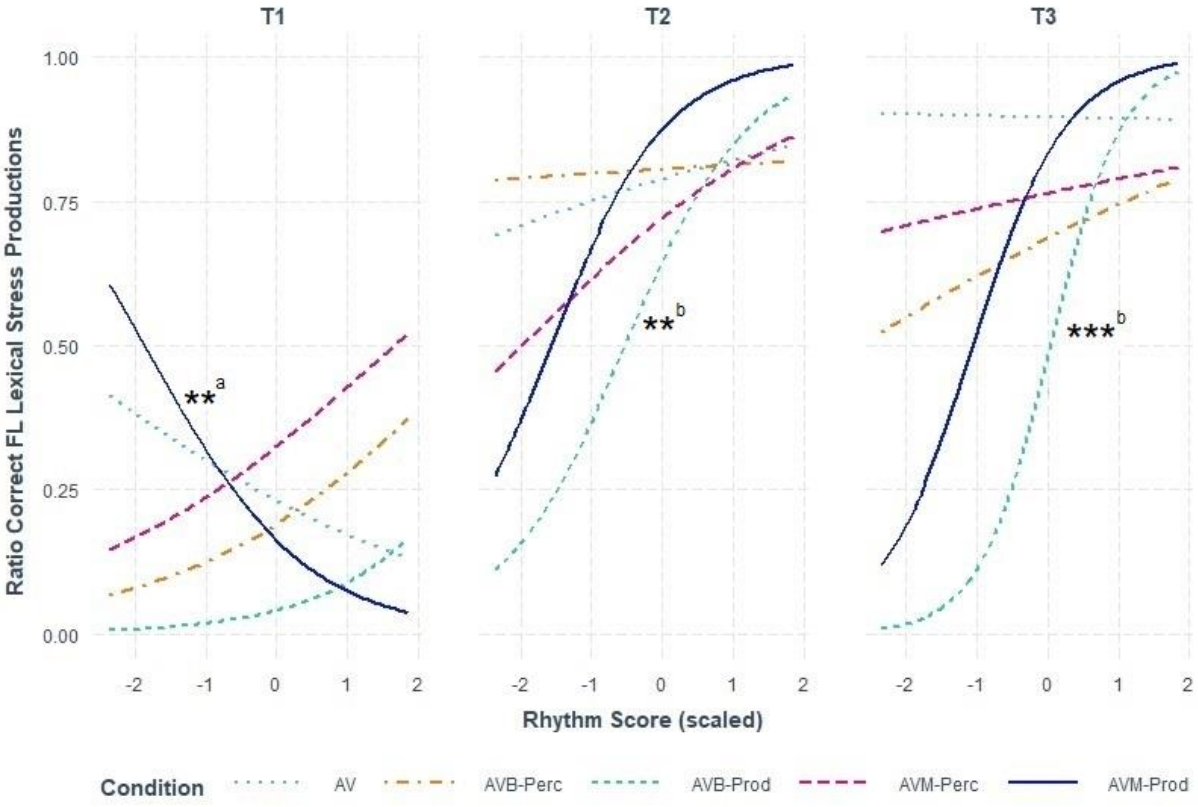
visible in the AV-M-prod condition), whereas in the AV, AV-B-perc, and AV-M-perc condition there was no clear relation. Similarly, Bonferroni corrected contrast analyses revealed that, at T3, the influence of rhythm aptitude on FL lexical stress production was significantly different in the AV-B-prod condition compared to the AV, AV-B-perc and AV-M-perc condition, respectively b = -2.02, SE = .64, p = .017, b = -1.70, SE = .47, p = .003, and b = 1.85, SE = .55, p = .008. Namely, whereas in the AV-B-prod condition, higher rhythm scores predicted better performance (which was also visible in the AV-M-prod condition), in the AV, AV-B-perc, and AV-M-perc this relation was less clear or non-existent (see Figure 16).

The significant three-way interaction also suggests that the relation between training condition and rhythm aptitude significantly varied across testing times. In the AV-M-prod condition, the influence of rhythm aptitude on producing correct FL lexical stress was significantly different at T1 compared to at T2 and T3, respectively, b = 1.24, SE = .48, p = .010 and b = 1.52, SE = .54, p = .005. Namely, whereas at T1, lower rhythm scores predicted better performance, at T2 and T3, higher rhythm scores predicted better performance (see Figure 16).

In short, in both the AV-B-prod and AV-M-prod condition, a higher rhythm aptitude predicted better performance on FL lexical stress production at T2 and T3, whereas in the other training conditions this relation was less clear. Therefore, rhythm aptitude might differently influence training conditions in which subjects are asked to produce gestures compared to the other training conditions.

Figure 16

The Influence of the Relation Between Training Condition and Rhythm Score on the Ratio Correct FL Lexical Stress Productions Separated by Time of Testing



Note. Higher rhythm scores (scaled) indicate higher rhythm aptitude. ***: $p < .001$, **: $p < .01$, *: $p < .05$.

[a]Predictions in the AV-M-prod condition at T1 were significantly different than at T2 and T3.

[b]At T2 and T3, predictions in the AV-B-prod condition were significantly different than in the AV, AV-B-perc, and AV-M-perc condition.

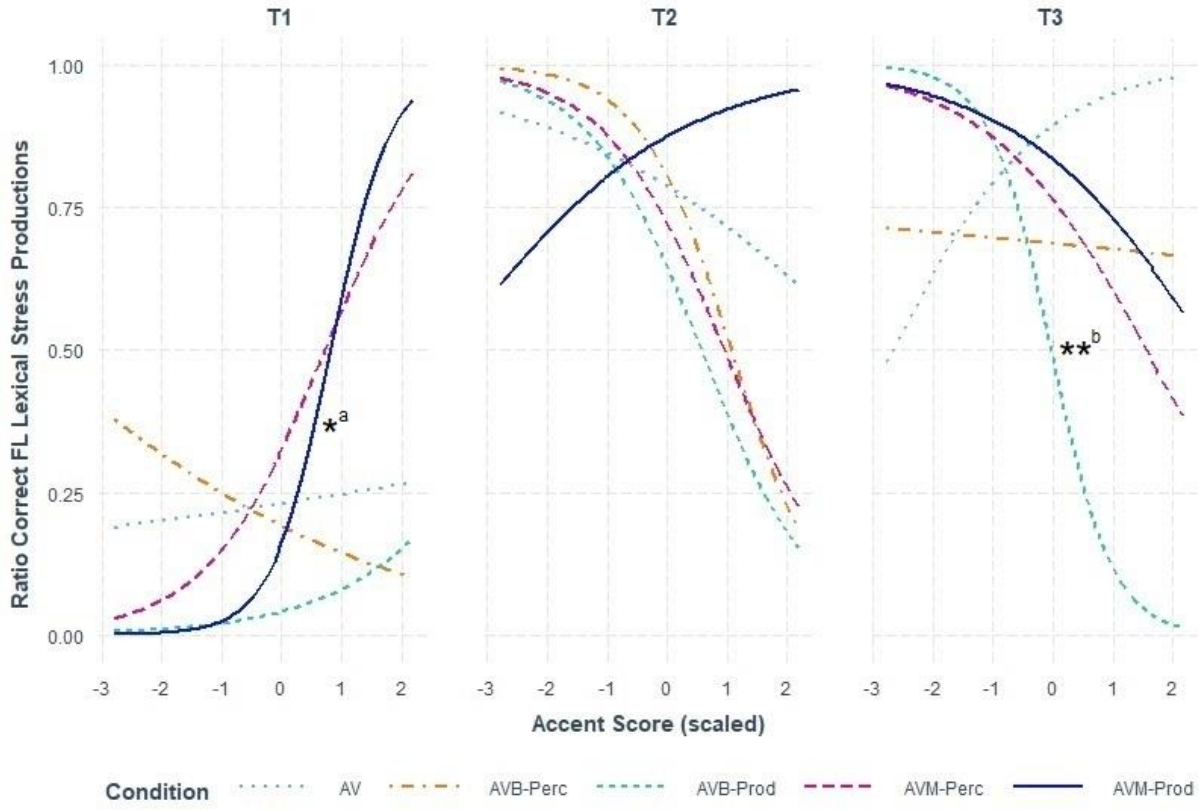
Fourth, there was a significant three-way interaction between *time of testing*, *training condition*, and *accent score* (see Table 11), which suggests that the relation between testing time and accent aptitude varied across training conditions. At T3, the influence of accent aptitude on producing correct FL lexical stress was significantly different in the AV-B-prod

condition compared to in the other training conditions, b = -1.98, SE = .69, p = .004. Bonferroni corrected contrast analyses revealed that, at T3, the influence of accent aptitude on FL lexical stress production was significantly different in the AV-B-prod condition compared to the AV and AV-B-perc condition, respectively b = 3.32, SE = 1.13, p = .033 and b = 2.51, SE = .82, p = .023. Namely, whereas in the AV-B-prod condition lower accent scores predicted better performance, in the AV condition, higher accent scores predicted better performance, and in the AV-B-perc condition accent scores did not predict performance on FL lexical stress production (see Figure 17).

The significant three-way interaction also suggests that the relation between training condition and accent aptitude significantly varied across testing times. In the AV-M-prod condition, the influence of accent aptitude on producing correct FL lexical stress was significantly different at T1 compared to at T2 and T3, b = 2.02, SE = .91, p = .026. Namely, at T1, higher accent scores predicted better performance, whereas at T2 and T3, this relation was less clear (see Figure 17). Although these findings are difficult to interpret, they seem to suggest that, at T3, especially in the AV-B-prod condition, but also in the AV-M-perc and AV-M-prod condition, lower accent scores predicted better performance on FL lexical stress production. Moreover, at T2, only in the AV-M-prod condition, higher accent scores predicted better performance.

Figure 17

The Influence of the Relation Between Training Condition and Accent Score on the Ratio Correct FL Lexical Stress Productions Separated by Time of Testing



Note. Higher accent scores (scaled) indicate higher accent aptitude. **: $p < .01$, *: $p < .05$.

[a]Predictions in the AV-M-prod condition at T1 were significantly different than at T2 and T3.

[b]At T3, predictions in the AV-B-prod condition were significantly different than in the AV and AV-B-perc condition.

Finally, several significant four-way interactions were found between *WM score* or *musical aptitude scores* on the one hand and *time of testing*, *training condition*, and *stress rule* on the other hand, and between *WM score* on the one hand and *time of testing*, *training condition*, and *musical aptitude scores* on the other hand. As these four-way interactions are difficult to interpret and are beyond the scope of this study, results will not be discussed here.

However, it should be noted that these interactions suggest that the influence of subjects' WM capacity and musical aptitude on the effectiveness of the different training conditions in FL lexical stress training differentiate depending on the stress rule being learned. Moreover, these interactions also suggest that musical aptitude and WM capacity interact in their influence on the effectiveness of the different training conditions in FL lexical stress training.

6. Discussion

The current study aimed to disentangle the influence of different gesture types (i.e., beat or metaphoric), different physical involvement levels (i.e., producing or perceiving gestures), and the learner characteristics WM capacity and musical aptitude on the effectiveness of FL lexical stress training by Dutch learners of Spanish. Four research questions were posed in relation to which relevant findings will be discussed. First, findings concerning the influence of gesture type (research question 1) and physical involvement level (research question 2) on the effectiveness of FL lexical stress training will be considered. Next, findings concerning the long-term influence (i.e., productions at T3) of training with gestures in learning FL lexical stress will be discussed (research question 4). Finally, relevant findings concerning the influence of WM capacity and musical aptitude on the effectiveness of FL lexical stress training will be considered (research question 3). Moreover, several findings that are not directly related the research questions will first be discussed here.

The present study found a significant improvement in learning outcome, indicating that subjects, irrespective of training condition, improved in their ability to produce correct FL lexical stress after a short training session, which was visible on both post-tests. These findings suggest that a short training session which focuses on one specific FL prosodic contrast, in this case lexical stress, can significantly improve FL speakers' productions.

Next, the present study found that, before training, stress rule (i.e., whether the target word included a written accent or not) significantly influenced whether FL lexical stress was produced correctly or not. At T1, subjects were significantly more likely to produce correct FL lexical stress in words without a written accent compared to in words with a written accent, but this difference disappeared after a short training session (at T2 and T3). This outcome is partly consistent with findings by Gutiérrez-Palma, Suárez-Coalla, and Cuetos (2020) who demonstrated that processing stress marks (i.e., written accents) might be

cognitively more demanding. What is surprising is that in the present study, this difference in performance between words with and without a written accent was no longer visible on the post-tests. A possible explanation might be that after a short training session which focusses on the different stress rules, the presence of a visual cue, such as a written stress mark, clearly guided the subjects in determining which syllable had to receive lexical stress, and thereby compensated for the more cognitively demanding processing of stress marks. Present findings are also partly in line with those of Van Maastricht et al. (2019) who reported that subjects produced significantly more correct FL lexical stress in words without a written accent not only before but also after lexical stress training. It is difficult to explain this discrepancy after training since Van Maastricht et al. (2019) and the present study used a highly similar experimental design. A possible explanation might be that, as lexical stress training only included one practice item in Van Maastricht et al. (2019), training was less effective than in the present study (which included two practice items) and could therefore not compensate for the more cognitively demanding processing of stress marks. Another related explanation is that, as lexical stress training is taken as a whole, without differentiating between the training conditions, producing gestures, which was only a training condition in the present study, might have had an important impact on the effectiveness of lexical stress training.

An unanticipated finding was that training with gestures (regardless of gesture type) was not significantly more effective compared to training without gestures in learning to produce FL lexical stress. Earlier studies on the effectiveness of gestures in learning FL prosody reported similar non-significant findings (e.g., Eng et al., 2013; Llanes-Coromina et al., 2018; Morett & Chang, 2015; Van Maastricht et al., 2019). The present study did report a descriptive trend in which most gesture training conditions, except for the AV-M-perc condition, demonstrated a greater improvement from T1 to T2 compared to the no gesture condition. These descriptive results agree with those obtained by Llanes-Coromina et al.

(2018) and Van Maastricht et al. (2019) who both reported a similar trend suggesting a greater learning benefit for training with gestures. The present trend is also consistent with other research which found a significant benefit for training with gestures in learning FL prosody (e.g., Baills et al., 2019; Gluhareva & Prieto, 2017; Hannah et al., 2017; Kelly et al., 2017; Yuan et al., 2018).

A possible explanation for the lack of significance in the present study seems to be a statistical power problem: A power analysis suggested that the present study is limited as substantially more subjects were required per training condition to reach statistical power (G*Power, version 3.1, Faul, Erdfelder, Buchner, & Lang, 2009). Another limitation of the present study is that in the statistical analysis no direct comparison was made between the four gesture training conditions on the one hand and the no gesture training condition on the other. Hence, the lack of significance may also (partly) be explained by statistical choices. The limited number of subjects per relevant training condition and the missing comparison between training with gestures versus training without gestures make it impossible to determine with certainty whether the non-existent difference between training with or without gestures is caused by a lack of statistical power, the type of statistical analysis, or by the actual absence of a significant difference. Therefore, future research should include more subjects per training condition to reach statistical power and make a direct comparison between training with gestures versus training without gestures. It might also be possible that the difference between finding a significant or non-significant benefit of gestures in learning FL prosody can be explained by the influence of gesture type (i.e., beat or metaphoric), physical involvement level (i.e., producing or perceiving gestures), and individual learner characteristics such as WM capacity and musical aptitude. Therefore, the influence of these different factors on the effectiveness of FL lexical stress training will be discussed below.

6.1 The influence of gesture type

As earlier studies reported mixed findings concerning the effectiveness of both beat and metaphoric gestures (e.g., Baills et al., 2019; Gluhareva & Prieto, 2017), the present study sought to determine whether gesture type influenced the effectiveness of FL lexical stress training (i.e., research question 1). We hypothesised that metaphoric gestures would be more effective compared to beat gestures as previous studies indicated more positive findings for metaphoric gestures in learning FL prosody (e.g., Van Maastricht et al., 2019; Yuan et al., 2018). Unfortunately, the current study found no significant influence of gesture type on the effectiveness of FL lexical stress training. However, as some descriptive trends were observed, it might be that the lack of significance was caused by the present study's statistical power problem.

One trend, for instance, was that subjects who perceived beat gestures in training demonstrated a greater improvement on FL lexical stress production from T1 to T2 compared to subjects who perceived metaphoric gestures during training. Similarly, subjects who produced beat gestures in training showed a slightly greater improvement from T1 to T2 compared to subjects who produced metaphoric gestures. These descriptive results suggest that immediately after lexical stress training, beat gestures might be more beneficial compared to metaphoric gestures in learning FL lexical stress.

These trends were unexpected, as they contradict our hypotheses, as well as findings by Van Maastricht et al. (2019) who suggested that metaphoric gestures were more effective than beat gestures in learning to produce FL lexical stress. This discrepancy might be partly attributed to the different types of beat gestures used in both studies. In Van Maastricht et al. (2019) beat gestures were made by moving one hand up and down whenever a stressed syllable occurred, whereas in the present study both hands together were moved vertically whenever there was a stressed syllable. By using both hands, it might be that, in the current

study, beat gestures were better visible and, therefore, more clearly indicated the appearance of stressed syllables, which enlarged the benefit of using beat gestures in learning FL lexical stress. The explanation for why beat gestures seem to be more effective in learning FL lexical stress compared to metaphoric gestures might be related to the diverging relations between speech and gesture in both gesture types. Beat gestures are commonly used to emphasise prosodic prominence in natural speech (Loehr, 2012; Pouw & Dixon, 2019; Wagner et al., 2014), whereas metaphoric gestures can be used to visually represent characteristics of specific prosodic contrasts in speech, in this case by representing the durational differences between stressed and unstressed syllables (i.e., long vs. short). Since metaphoric gestures are not aligned with prosodic prominence in speech, it might be that by making the beat gestures more distinctive (by using both hands) subjects have noticed this close alignment between gesture and speech, whereby the beat gestures have surpassed the 'representational benefit' of metaphoric gestures, and therefore became more effective in learning FL lexical stress.

The trend, that beat gestures seem to be more effective than metaphoric gestures is also somewhat surprising as previous studies exploring the effectiveness of either beat or metaphoric gestures in learning FL prosody reported more positive findings for training with metaphoric gestures (e.g., Baills et al., 2019; Hannah et al., 2017). As descriptives in the present study suggested better learning outcomes for training with beat gestures, it seems possible that the differences in effectiveness between beat and metaphoric gestures found in earlier studies were caused by some other underlying factors. For instance, it might be that whether learners' improvements on FL prosody were measured with a production or perception task significantly influenced the effectiveness of the different gesture types (e.g., Li et al., 2020). Another possibility is that what can be considered as the most effective gesture type in training depends on the FL prosodic contrast being learned (e.g., Eng et al., 2013; Kelly et al., 2017).

Present descriptive results, while preliminary, suggest that the effectiveness of gestures in FL prosody training might depend on the relation between gesture type, type of experimental task, and type of prosodic contrast. However, a major limitation of the current study is that it lacks statistical significance. Therefore, present findings should be interpreted with caution and future research is required to establish the viability of the findings. Another limitation of the present study is that the training videos used during lexical stress training differentiated in length: The videos in the metaphoric gesture conditions (i.e., AV-M-perc and AV-M-prod) were slightly longer compared to the videos in the beat gesture conditions (i.e., AV-B-perc and AV-B-prod) due to a short pause between producing the subject and adjective in the video-recordings. This pause was therefore added into the dubbed audio-file. The almost unnoticeable difference in length between training videos might have influenced the effectiveness of the training conditions which is something that should be controlled for in future research. Moreover, despite these promising results, important questions remain. Future research should, for instance, explore whether the effectiveness of beat and metaphoric gestures is dependent on whether subjects are asked to produce or perceive the prosodic contrasts (i.e., type of experimental task), or they should investigate whether the effectiveness of beat and metaphoric gestures depends on the type of FL prosodic contrast being learned.

6.2 The influence of physical involvement level

The present study also aimed to determine the influence of physical involvement level on the effectiveness of gestures in FL lexical stress training (i.e., research question 2). In line with the theories of embodied cognition and multimodality, we hypothesised that producing gestures, which involves more of the physical body and includes more modalities, would be more effective in learning to produce FL lexical stress compared to merely perceiving gestures (e.g., Kendon, 2004; McNeill, 1992; Wilson, 2002). Unfortunately, the present study

did not find any significant differences in effectiveness between producing or perceiving gestures in FL lexical stress training. However, as the lack of significant findings might be explained by the current study's statistical power problem, some descriptive trends will be discussed.

One observed trend was that producing (beat or metaphoric) gestures during FL lexical stress training resulted in a greater learning improvement from T1 to T2 compared to perceiving (beat or metaphoric) gestures during training. This is in line with findings by Baills (2016) and Kushch et al. (2018) who both reported that producing gestures resulted in significantly better learning outcomes compared to perceiving gestures. The descriptive trend also supports the theories of embodied cognition and multimodality. However, it should be noted that a similar non-significant difference between producing and perceiving gestures in learning FL prosody was reported in Baills et al. (2019) who did not report on any descriptive trends.

A limitation of the present study is that subjects in the gesture production conditions were instructed to produce gestures during lexical stress training, whereas subjects were allowed but not prompted to produce gestures during the post-tests. Therefore, there might be a greater difference in learning outcome between producing versus perceiving gestures in learning FL prosody when subjects are instructed to produce gestures during the post-tests, which is something that should be explored in future research. Moreover, as current findings lack statistical power and should be interpreted with caution, further research is required to determine whether physical involvement level influences the effectiveness of gestures in FL prosody training.

6.3 The long-term effects

Another objective of the present study was to determine whether there were any long-term effects of using gesture in FL lexical stress training (i.e., research question 4). As studies in FL vocabulary learning (e.g., Macedonia & Klimesch, 2014; Macedonia, et al, 2011) reported long-term memory enhancing benefits for training with gestures, we hypothesised that training with gestures would also benefit learning FL lexical stress in the long-term. However, similar to the short-term findings (i.e., from T1 to T2), the current study did not report any significant differences in long-term improvement between training with and without gestures on producing correct FL lexical stress (i.e., from T1 to T3 and from T2 to T3). Moreover, no significant differences in long-term improvement could be reported for training with beat versus metaphoric gestures, and for producing versus perceiving gestures in training. However, a visual inspection of the data indicated several descriptive trends.

One observed trend was that, whereas subjects in all training conditions improved from T1 to T3, only subjects who were trained without gestures and subjects who perceived metaphoric gestures in training improved from T2 to T3. Surprisingly, these observations suggest that FL lexical stress training without gestures was more beneficial in the long-term compared to training with gestures. A possible explanation might be that although the link between the gestures and the FL prosodic contrast was strong enough to benefit learning immediately after training (i.e., from T1 to T2 training with gestures seems to improve learning FL lexical stress more than training without gestures) the link was too indirect to benefit learning in the long-term (e.g., Kelly et al., 2017) and therefore lead to decrements in the learning process.

The descriptive results also indicated that subjects who produced beat gestures in training, closely followed by subjects who perceived beat gestures in training, showed the largest decrease in performance from T2 to T3. These observations suggest that beat gestures

were less effective compared to metaphoric gestures in learning to produce FL lexical stress in the long-term. This was unexpected, as beat gestures were considered to be most effective immediately after lexical stress training. A possible explanation might be that the acoustic alignment between beat gestures and speech prosody was distinctive enough for beat gestures to improve the production of FL lexical stress immediately after training, but, as this acoustic alignment was not remarkable, beat gestures did not make a distinctive enough memory trace to be used in the long-term (e.g., Macedonia & Klimesch, 2014). In metaphoric gestures the visual representation of the FL prosodic contrast might be more notable to the subjects, making a more distinctive memory trace, and therefore enabling subjects to recall the prosodic contrast in the long-term.

Another observed trend was that perceiving (beat or metaphoric) gestures in training appeared to result in a smaller decrease in performance from T2 to T3 compared to producing (beat or metaphoric) gestures, suggesting that in the long-term perceiving gestures was more effective compared to producing gestures in FL lexical stress training. Surprisingly, this trend contradicts findings immediately after training, as at T2, producing gestures in training appeared to be more effective compared to perceiving gestures. This trend also contradicts the theories of embodied cognition and multimodality as producing gestures did not result in a greater learning benefit (e.g., Kendon, 2004; McNeill, 1992; Wilson, 2002). It seems possible that these observations are related to a limitation in the present experimental design as producing gestures was only obligatory during a short training session and not during the post-tests. Therefore, it might be that, whereas immediately after lexical stress training producing gestures in training was more effective compared to perceiving gestures, in the long-term this effect of producing gestures had disappeared as lexical stress training at T3 was relatively long ago, whereas T2 immediately preceded it.

The descriptive findings in the present study raise the possibility that the overall influence of training with gestures and the influence of factors such as gesture type and physical involvement level might be different depending on testing time. However, a note of caution is due here as findings were non-significant. Moreover, a limitation of the present study was that the long-term effect was measured relatively shortly after T2 (i.e., about forty minutes later) and therefore findings on T3 might not reflect the actual long-term influence of gestures. Future research is required to establish the viability of these long-term findings. Moreover, further research should be undertaken to explore whether the influence of physical involvement level might be different when subjects are instructed to also produce gestures during the post-tests.

6.4 The influence of learner characteristics

The final objective of the present study concerned the influence of the learner characteristics WM capacity and musical aptitude on the effectiveness FL lexical stress training with gestures (research question 3).

6.4.1 WM capacity

As WM capacity is seen as a measure of language aptitude (e.g., Christiner & Reitener, 2013, 2015, 2016), subjects with a higher WM capacity were hypothesised to be better in producing FL lexical stress compared to subjects with a lower WM capacity. Surprisingly, the present study found that WM capacity was not significantly related to subjects' learning outcomes and improvements over time in learning to produce FL lexical stress. This was contrary to findings by Zhang et al. (2020) who reported that subjects with a higher WM capacity improved significantly more on producing FL lexical stress. However, it should be noted that in the present study a descriptive trend was visible in which subjects with a higher WM

capacity were slightly better in producing FL lexical stress compared to subjects with a lower WM capacity. Therefore, it might be, that the lack of significant findings is caused by the power problem in the current study. On the other hand, it could be that WM capacity does not significantly influence FL lexical stress productions, which is consistent with findings by Li et al. (2020) who reported that WM capacity did not significantly influence learning FL prosody. As the current study lacks statistical power, it cannot be concluded that WM capacity influences learning FL lexical stress, however, since some interesting trends were visible, more research, with a larger sample size, should be conducted.

Although WM capacity was not significantly related to subjects' learning outcomes and improvements over time, there was a significant three-way interaction between WM capacity on the one hand and testing time and training condition on the other hand. Earlier studies proposed contradicting theories about the influence of subjects' WM capacity on training with gestures: Some claim that gestures lighten the WM load, suggesting that especially subjects with a lower WM capacity benefit from training with gestures (Paivio, 1991), whereas others claim that gestures might overload the WM system in subjects with a lower WM capacity, suggesting that only subjects with a higher WM capacity benefit from training with gestures (Mayer & Moreno, 2003). Moreover, according to Kelly et al. (2017) the level of abstractedness between speech and gesture might be important, suggesting that if the relation between speech and gesture is unclear gestures are more likely to overload subjects' WM systems.

The present study found that, whereas the influence of WM capacity on learning FL lexical stress significantly differentiated between training conditions before lexical stress training, these significant differences disappeared after training (at T2 and T3). This finding seems to reflect some methodological problems in the present study as, before any training was given, subjects in the different training conditions significantly differed in how their WM

capacity influenced their FL lexical stress productions. Therefore, the present study seems to be limited in that there is great variation between subjects. Although some subjects did not meet the pre-defined experimental criteria (e.g., no experience with the Spanish language and no audio-visual impairments) these criteria did not cause any significant differences between subjects. It is therefore unknown what has caused the variation between subjects which is something that should be considered in further research.

Besides this significant difference at T1, descriptive findings revealed some interesting differences between training conditions after FL lexical stress training. The descriptives showed that, at T2, subjects with a lower WM capacity seemed to benefit more from producing gestures, whereas subjects with a higher WM capacity seemed to benefit more from perceiving gestures, suggesting that the most effective physical involvement level in FL lexical stress training might be dependent on subjects' WM capacities. This seems to be in line with Mayer and Moreno's (2003) information overload hypothesis and Kelly et al.'s (2017) theory about level of abstractedness. It might be that producing gestures is most beneficial in subjects with lower WM capacities, as in *gesture production* the relation between speech and gesture becomes more evident to the subjects since they have to actively reproduce the gestures. This gesture reproduction might make processing less cognitively demanding and therefore does not overload the WM systems in subjects with lower WM capacities. In *gesture perception*, the relation between speech and gesture might be more abstract to the subjects as subjects merely perceive the gestures, which might make processing more cognitively demanding, resulting in gesture perception to only be beneficial in subjects with high WM capacities.

Another observed trend in the present study was that, at T3, especially subjects with a lower WM capacity appeared to benefit from FL lexical stress training, as, irrespective of training condition, they outperformed subjects with a higher WM capacity, which is in line

with Paivio's (1991) dual coding theory. However, surprisingly, subjects with a lower WM capacity also benefitted more from training without gestures. Therefore, another possible explanation for these findings might be that, in a complex task such as this, in which subjects have to speak in a FL and recollect gestures to adhere to FL lexical stress rules, subjects with a lower WM capacity simply choose one task to focus on, in this case speech production, resulting in better performances, whereas subjects with a higher WM capacity might still try to focus on their speech production while also trying to recollect the gestures, which overloads their WM systems resulting in worse performances.

Present findings concerning the influence of WM capacity on the effectiveness of gestures in learning FL prosody are inconclusive. Not only are findings on the two post-tests non-significant and therefore speculative, but the descriptive results also support conflicting theories: Whereas observed trends at T2 are in line with Mayer & Moreno's (2003) information overload hypothesis, trends at T3 are consistent with Paivio's (1991) dual coding theory. Another limitation of the present study is that the WM task was originally developed to measure WM capacity in children (i.e., AWMA (Alloway, 2007)). Although the task had been used to measure WM capacity in adults before (Dunning & Holmes, 2014) and seems to differentiate between subjects' WM capacities in the present study, another WM task might reveal more subtle differences between subjects. In future research, with more statistical power and another WM task, the influence of WM capacity on the effectiveness of gestures in learning FL prosody should be further explored.

6.4.2 Musical aptitude

Whereas earlier studies suggested that subjects with a higher musical aptitude should improve more on learning FL prosody (Krauss & Chandrasekaran, 2010), the present study reported no significant influence of musical aptitude on subjects' learning outcomes and improvements

over time in learning to produce FL lexical stress. This was also contrary to findings by Yuan et al. (2018) who reported that subjects with a higher musical aptitude showed better learning outcomes in learning FL phrasal stress. However, descriptive results suggest that, in line with Yuan et al. (2018), subjects with a higher musical aptitude seem to produce more correct FL lexical stress. As the current study lacks statistical power, it cannot be concluded that musical aptitude influences learning FL lexical stress, however, since some interesting trends were visible, more research, with a larger sample size, should be conducted.

Although musical aptitude scores were not significantly related to subjects' learning outcomes and improvements over time, there were several significant three-way interactions between musical aptitude scores on the one hand and testing time and training condition on the other hand. First, concerning melody aptitude, melody scores significantly influenced differences between training conditions at T2 and T3, and not at T1, which was to be expected, as, at T1, subjects had not yet received training. One interesting finding was that, at T2, subjects with a lower melody aptitude seemed to benefit most from (producing or perceiving) metaphoric gestures, whereas subjects with a higher melody aptitude seemed to benefit most from (producing or perceiving) beat gestures. At T3, these differences had disappeared, though producing beat gestures was still most beneficial in subjects with a higher melody aptitude, and perceiving metaphoric gestures was still most beneficial for subjects with a lower melody aptitude. Together, these findings suggest that the effectiveness of different gesture types might be related to subjects' melody aptitudes. A possible explanation might be that, since beat gestures are related to the acoustic properties of speech (e.g., Loehr, 2012), beat gestures result in greater learning benefits for subjects who are better at perceiving acoustics (i.e., subjects with higher musical abilities), whereas metaphoric gestures, which are not related to the acoustics of speech, result in greater learning benefits for subjects with weaker musical abilities.

Second, concerning rhythm aptitude, one interesting finding was that the influence of rhythm scores on producing correct FL lexical stress differed significantly before and after training. As expected, there were no significant differences between training conditions at T1, whereas at T2 and T3, subjects with a higher rhythm aptitude benefitted most from producing gestures in training, and subjects with a lower rhythm aptitude benefitted more from perceiving gestures. These findings suggest that the influence of physical involvement level might depend on rhythm aptitude and that subjects with a lower rhythm aptitude benefit more from less physical involvement. This seems counterintuitive, as it may be expected that producing gestures would especially benefit subjects with a lower musical aptitude since subjects are not only presented with the visualised lexical stress patterns (as in the gesture perception conditions) but are also forced to reproduce those patterns. Subjects with a higher musical aptitude may be expected to benefit more from gesture perception, as they do not necessarily need this extra information from gesture production. A possible explanation for these unexpected findings is that both producing gestures and producing speech during FL lexical stress training was too cognitively demanding for subjects with a lower rhythm aptitude, resulting in better performances after gesture perception, whereas for subjects with a higher rhythm aptitude it was less cognitively demanding to produce both gestures and speech during training, resulting in better performances after gesture production (e.g., Kendon, 2004; McNeill, 1992; Wilson, 2002).

Third, concerning accent aptitude, another interesting finding was that the influence of accent scores on producing FL lexical stress significantly differed before and after training. As expected, there were no significant differences between training conditions at T1, whereas, at T2, subjects with a lower accent aptitude benefitted more from all training conditions compared to subjects with a higher accent aptitude, except from producing metaphoric gestures in training, which appeared to be most effective in subjects with a higher accent

aptitude. Moreover, at T3, subjects with a lower accent aptitude benefitted most from producing and perceiving metaphoric gestures and producing beat gestures in training, whereas subjects with a higher accent aptitude benefitted most from training without gestures. These findings are difficult to interpret but seem to suggest that especially subjects with a lower accent aptitude benefitted from training with gestures. Therefore, it might be, that while subjects with a higher accent aptitude are already able to hear the prosodic patterns in speech, subjects with a lower accent aptitude need more visual information to see which syllables need to be stressed.

Although the influence of the different musical aptitude measures varies, they seem to indicate that subjects with a lower musical aptitude need more visual information compared to subjects with a higher musical aptitude in learning FL prosody, but there is an important trade-off point: Whereas subjects with a lower melody aptitude benefitted more from metaphoric gestures, which are not acoustically related to speech but provide more information about the specific prosodic contrast compared to beat gestures, and subjects with a lower accent aptitude seemed to benefit more from gestural training in general, subjects with a lower rhythm aptitude did not benefit more from gesture production. Therefore, subjects with a lower musical aptitude seem to benefit from more visual information, as long as they can cognitively process this additional information. Reproducing gestures might be cognitively too demanding for subjects with low musical aptitudes, who already have to focus on the musical properties, and, therefore, perceiving gestures, which is less cognitively demanding, is more beneficial.

These findings contradict those of Li et al. (2020) who found that musical aptitude did not significantly predict the effectiveness of training with metaphoric gestures in learning FL vowel-length contrasts. Similarly, present findings contradict results by Yuan et al. (2018) who reported that only subjects with moderate musical abilities benefitted from training with

metaphoric gestures in learning FL phrasal stress. A possible explanation for these diverging findings might be that the relation between gesture type, physical involvement level, and type of prosodic contrast is important in determining the influence of musical aptitude on learning FL prosody. Another related possibility is that the type of musical aptitude task influenced the results. Although the same experimental task was used in all three studies (i.e., PROMS (Law & Zentner, 2012)), they differentiated in which subtests were conducted: Li et al. (2020) measured pitch and rhythm aptitude and Yuan et al. (2018) measured pitch and melody aptitude, whereas the present study measured melody, rhythm, and accent aptitude. Future research is needed to disentangle the influence of gesture type, physical involvement level, type of prosodic contrast, and type of experimental task to determine the influence of musical aptitude on learning FL prosody.

6.4.3 Four-way interactions

The present study found several significant four-way interactions between *WM score* or *musical aptitude scores* on the one hand and *time of testing*, *training condition*, and *stress rule* on the other hand, and between *WM score* on the one hand and *time of testing*, *training condition*, and *musical aptitude scores* on the other hand. However, as disentangling these interactions was not within the scope of this study, these interactions will not be discussed here. Though, future studies should take into consideration that the influence of learner characteristics seems to differentiate depending on the stress rule being learned and that multiple learner characteristics seem to interact in their influence on FL lexical stress training.

7. Conclusion

The present study aimed to disentangle the influence of different gesture types (i.e., beat or metaphoric), different physical involvement levels (i.e., producing or perceiving gestures), and learner characteristics WM capacity and musical aptitude on the effectiveness of FL lexical stress training by Dutch learners of Spanish. Results have shown that, irrespective of training condition, subjects significantly improved on their FL lexical stress productions from T1 to T2 and from T1 to T3. Although differences between training conditions were non-significant, the present study identified several descriptive trends suggesting that the effectiveness of gesture type and physical involvement level in FL lexical stress training might differentiate depending on testing time. Moreover, the present study reported several significant three-way interactions between WM capacity or musical aptitude on the one hand and testing time and training condition on the other hand. Hence, the effectiveness of gesture type and physical involvement level in FL lexical stress training was significantly influenced by WM capacity and musical aptitude. Taken together, present findings suggest that in learning FL prosody, in which the relation between speech and gesture might not be immediately apparent, learner characteristics and methodological choices, such as gesture type and physical involvement level, are important in determining the gestural benefit and should be carefully considered in future research.

8. References

Alloway, T. P. (2007). *Automated working memory assessment: Manual*. London: Pearson Assessment.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience, 4*(10), 829-839.

Baills, F. (2016). Producing pitch gestures enhances the acquisition of Mandarin Chinese tones (Unpublished master's thesis). Universitat Pompeu, Barcelona, Spain.

Baills, F., Suárez-González, N., González-Fuente, S., & Prieto, P. (2019). Observing and producing pitch gestures facilitates the learning of mandarin Chinese tones and words. *Studies in Second Language Acquisition, 41*(1), 33-58. doi: 10.1017/S0272263118000074

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2020). lme4: Linear mixed-effects models using 'Eigen' and S4. R package version 1.1-26. https://cran.r-project.org/web/packages/lme4/index.html

Behrman, A. (2014). Segmental and prosodic approaches to accent management. *American Journal of Speech-Language Pathology, 23*(4), 546-561. doi: 10.1044/2014_AJSLP-13-0074

Bosker, H. R., & Peeters, D. (2021). Beat gestures influence which speech sounds you hear. *Proceedings of the Royal Society B, 288*(1943). doi: 10.1098/rspb.2020.2419

Broersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer [Computer program]. Version 6.0.37, retrieved 17 April 2018 from http://www.praat.org/

Christiner, M., & Reiterer, S. M. (2013). Song and speech: Examining the link between singing talent and speech imitation ability. *Frontiers in Psychology, 4*, 874-885. doi: 10.3389/fpsyg.2013.00874

Christiner, M., & Reiterer, S. M. (2015). A Mozart is not a Pavarotti: Singers outperform instrumentalists on foreign accent imitation. *Frontiers in Human Neuroscience, 9*, 482-490. doi: 10.3389/fnhum.2015.00482

Christiner, M., & Reiterer, S. (2016). Music, song and speech. In G. Granena, D.O. Jackson, & Y. Yilmaz (Eds.), *Cognitive individual differences in second language processing and acquisition* (pp. 131-147). Amsterdam/Philadelphia: John Benjamins B.V.

Christiner, M., & Reiterer, S. (2018). Early influence of musical abilities and working memory on speech imitation abilities: Study with pre-school children. *Brain Sciences, 8*(9), 169-185. doi: 10.3390/brainsci8090169

Clark, E., & Paran, A. (2007). The employability of non-native-speaker teachers of EFL: A UK survey. *System, 35*(4), 407-430. doi: 10.1016/j.system.2007.05.002

Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition, 106*(2), 1047-1058. doi: 10.1016/j.cognition.2007.04.010

Cook, S. W., Yip, T. K., & Goldin-Meadow, S. (2012). Gestures, but not meaningless movements, lighten working memory load when explaining math. *Language and Cognitive Processes, 27*(4), 594-610. doi: 10.1080/01690965.2011.567074

Crosswhite, K. (n.d.). Text Grid Maker (Modified by M. Antoniou & E. Doty) [Praat script]. Retrieved from https://www.ub.edu/phoneticslaboratory/index.html

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19*(1). doi: 10.1017/S0272263197001010

Derwing, T., & Munro, M. (2005). Preparation of teachers of English as a second language: Putting the NS/NNS debate in context. In E. Llurda (Ed.), *Non-native language teachers. Perceptions, challenges and contributions to the profession* (pp. 179-191). Springer.

Derwing, T., Munro, M., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning, 48*(3), 393–410. doi: 10.1111/0023-8333.00047

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

Dunning, D. L., & Holmes, J. (2014). Does working memory training promote the use of strategies on untrained working memory tasks?. *Memory & Cognition, 42*(6), 854-862. doi: 10.3758/s13421-014-0410-5

Eng, K., Hannah, B., Leong, L., & Wang, Y. (2013). Can co-speech hand gestures facilitate learning of non-native tones?. *Proceedings of Meetings on Acoustics ICA2013, 19*(1). doi: 10.1121/1.4799746

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.

Gamer, M., Lemon, J., & Fellows Puspendra Signh, I. (2019). irr: Various coefficients of interrater reliability and agreement. R package version 0.84.1. https://cran.r-project.org/web/packages/irr/index.html

Gluhareva, D., & Prieto, P. (2017). Training with rhythmic beat gestures benefits L2 pronunciation in discourse-demanding situations. *Language Teaching Research, 21*(5), 609-631. doi: 10.1177/1362168816651463

Goldin-Meadow, S., & Butcher, C. (2003). Pointing toward two-word speech in young children. In S. Kita (Ed.), *Pointing: Where language, culture, and cognition meet* (pp. 93-116). New York: Psychology Press.

Gordon, J., Darcy, I., & Ewert, D. (2013). Pronunciation teaching and learning: Phonetic instruction in the L2 classroom. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 4th pronunciation in second language learning and teaching conference* (pp. 194–206). Ames: IA.

Grace-Martin, K. (2019). *Specifying fixed and random factors in mixed models*. The analysis factor. https://www.theanalysisfactor.com

Gutiérrez-Palma, N., Suárez-Coalla, P., & Cuetos, F. (2020). Stress assignment in reading aloud in Spanish. *Applied Psycholinguistics, 41*(4), 753-769. doi: 10.1017/S014271642000020X

Hannah, B., Wang, Y., Jongman, A., Sereno, J. A., Cao, J., & Nie, Y. (2017). Cross-modal association between auditory and visuospatial information in Mandarin tone perception in noise by native and non-native perceivers. *Frontiers in Psychology, 8*. doi: 10.3389/fpsyg.2017.02051

Harrell, F.E. (2021). rms: Regression modelling strategies. R package version 6.2-0. https://cran.rproject.org/web/packages/rms/index.html

Hendriks, B., Van Meurs, F., & Hogervorst, N. (2016). Effects of degree of accentedness in lecturers' Dutch-English pronunciation on Dutch students' attitudes and perceptions of comprehensibility. *Dutch Journal of Applied Linguistics, 5*(1). doi: 10.1075/dujal.5.1.01hen

Hirata, Y. (2004). Computer assisted pronunciation training for native English speakers learning Japanese pitch and durational contrasts. *Computer Assisted Language Learning, 17*(3-4), 357-376. doi: 10.1080/0958822042000319629

Hirata, Y., & Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research, 53*(2), 298-310. doi: 10.1044/1092-4388

Hirata, Y., Kelly, S. D., Huang, J., & Manansala, M. (2014). Effects of hand gestures on auditory learning of second-language vowel length contrasts. *Journal of Speech, Language, and Hearing Research, 57*(6), 2090-2101. doi: 10.1044/2014_JSLHR-S-14-0049

Hoetjes, M., & Van Maastricht, L. (2020). Using gesture to facilitate L2 phoneme acquisition: The importance of gesture and phoneme complexity. *Frontiers in Psychology, 11*. doi: 10.3389/fpsyg.2020.575032

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science, 16*(5), 367-371. doi: 10.1111/j.0956-7976.2005.01542.x

Kelly, S., Bailey, A., & Hirata, Y. (2017). Metaphoric gestures facilitate perception of intonation more than length in auditory judgments of non-native phonemic contrasts. *Collabra: Psychology, 3*(1). doi: 10.1525/collabra.76

Kelly, S. D., Hirata, Y., Manansala, M., & Huang, J. (2014). Exploring the role of hand gestures in learning novel phoneme contrasts and vocabulary in a second language. *Frontiers in Psychology, 5*, 673. doi: 10.3389/fpsyg.2014.00673

Kelly, S. D., & Lee, A. L. (2012). When actions speak too much louder than words: Hand gestures disrupt word learning when phonetic demands are high. Language and *Cognitive Processes, 27*(6), 793-807. doi: 10.1080/01690965.2011.581125

Kelly, S.D., McDevitt, T. & Esch, M. (2009). Brief training with co-speech language lends a hand to word learning in a foreign language. *Language and Cognitive Processes, 24*(2), 313-334. doi: 10.1080/01690960802365567

Kendon, A. (1972). Some relationships between body motion and speech. In Siegman, A.W., Pope, B., *Studies in dyadic communication* (pp. 177-210). New York, USA: Pergamon Press.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process utterance. In M.R. Key (Ed.), *The relationship of verbal and nonverbal communication* (pp. 207-222). The Hague: Mouton Publishers.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.

Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), *Language and gesture* (pp. 162-185). Cambridge: Cambridge University Press.

Kita, S. (Ed.). (2003). *Pointing: Where language, culture, and cognition meet*. Psychology Press.

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language, 48*(1), 16-32. doi: 10.1016/S0749-596X(02)00505-3

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language, 57*(3), 396-414. doi: 10.1016/j.jml.2007.06.005

Kraus, N., & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews Neuroscience, 11*(8), 599-605.

Kushch, O., Gluhareva, D. Borràs-Comes, J, Perez, P.,& Prieto, P. (2018). The role of beat gesture production in L2 pronunciation training. Submitted for publication to *Bilingualism: Language and Cognition*

Law, L. N., & Zentner, M. (2012). Assessing musical abilities objectively: Construction and validation of the Profile of Music Perception Skills. *PloS One, 7*(12). doi: 10.1371/journal.pone.0052508

Lennes, M. (2003). Extracts F0 from points [Praat script]. Retrieved from https://www.ub.edu/phoneticslaboratory/praat-scripts.html

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*(6), 1093-1096. doi: 10.1016/j.jesp.2010.05.025

Li, P., Baills, F., & Prieto, P. (2020). Observing and producing durational hand gestures facilitates the pronunciation of novel vowel-length contrasts. *Studies in Second Language Acquisition, 42*(5). doi: 10.1017/S0272263120000054

Llanes-Coromina, J., Prieto, P., Rohrer, P.L. (2018). Brief training with rhythmic beat gestures helps L2 pronunciation in a reading aloud task. In Klessa K, Bachan J, Wagner A, Karpiński M, Śledziński D., *Proceedings of the 9th international conference on speech prosody* (pp. 498-502). Poznań, Poland: ISCA. doi: 10.21437/SpeechProsody.2018-101

Loehr, D. P. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology, 3*(1), 71-89. doi: 10.1515/lp-2012-0006

Long, J.A. (2020). interactions: Comprehensive, user-friendly, toolkit for probing interactions. R package version 1.1.3. https://cran.r-project.org/web/packages/interactions/index.html

Macedonia, M., & Klimesch, W. (2014). Long-term effects of gestures on memory for foreign language words trained in the classroom. *Mind, Brain, and Education, 8*(2), 74-88. doi: 10.1111/mbe.12047

Macedonia, M., Müller, K., & Friederici, A. D. (2011). The impact of iconic gestures on foreign language word learning and its neural substrate. *Human Brain Mapping, 32*(6), 982-998. doi: 10.1002/hbm.21084

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44*(2), 314-324. doi: 10.3758/s13428-011-0168-7

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 43-52. doi: 10.1207/S15326985EP3801_6

McNeill, D. (1985). So you think gestures are nonverbal?. *Psychological Review, 92*(3), 350-371. doi: 10.1037/0033-295X.92.3.350

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago press.

McNeill, D. (2008). *Gesture and thought*. University of Chicago press.

Morett, L. M. (2018). In hand and in mind: Effects of gesture production and viewing on second language word learning. *Applied Psycholinguistics, 39*(2), 355-382. doi: 10.1017/S0142716417000388

Morett, L.M., & Chang, L.Y. (2015). Emphasising sound and meaning: Pitch gestures enhance Mandarin lexicon tone acquisition. *Language, Cognition, and Neuroscience, 30*(3), 347-353. doi: 10.1080/23273798.2014.923105

Munro, M.J. & Derwing, T.M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45,* 73-97. doi: 10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 49*, 285-310. doi:10.1111/0023-8333.49.s1.8

Niedenthal, P. M. (2007). Embodying emotion. *Science, 316*(5827), 1002-1005. doi: 10.1126/science.1136930

Ortega-Llebaria, M. (2006). Phonetic cues to stress and accent in Spanish. In M. Díaz-Campos (Ed.), *Selected proceedings of the 2nd conference on laboratory approaches to Spanish phonetics and phonology* (pp. 104-118). Somerville, MA: Cascadilla Proceedings Project.

Ortega-Llebaria, M., del Mar Vanrell, M., & Prieto, P. (2010). Catalan speakers' perception of word stress in unaccented contexts. *The Journal of the Acoustical Society of America, 127*(1), 462-471. doi: 10.1121/1.3268506

Özer, D., & Göksun, T. (2020). Visual-spatial and verbal abilities differentially affect processing of gestural vs. spoken expressions. *Language, Cognition and Neuroscience, 35*(7), 896-914. doi: 10.1080/23273798.2019.1703016

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology, 45*(3), 255. doi: 10.1037/h0084295

Phonetics Laboratory (n.d.). Form extract and save intervals [Praat script]. Retrieved from https://www.ub.edu/phoneticslaboratory/praat-scripts.html

Pouw, W., & Dixon, J. A. (2019). Quantifying gesture-speech synchrony. *In the 6th gesture and speech in interaction conference* (pp. 75-80). Universitaetsbibliothek Paderborn.

Prieto, P., del Mar Vanrell, M., Astruc, L., Payne, E., & Post, B. (2012). Phonotactic and phrasal properties of speech rhythm. Evidence from Catalan, English, and Spanish. *Speech Communication, 54*(6), 681-702. doi: 10.1016/j.specom.2011.12.001

Quin-Allen, L. (1995). The effects of emblematic gestures on the development and access of mental representations of French expressions. *The Modern Language Journal, 79*(4), 521-529. doi: 10.1111/j.1540-4781.1995.tb05454.x

R Core Team (2019). R: A language and environment for statistical computing. [Computer Program]. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from: https://www.R-project.org/

Revelle, W. (2020). psych: Procedures for psychological, psychometric, and personality research. R package version 2.0.12. https://cran.rproject.org/web/packages/psych/index.html

Rietveld, T., & van Heuven, V. J. (2009). *Algemene fonetiek*. Bussum, The Netherlands: Coutinho.

Robinson, P. (2012). Aptitude in second language acquisition. *Annual Review of Applied Linguistics, 25*, 46-73. doi: 10.1017/S0267190505000036

RStudio Team (2020). RStudio: Integrated Development for R. [Computer Program]. RStudio, PBC. Boston, MA. Retrieved from: http://www.rstudio.com/

Sweller, N., Shinooka-Phelan, A., & Austin, E. (2020). The effects of observing and producing gestures on Japanese word learning. *Acta Psychologica, 207*. doi: 10.1016/j.actpsy.2020.103079

Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture, 8*(2), 219-235. doi: 10.1075/gest.8.2.06tel

Van Maastricht, L., Hoetjes, M., & van Drie, E. (2019). Do gestures during training facilitate L2 lexical stress acquisition by Dutch learners of Spanish? *Proceedings of the 15th international conference on auditory-visual speech processing* (AVSP2019).

Van Maastricht, L., Krahmer, E., & Swerts, M. (2016). Native speaker perceptions of (non-) native prominence patterns: Effects of deviance in pitch accent distributions on accentedness, comprehensibility, intelligibility, and nativeness. *Speech Communication, 83*, 21-33. doi: 10.1016/j.specom.2016.07.008

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Elsevier, 57*, 209-232. doi: 10.1016/j.specom.2013.09.008

Wen, Z., & Skehan, P. (2011). A new perspective on foreign language aptitude research: building and supporting a case for "working memory as language aptitude". *Ilha do*

Desterro: A Journal of English Language, Literatures in English and Cultural Studies, 60, 15-44. doi:10.5007/2175-8026.2011n60p015

Warnes, G.R., Bolker, B., Lumley, T., & Johnson, R.C. (2018). gmodels: Various Programming Tools for Model Fitting. R package version 2.18.1. https://cran.r-project.org/web/packages/gmodels/index.html

Wickham, H., Chang, W., Henry, L., Pedersen, T.L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., & Dunnington, D. (2020). ggplot2: Create elegant data visualisations using the grammar of graphics. R package version 3.3.3. https://cran.r-project.org/web/packages/ggplot2/index.html

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review, 9*(4), 625-636. doi: 10.3758/BF03196322

Yuan, C., González-Fuente, S., Baills, F., & Prieto, P. (2018). Observing pitch gestures favors the learning of Spanish intonation by Mandarin speakers. *Studies in Second Language Acquisition, 41*(1), 5-32. doi: 10.1017/S0272263117000316

Zhang, Y., Baills, F., & Prieto, P. (2020). Hand-clapping to the rhythm of newly learned words improves L2 pronunciation: Evidence from training Chinese adolescents with French words. *Language Teaching Research, 24*(5), 666-689. doi: 10.1177/1362168818806531

Zwaan, R. A. (2004). The immersed experiencer: Toward an embodied theory of language comprehension. In B.H. Ross, *The psychology of learning and motivation: Advances in research and theory, 44* (pp. 35-62). Elsevier Academic Press.

9. Appendices

Appendix A - Example recruitment poster

Radboud University

Viva España! Stoomcursus Spaans in 60 minuten

Locatie: Erasmusgebouw 12.18A

Heb je altijd al willen oefenen met het Spaans? Of heb je een uurtje over en wil je graag 1 proefpersoonuur of een VVV-bon van €10, - verdienen? Schrijf je dan in voor de studie Viva España!

Verdien 1 ppu of een VVV-bon t.w.v. €10, - binnen 60 minuten

Na deelname krijg je natuurlijk ook iets lekkers!

Schrijf je in voor een tijdslot op SONA of stuur een mail naar **…**

[CLS lab] Behavioural: Viva España! Stoomcursus Spaans in 60 minuten. | 10 euro/1 ppu

Je bent helaas uitgesloten van deelname aan dit onderzoek wanneer je:

- Spaanse les hebt gehad
- Langere tijd (> 3 maanden) in een Spaanstalig land bent verbleven
- ***Geen*** moedertaal spreker van het Nederlands bent
- Meertalig bent opgevoed

Appendix B - Information document

INFORMATIEDOCUMENT

Naam onderzoek: Viva España! Stoomcursus Spaans in 60 minuten.

Verantwoordelijke onderzoeker: Marieke Hoetjes & Lieke van Maastricht

Doel en procedure van het onderzoek

In dit onderzoek gaat u verschillende korte taken uitvoeren. U gaat onder andere meermaals een aantal Spaanse zinnen voorlezen. Tussen deze zinnen door zult u een training krijgen. Daarnaast gaat u oefenen met het vergelijken van verschillende melodieën en met het benoemen van getallenreeksen. Als u wilt, kunnen we na afloop uitleggen waar we onderzoek naar doen. Daarnaast willen wij u vragen het experiment na afloop niet onderling te bespreken.

Risico's en ongemakken

Er zijn geen risico's voor uw gezondheid of uw veiligheid.

Vertrouwelijkheid van de onderzoeksgegevens

De gegevens die we in dit onderzoek verzamelen, zullen door wetenschappers gebruikt worden voor artikelen en presentaties. Natuurlijk maken we deze gegevens volledig anoniem en bewaren we ze volgens de aan de Radboud Universiteit geldende regels. Uitgangspunt is dat de anoniem gemaakte data tenminste 10 jaar ten behoeve van de wetenschappelijke gemeenschap opvraagbaar zijn. In dit onderzoek worden audio- en video-opnames gemaakt. U krijgt van ons een formulier waarop u toestemming voor het maken en gebruiken van deze opnames kunt geven.

Vrijwilligheid

U doet vrijwillig mee aan dit onderzoek. Daarom kunt u op elk moment tijdens het onderzoek uw deelname stopzetten. Alle gegevens die we bij u verzameld hebben, worden dan definitief verwijderd. Ook na het onderzoek (tot 24 uur na deelname) kunt u dit aan ons doorgeven.

Vergoeding

Als dank voor uw medewerking, ontvangt u van ons *1 proefpersoonuur* of een VVV-bon ter waarde van *10 euro*.

Wie kan niet aan dit onderzoek meedoen?

Wanneer u een opleiding of cursus Spaans volgt of hebt gevolgd wordt u uitgesloten van dit onderzoek. Ook wordt u uitgesloten wanneer u langere tijd (> 3 maanden) in een Spaanstalig land heeft gewoond. Daarnaast mogen moedertaalsprekers van het Spaans niet meedoen aan dit onderzoek. Wanneer u tweetalig bent opgevoed wordt u ook uitgesloten van dit onderzoek. Verder wordt geadviseerd niet deel te nemen wanneer u auditieve of visuele problemen heeft die niet middels een apparaat gecorrigeerd kunnen worden.

Nadere inlichtingen

Als u graag verdere informatie over het onderzoek wilt hebben, nu of in de toekomst, kunt u contact opnemen met *Lisette van der Heijden* (telefoon: …; e-mail: …).

Voor eventuele klachten over dit onderzoek kunt u contact opnemen met:

Margret van Beuningen, secretaris Ethische Toetsingscommissie Geesteswetenschappen Radboud Universiteit.

Appendix C - Informed consent

TOESTEMMINGSVERKLARING

Naam onderzoek: *Viva España! Stoomcursus Spaans in 60 minuten.*

Verantwoordelijke onderzoeker: *Marieke Hoetjes & Lieke van Maastricht*

Verklaring deelnemer

Ik heb uitleg gekregen over het doel van het onderzoek. Ik heb vragen mogen stellen over het onderzoek. Ik neem vrijwillig deel aan het onderzoek. Ik begrijp dat ik op elk moment tijdens het onderzoek mag stoppen als ik dat wil. Ik begrijp hoe de gegevens van het onderzoek bewaard zullen worden en waarvoor ze gebruikt zullen worden. Ik stem in met deelname aan het onderzoek. Daarbij geef ik toestemming om audio- en video-opnamen van mij te maken voor dit onderzoek, deze opnamen op te slaan in een databank, volgens de geldende regels van de Radboud Universiteit Nijmegen, en beschikbaar te houden voor wetenschappelijk onderzoek.

Toestemming gebruik audio- en video-opnamen

Ik geef toestemming om (s.v.p. aankruisen wat van toepassing is):

Ja Nee

☐ ☐ deze audio- en video-opnamen af te spelen op congressen en voor onderwijsdoeleinden.

overige opmerkingen:

……

Ja Nee

☐ ☐ screenshots van deze video-opnamen te gebruiken in publicaties over dit onderzoek.

overige opmerkingen:

……

Naam: …….

Geboortedatum:……....

Handtekening: .....

Datum:…….....

Verklaring uitvoerend onderzoeker (zie ommezijde)

Verklaring uitvoerend onderzoeker

Ik verklaar dat ik de hierboven genoemde persoon juist heb geïnformeerd over het onderzoek en dat ik mij houd aan de richtlijnen voor onderzoekers zoals verwoord in het protocol van de Ethische Toetsingscommissie Geesteswetenschappen.

Naam:…….....……

Handtekening: .....

Datum:…….....

Appendix D - Filler and target items in read-aloud tasks and lexical stress training

Target items

No.	Item	Phrase	Rule	Spanish stress	Dutch stress	Code
O1	elefante	El elefante gris	1	ele**fan**te	**o**lifant	R1_O1
1	clima	El clima frío	1	**cli**ma	kli**maat**	R1_E1
2	vacaciones	Las vacaciones largas	1	vaca**cio**nes	va**kan**tie	R1_E2
3	horizonte	El horizonte bonito	1	hori**zon**te	**ho**rizon	R1_E3
O2	radiador	El radiador moderno	2	radia**dor**	radi**a**tor	R2_O2
4	profesor	El profesor genial	2	profe**sor**	pro**fe**sor	R2_E4
5	ventilador	El ventilador metálico	2	ventila**dor**	venti**la**tor	R2_E5
6	internet	El internet grande	2	inter**net**	**in**ternet	R2_E6
7	mamut	El mamut enorme	2	ma**mut**	**ma**moet	R2_E7
O3	fórmula	La fórmula difícil	3	**fór**mula	for**mu**le	R3_O3
8	análisis	El análisis difícil	3	a**ná**lisis	ana**ly**se	R3_E8
9	centímetro	El centímetro largo	3	cen**tí**metro	**cen**timeter	R3_E9
10	pirámides	Los pirámides grandes	3	pi**rá**mides	pira**mi**des	R3_E10
11	cardiólogos	Los cardiólogos inteligentes	3	car**dió**logos	cardio**lo**gen	R3_E11
12	máquina	La máquina roja	3	**má**quina	ma**chi**ne	R3_E12
13	sofá	El sofá rosa	3	so**fá**	**so**fa	R3_E13
14	automóvil	El automóvil negro	3	auto**mó**vil	automo**biel**	R3_E14

Filler items

No.	Item	Phrase	Rule	Spanish stress	Dutch stress	Code
1	formulario	El formulario complicado	1	formu**la**rio	formu**lier**	R1_F1
2	mango	El mango verde	1	**man**go	**man**go	R1_F2
3	planeta	La planeta bonita	1	pla**ne**ta	pla**neet**	R1_F3
4	tomates	Los tomates frescos	1	to**ma**tes	to**ma**ten	R1_F4
5	princesas	Las princesas guapas	1	prin**ce**sas	prin**se**ssen	R1_F5
6	parasol	El parasol rosa	2	para**sol**	para**sol**	R2_F6
7	universidad	La universidad grande	2	universi**dad**	universi**teit**	R2_F7
8	brócoli	El brócoli verde	3	**bró**coli	**bro**ccoli	R3_F8
9	político	El político famoso	3	po**lí**tico	po**li**ticus	R3_F9
10	cámara	La cámara cara	3	**cá**mara	**ca**mera	R3_F10
11	números	Los números romanos	3	**nú**meros	**num**mers	R3_F11
12	históricos	Los históricos viejos	3	his**tó**ricos	his**to**rici	R3_F12
13	fútbol	El fútbol naranja	3	**fút**bol	**voet**bal	R3_F13
14	saxofón	El saxofón azul	3	saxo**fón**	saxo**foon**	R3_F14

Example items (in lexical stress training)

No.	Item	Phrase	Rule	Spanish stress	Dutch stress
1	computadora	La computadora violeta	1	computa**do**ra	com**pu**ter
2	doctor	El doctor genial	2	doc**tor**	**dok**ter
3	teléfono	El teléfono móvil	3	te**lé**fono	tele**foon**

Appendix E - Randomised items (version A and B) in read-aloud tasks

T1-A	T1-B	T2-A	T2-B	T3-A	T3-B
máquina	pirámides	*históricos*	análisis	*político*	mango
elefante	*princesas*	radiador	*tomates*	internet	ventilador
fútbol[a]	profesor	*formulario*	*fútbol*	*planeta*	*vacaciones*
clima	automóvil	mamut	internet	pirámides	*fútbol*
históricos	*tomates*	ventilador	*cámara*	*históricos*	máquina
formulario	internet	*saxofón*	automóvil	cardiólogos	mamut
vacaciones	cardiólogos	pirámides	elefante	*parasol*	*formulario*
brócoli	*universidad*	*princesas*	*números*	*princesas*	análisis
fórmula	*político*	máquina	*parasol*	radiador	elefante
mango	centímetro	horizonte	cardiólogos	*cámara*	*números*
horizonte	*cámara*	*universidad*	sofá	sofá	clima
planeta	radiador	*brócoli*	*mango*	*tomates*	centímetro
sofá	*saxofón*	profesor	centímetro	automóvil	*universidad*
ventilador	mamut	*político*	clima	horizonte	fórmula
parasol	análisis	vacaciones	fórmula	*brócoli*	profesor
números	*números*	*planeta*	*planeta*	*saxofón*	*saxofón*
análisis	*parasol*	fórmula	vacaciones	profesor	*brócoli*
mamut	ventilador	clima	*político*	fórmula	horizonte
saxofón	sofa	centímetro	profesor	*universidad*	automóvil
radiador	*planeta*	*mango*	*brócoli*	centímetro	*tomates*
cámara	horizonte	sofá	*universidad*	clima	sofá
centímetro	*mango*	cardiólogos	horizonte	*números*	*cámara*
político	fórmula	*parasol*	máquina	elefante	radiador

T1-A	T1-B	T2-A	T2-B	T3-A	T3-B
universidad	*brócoli*	*números*	*princesas*	análisis	*princesas*
cardiólogos	vacaciones	elefante	pirámides	*formulario*	*parasol*
internet	*formulario*	automóvil	*saxofón*	mamut	cardiólogos
tomates	*históricos*	*cámara*	ventilador	máquina	*históricos*
automóvil	clima	internet	mamut	*fútbol*	pirámides
profesor	*fútbol*	*fútbol*	*formulario*	vacaciones	*planeta*
princesas	elefante	*tomates*	radiador	ventilador	internet
pirámides	máquina	análisis	*históricos*	*mango*	*político*

[a]Filler items are in italics.

Appendix F - Questionnaire

Vragenlijst

1. Participant SONA nummer: _____

2. Geslacht: ☐ Man ☐ Vrouw

3. Geboortedatum en leeftijd: _____

4. Geboorteplaats: _____

5. Huidige woonplaats: _____

6. Nationaliteit vader: _____

7. Nationaliteit moeder: _____

8. Ben je op dit moment student? Zo ja, hoe heet je opleiding?

_____

9. Welke taal/talen werden er thuis gesproken toen je opgroeide?

_____

10. Welke tweede, derde, vierde, etc. talen heb je geleerd na je moedertaal? Geef aan op welk niveau je ze beheerst.

Taal	*Beginner*	*Intermediair*	*Gevorderd*
1.	A1 / A2	B1 / B2	C1 / C2
2.	A1 / A2	B1 / B2	C1 / C2
3.	A1 / A2	B1 / B2	C1 / C2
4.	A1 / A2	B1 / B2	C1 / C2
5.	A1 / A2	B1 / B2	C1 / C2
6.	A1 / A2	B1 / B2	C1 / C2

11. Ben je wel eens naar een Spaanssprekend land geweest *met als doel de taal te leren*?
Ja / Nee

11a. Wanneer? _____

11b. Waar? _____

11c. Hoe lang? _____ < 3 maanden; _____ tussen 3 en 6 maanden; _____ > 6 maanden.

12. Ben je wel eens naar een land geweest waar ze een andere vreemde taal spreken (niet Spaans) *met als doel de taal te leren*?

Ja / Nee

	Verblijf 1	*Verblijf 2*	*Verblijf 3*
Land/Regio			
Taal			
Doel van verblijf			
Datum			

13. Heb je wel eens Spaanse les gehad?

Ja / Nee

13a. Wanneer? _____

13b. Waar? _____

13c. Hoe lang? _____< 6 maanden; _____ tussen 6 en 12 maanden; _____ > 12 maanden.

14. Heb je muziekles/zangles gehad, buiten de verplichte muziekles op school?

_____

14a. Zo ja, welke instrumenten speel je (tel zang als instrument)?

Instrument	Hoe lang al?

14b. Zo ja, hoeveel uur in de week speel je je instrument(en)?

_____

14c. Zo ja, hoe lang speel je dit instrument/deze instrumenten al?

_____

15. Zit je in een band, koor, of andere groep waarin samen muziek wordt gemaakt?

_____

15b. Werk je hierbij met een dirigent?

_____

15c. Zo ja, in wat voor muziek groep zit je?

_____

15d. Zo ja, hoeveel uur in de week speel je hier mee?

_____

15e. Zo ja, hoe lang speel je hier al mee?

_____

16. Heb je auditieve of visuele problemen die niet middels een apparaat gecorrigeerd kunnen worden?

Ja / Nee

17. Heb je dyslexie?

Ja / Nee

18. Is het je opgevallen dat de spreker gebaarden?

18b. Weet je nog wat voor gebaar dat was?

Appendix G - Stepwise protocol experimental procedure

Voorbereiding

- Voldoende VVV-bonnen + snoep aanwezig
- Informatie document, toestemmingsformulier, vragenlijst, score formulier WM taak, vertaalde instructies muziek taak
- Luxaflex halfdicht
- Start laptop op
- Video + Audio-recorder op plek leggen

Binnenkomst

- Informatie document – toestemmingsformulier
- Vraag: Heb je al ervaring met het Spaans? Zo ja, wat is je ervaring?
- *Je gaat in dit experiment een aantal verschillende taken doen. In totaal zal het ongeveer een uur duren. Halverwege kan je even pauzeren. We gaan nu beginnen met de eerste taak.*

Zet camera + audio-recorder aan!

Read-aloud taak (T1 + T2) & Lexical stress training

- Noteer proefpersoon – versie
- PowerPoint T1 + T2
- *In deze taak ga je zinnen voorlezen in het Spaans. Bij elke zin zie je een afbeelding die de betekenis van de zin duidelijk maakt. Deze betekenis hoef je niet te leren.*
- *Probeer de zinnen zo goed mogelijk hardop voor te lezen. Je mag meerdere pogingen doen. Je gaat automatisch naar de volgende zin. Er volgt eerst een voorbeeld.*
- Training:
    - Ga naar de volgende slide
    - Koptelefoon op
    - *Je gaat nu oefenen met de zinnen die je net hebt gelezen. Op de volgende slides volgt een uitleg over hoe beklemtoning werkt in het Spaans. Lees deze uitleg goed door en oefen met de voorbeelden.* ***Probeer de voorbeelden ook hardop na te zeggen.*** *Let goed op, want deze regels zal je zo moeten toepassen in de Spaanse zinnen. Je kan zelf door deze training heen klikken met de muis.*

- Geen instructie voor T2 – Alleen wanneer proefpersoon vragen heeft!

<u>Zet audio-recorder op pauze!</u>

Muziek taak

- Open taak
- NL instructie op papier + gesproken instructie
- Je gaat nu een korte muziektaak doen. Hierin ga je telkens twee geluidsfragmenten horen en een derde vergelijkingsfragment. Het is jouw taak om te bepalen of het vergelijkingsfragment hetzelfde of anders is van de eerder gehoorde fragmenten. Deze taak duurt ongeveer 10 minuten.
- Koptelefoon op
- Proefpersoon-nummer: P…

Pauze

- Niet op telefoon
- Snoep
- Wc-bezoek

<u>Zet audio-recorder aan!</u>

WM taak

- Taak in OpenSesame
    - Voer pp nummer in
    - Toetsenbord: 8 = goed / 9 = fout
- *Je krijgt zo een aantal getallen te horen. Je moet deze getallen andersom nazeggen. Je begint dus met het getal dat je als laatste hebt gehoord. Als je 2 – 7 hoort, dan zeg jij 7 – 2. Als je 9 – 3 hoort, dan zeg jij 3 – 9. Je gaat nu eerst oefenen.*
- Noteer score op los formulier

Read-aloud taak (T3)

- PowerPoint T3
- Selecteer de juiste versie

- *Je gaat nu dezelfde Spaanse zinnen uit de eerste taak (in een andere volgorde) nogmaals op het scherm zien verschijnen. Probeer de zinnen zo goed mogelijk hardop voor te lezen. Je mag meerdere pogingen doen. Je gaat automatisch naar de volgende zin.*

Vragenlijst

- Neem vragenlijst mondeling af

Zet audio-recording en video-recording op PC!

Appendix H - Praat scripts

Extract target words

```
form Extract_and_save_intervals
	comment Select the Sound you want to extract the intervals from
	comment
	comment Do you want to extract all interval?
	comment If you don`t write here the label of the intervals you don`t want to extract.
	word do_not_extract ""
	comment If you do not want to extract empty intervals write ""

	word prefix
	word suffix
endform

marca_de_silencio$ = do_not_extract$
informante$ = prefix$
repeticion$ = suffix$
folder$ = chooseDirectory$ ("Choose the folder where the sounds will be saved:")
#falta que lo haga del sonido seleccionado
@extract_and_save_intervals: folder$

procedure extract_and_save_intervals: .directoryOutput$
	#.directoryOutput$ = chooseDirectory$: "Choose a directory to save all the new files in"
	if .directoryOutput$ <> ""
		.numberOfSelectedSounds = numberOfSelected ("Sound")
		if .numberOfSelectedSounds = 0
		pause Select the Sound you want to extract from and click OK
		endif
		.base$ = selected$ ("Sound")
		selectObject: "TextGrid " + .base$

		#compruebo que el grid sea de ese Sound
		.soundDur = Get total duration
		.gridDur = Get total duration
		if .soundDur<>.gridDur
			El sonido y el TextGrid tienen duraciones diferentes. Â¿Seguro que son del mismo
archivo?
		endif
```

```
		#me apetecia que el margen de los archivos fuera silencio puro, pero para limpiarlos no irÃ¡ bien...
		#selectObject: "Sound " + .base$
		# sampl_frequ = Get sampling frequency
		# silence = Create Sound from formula: "silence", 1, 0, 0.2, sampl_frequ, "0"
		# silence2 = Create Sound from formula: "silence", 1, 0, 0.2, sampl_frequ, "0"
		.numberOfIntervals = Get number of intervals: 1
		for .interval from 1 to .numberOfIntervals

			selectObject: "TextGrid " + .base$
			.codigo$ = Get label of interval: 1, .interval
			if .codigo$ <> marca_de_silencio$
			echo Extrayendo intervalo '.interval' '.codigo$'
				.int_start = Get start point: 1, .interval
				.int_end = Get end point: 1, .interval
				selectObject: "Sound " +.base$
				#extraigo los intervalos con un margen porque sino van muy justos, en el caso de que sea el primer sonido o el ÃƒÂºltimo no hay posibilidad de silencio
				if .interval = 1
					Extract part: .int_start, .int_end+0.2, "rectangular", 1, 0
				elif .interval = .numberOfIntervals
					Extract part: .int_start-0.2, .int_end, "rectangular", 1, 0
				else
					Extract part: .int_start-0.2, .int_end+0.2, "rectangular", 1, 0
				endif
				Save as WAV file: .directoryOutput$ + "/" + informante$ + .codigo$ + repeticion$ + ".wav"
				Remove
			endif
		endfor
	endif
endproc
```

Randomise and annotate files

```
## text grid maker.praat
## Originally created by the excellent Katherine Crosswhite
## Script modified by Mark Antoniou
## Modified by Eric Doty

## What does it do?
## This script opens all files in a directory. It creates a TextGrid for each of sound file, then opens the sound file and the TextGrid into the editor so you can add boundaries and labels.

## Leaving the "Word" field blank will open all sound files in a directory. By specifying a Word, you can open only those files that begin with a particular sequence of characters. For example, only tokens whose filenames begin with ba.

## The script will skip over any files in the directory that already have an associated .TextGrid file

# The following four lines will create a dialog box, asking for the directory location you want to use. The two variables, "Directory" and "Word" will be used later in the script, where they are referred to as "directory$" and "word$", the dollar sign indicating that they are both string variables.

form Enter directory and search string
# Be sure not to forget the slash (Windows: backslash, OSX: forward slash) at the end of the directory name.
	sentence Directory C:\Users\u537205\Downloads\Coded data - tweede helft\
	sentence Word
	sentence Filetype wav
endform

clearinfo

# Make a list of all sound files in the directory.
Create Strings as file list... file-list 'directory$"word$'*.'filetype$'

# Loop for all files.
number_of_files = Get number of strings
for x from 1 to number_of_files
x = randomInteger (1,2735)

# Now we will set up a string variable called "current_file$" and use it to store the first filename from the list.
	select Strings file-list
	current_file$ = Get string... x
```

```
# Now that we have the filename, we read in that file:
	Read from file... 'directory$"current_file$'

# A variable called "object_name$" will have the name of the sound object. This is equivalent to the filename minus the extension. This will be useful for referring to the sound object later.
	object_name$ = selected$ ("Sound")

# Check if TextGrid exists for the file
	textgrid_filename$ = directory$ + object_name$ + ".TextGrid"
	if not fileReadable (textgrid_filename$)

# Now create a TextGrid for the current sound file. It will have only one tier named "segments". You can have multiple tiers, each with its own name. For example, I could've made three tiers by saying To TextGrid... "utterances words segments".
		To TextGrid... "segments"

# Since we have just created a TextGrid, it is automatically selected. We need both the TextGrid and the sound object to be selected together, so we must add the sound object to the selection.
		plus Sound 'object_name$'

# We want to open those two selected objects (Sound object and Textgrid object) in the editor.
		Edit

# The script will pause, allowing the user to enter the appropriate marks using the mouse and keyboard. Note that the user does not need to save the textgrid. They will click on "continue" to move to the next sound.
		pause Mark your segments.

# We will save the TextGrid object, so that the user doesn't have to do it for each file. First, deselect the sound object, leaving only the TextGrid selected.
		minus Sound 'object_name$'

# Save the textgrid, giving it the same filename as the sound file, and the extension ".TextGrid".
		Save as text file... 'directory$"object_name$'.TextGrid
	endif

#  End the loop, and go on to the next file. To conserve memory, first remove the objects that we are through with. I like to do this by selecting all the objects in the list, then deselecting any we will still be using, such as the list of filenames.
	select all
	minus Strings file-list
```

```
        Remove

# This specifies the end of the loop.
endfor

# Clean up the Praat objects window.
select Strings file-list
Remove

# Display a message letting you know that you've reached the end of the list.
printline TextGrids have been created for 'word$'.'filetype$' files in
printline 'directory$'
```

Create large datafile

```
# This script opens each sound file in a directory, looks for a corresponding TextGrid in (possibly) a different directory, and extracts f0, F1, and F2 from the midpoint(s) of any labelled interval(s) in the specified TextGrid tier. It also extracts the duration of the labelled interval(s). All these results are written to a tab-delimited text file.
# The script is a modified version of the script "collect_formant_data_from_files.praat" by Mietta Lennes, available here: http://www.helsinki.fi/~lennes/praat-scripts/
# The modifications were done by Dan McCloy (drmccloy@uw.edu) in December 2011.

# This script is distributed under the GNU General Public License.
# Copyright 4.7.2003 Mietta Lennes

form Get pitch formants and duration from labeled segments in files
	comment Directory of sound files. Be sure to include the final "/"
	text sound_directory /home/dan/Desktop/sound files/
	sentence Sound_file_extension .wav
	comment Directory of TextGrid files. Be sure to include the final "/"
	text textGrid_directory /home/dan/Desktop/text grids/
	sentence TextGrid_file_extension .TextGrid
	comment Full path of the resulting text file:
	text resultsfile /home/dan/Desktop/resultsfile.txt
	comment Which tier do you want to analyze?
	integer Tier 1
	comment Formant analysis parameters
	positive Time_step 0.01
	integer Maximum_number_of_formants 5
	positive Maximum_formant_(Hz) 5500
	positive Window_length_(s) 0.025
	real Preemphasis_from_(Hz) 50
	comment Pitch analysis parameters
	positive Pitch_time_step 0.01
	positive Minimum_pitch_(Hz) 75
	positive Maximum_pitch_(Hz) 300
endform

# Make a listing of all the sound files in a directory:
Create Strings as file list... list 'sound_directory$'*'sound_file_extension$'
numberOfFiles = Get number of strings

# Check if the result file exists:
```

```
if fileReadable (resultsfile$)
	pause The file 'resultsfile$' already exists! Do you want to overwrite it?
	filedelete 'resultsfile$'
endif

# Create a header row for the result file: (remember to edit this if you add or change the analyses!)
header$ = "Filename	TextGridLabel	duration f0_midpoint	F1_midpoint
	F2_midpoint'newline$"'
fileappend '"resultsfile$"' 'header$'

# Open each sound file in the directory:
for ifile to numberOfFiles
	filename$ = Get string... ifile
	Read from file... 'sound_directory$"filename$'

	# get the name of the sound object:
	soundname$ = selected$ ("Sound", 1)

	# Look for a TextGrid by the same name:
	gridfile$ = '"textGrid_directory$"soundname$"textGrid_file_extension$"'

	# if a TextGrid exists, open it and do the analysis:
	if fileReadable (gridfile$)
		Read from file... 'gridfile$'

		select Sound 'soundname$'
		To Formant (burg)... time_step maximum_number_of_formants maximum_formant
window_length preemphasis_from

		select Sound 'soundname$'
		To Pitch... pitch_time_step minimum_pitch maximum_pitch

		select TextGrid 'soundname$'
		numberOfIntervals = Get number of intervals... tier

		# Pass through all intervals in the designated tier, and if they have a label, do the analysis:
		for interval to numberOfIntervals
			label$ = Get label of interval... tier interval
			if label$ <> ""
				# duration:
```

```
                                start = Get starting point... tier interval
                                end = Get end point... tier interval
                                duration = end-start
                                midpoint = (start + end) / 2

                                # formants:
                                select Formant 'soundname$'
                                f1_50 = Get value at time... 1 midpoint Hertz Linear
                                f2_50 = Get value at time... 2 midpoint Hertz Linear

                                # pitch:
                                select Pitch 'soundname$'
                                f0_50 = Get value at time... midpoint Hertz Linear

                                # Save result to text file:
                                resultline$ = "'soundname$'        'label$'   'duration'        'f0_50'
        'f1_50'   'f2_50"newline$"'
                                fileappend "'resultsfile$'" 'resultline$'

                                # select the TextGrid so we can iterate to the next interval:
                                select TextGrid 'soundname$'
                        endif
                endfor
                # Remove the TextGrid, Formant, and Pitch objects
                select TextGrid 'soundname$'
                plus Formant 'soundname$'
                plus Pitch 'soundname$'
                Remove
        endif
        # Remove the Sound object
        select Sound 'soundname$'
        Remove
        # and go on with the next sound file!
        select Strings list
endfor

# When everything is done, remove the list of sound file paths:
Remove
```

Appendix I - R scripts

Calculate IRR

```
### CALCULATING INTER RATER RELIABILITY - READ ALOUD TASK 10% DATA
# Install packages
install.packages("lmerTest", "lme4", "Matrix", "lpSolve", "irr")
library("lmerTest", "lme4", "Matrix", "lpSolve", "irr")

# Importing IRR data in R - 10% data
IRRReadAloudData = read.csv("Results_Lieke(rater1)_Lisette(rater2).csv", sep = ";", header = TRUE)

# Inspect the data
data("IRRReadAloudData", package = "irr")
head(IRRReadAloudData[, 1:3])

# Cohen's Kappa - to asses Inter Rater Reliability
irr::kappa2(IRRReadAloudData[, 2:3], weight = "unweighted")

# Importing IRR data in R - for 2nd and 3rd trial 10% data
IRRReadAloudData_v2_3 = read.csv("Results_Lieke(rater1)_Lisette(rater2)_poging2_verbeteringen.csv", sep =
";", header = TRUE)

# Inspect the data
data("IRRReadAloudData_v2_3", package = "irr")
head(IRRReadAloudData_v2_3[, 1:4])

# Cohen's Kappa - to asses Inter Rater Reliability at Trial 2
irr::kappa2(IRRReadAloudData_v2_3[, 2:3], weight = "unweighted")

# Cohen's Kappa - to asses Inter Rater Reliability at Trial 3
irr::kappa2(IRRReadAloudData_v2_3[,c(2,4)], weight = "unweighted")
```

Calculate correlations, Cronbach's Alpha, and PCA

```
### CORRELATIONS - MUSICAL APTITUDE TASK
# Install packages
install.packages("Hmisc", "ggplot2", "polycor")
library("Hmisc", "ggplot2", "polycor")

# Load data in dataframe
MusicDataCor <- read.delim("Results_Musictask_v2.csv", sep = ";", header = TRUE)

# Remove total + participant column from dataframe
MusicDataCor_MRA <- MusicDataCor[,c(2,3,4)]

# Change factor into numeric values for correlation
MusicDataCor_MRA$Melody <- as.integer(MusicDataCor$Melody)
MusicDataCor_MRA$Rhythm <- as.integer(MusicDataCor$Rhythm)
MusicDataCor_MRA$Accent <- as.integer(MusicDataCor$Accent)

# Draw normality curve - histogram
# Melody
hist.melody <- ggplot2::ggplot(MusicDataCor_MRA, ggplot2::aes(Melody)) + ggplot2::labs(legend.position =
"none") + ggplot2::geom_histogram(ggplot2::aes(y = ..density..), colour = "black", fill = "white") +
ggplot2::labs(x = "Melody score", y = "Density")
hist.melody
hist.melody + stat_function(fun = dnorm, args = list(mean = mean(MusicDataCor_MRA$Melody, na.rm =
TRUE), sd = sd(MusicDataCor_MRA$Melody, na.rm = TRUE)), colour = "black", size = 1)

# Rhythm
hist.rhythm <- ggplot2::ggplot(MusicDataCor_MRA, ggplot2::aes(Rhythm)) + ggplot2::labs(legend.position =
"none") + ggplot2::geom_histogram(ggplot2::aes(y = ..density..), colour = "black", fill = "white") +
ggplot2::labs(x = "Rhythm score", y = "Density")
hist.rhythm
hist.rhythm + stat_function(fun = dnorm, args = list(mean = mean(MusicDataCor_MRA$Rhythm, na.rm =
TRUE), sd = sd(MusicDataCor_MRA$Rhythm, na.rm = TRUE)), colour = "black", size = 1)

# Accent
hist.accent <- ggplot2::ggplot(MusicDataCor_MRA, ggplot2::aes(Accent)) + ggplot2::labs(legend.position =
"none") + ggplot2::geom_histogram(ggplot2::aes(y = ..density..), colour = "black", fill = "white") +
ggplot2::labs(x = "Accent score", y = "Density")
hist.accent
```

```
hist.accent + stat_function(fun = dnorm, args = list(mean = mean(MusicDataCor_MRA$Accent, na.rm = TRUE), sd = sd(MusicDataCor_MRA$Accent, na.rm = TRUE)), colour = "black", size = 1)

# Draw Q-Q plot
qqplot.melody <- qplot(sample = MusicDataCor_MRA$Melody, stat = "qq")
qqplot.melody
qqplot.rhythm <- qplot(sample = MusicDataCor_MRA$Rhythm, stat = "qq")
qqplot.rhythm
qqplot.accent <- qplot(sample = MusicDataCor_MRA$Accent, stat = "qq")
qqplot.accent

# Quantifying normality scores
pastecs::stat.desc(cbind(MusicDataCor_MRA$Melody, MusicDataCor_MRA$Rhythm, MusicDataCor_MRA$Accent), basic = FALSE, norm = TRUE)
## Skew & Kurtosis should be 0 in Normal distribution. The further from 0, the more likely the data is not normally distributed
## Positive values skew: a pile-up on the left of the normal distribution
## Negative values kurtosis: a flat and light-tailed distribution
## skew.2SE & kurt.2SE < 1, so not significant, so normally distributed
## Normtest.W: Shapiro-Wilk test, normtest.p: p-value Shapiro-Wilk-test --> not significant, so data is normally distributed

# Pearson's correlations
cor(MusicDataCor_MRA)
## A weak to moderate positive relation between variables.

# create matrix to calculate p-values of correlations
MusicDataCor_MRAMatrix <- as.matrix(MusicDataCor_MRA)
rcorr(MusicDataCor_MRAMatrix)
## significant p-value indicates that the probability of getting a correlation coefficient this big in this sample is the null hypothesis (i.e., that there is no relation between variables) were true is very low
# get confidence intervals
correlation_Melody_Rhythm <- cor.test(x = MusicDataCor_MRA$Melody, y = MusicDataCor_MRA$Rhythm, method = "pearson")
correlation_Melody_Rhythm
correlation_Melody_Accent <- cor.test(x = MusicDataCor_MRA$Melody, y = MusicDataCor_MRA$Accent, method = "pearson")
correlation_Melody_Accent
correlation_Accent_Rhythm <- cor.test(x = MusicDataCor_MRA$Accent, y = MusicDataCor_MRA$Rhythm, method = "pearson")
```

```
correlation_Accent_Rhythm
## confidence interval does not cross 0, so likely that variables are positively related

# calculating R^2
cor(MusicDataCor_MRA)^2
# in percentages
cor(MusicDataCor_MRA)^2 * 100
## E.G. So, R^2 for the correlation between rhythm and melody is .18, indicating that rhythm only accounts for 18% of the variation.

## CALCULATING CRONBACH'S ALPHA - MUSICAL APTITUDE TASK
install.packages("psych")
library("psych")
# create datasets from subscales
MusicalAptitudeScale <- MusicDataCor_MRAMatrix[, c(1, 2, 3)]

# calculate cronbach's alpha
psych:: alpha(MusicalAptitudeScale)
## alpha of .7, probably indicates good reliability
## reliability if an item is dropped --> reliability does not improve by deleting an item, so all items stay in scale
## r.drop --> correlation of that item with the scale total if that item isn't included in the scale total (item-rest correlation) --> all items seem to correlate with the scale total
## The musical aptitude subscale had a high reliability with a Cronbach's alpha = .71.

### PRINCIPAL COMPONENT ANALYSIS – MUSICAL APTITUDE TASK
# Install packages
install.packages("corpcor", "GPArotation", "psych")
library("corpcor", "GPArotation", "psych")

# Load data in dataframe
MusicDataPCA <- read.delim("Results_Musictask_v2.csv", sep = ";", header = TRUE)
# Remove total column from dataframe
MusicDataPCA_MRA <- MusicDataPCA[,c(2,3,4)]

str(MusicDataPCA_MRA) # change factor into numeric values for correlation
MusicDataPCA_MRA$Melody <- as.integer(MusicDataPCA$Melody)
MusicDataPCA_MRA$Rhythm <- as.integer(MusicDataPCA$Rhythm)
MusicDataPCA_MRA$Accent <- as.integer(MusicDataPCA$Accent)
```

```
# Create correlation matrix
MusicDataPCA_MRAMatrix <- cor(MusicDataPCA_MRA)

# Bartlett's test
psych::cortest.bartlett(MusicDataPCA_MRA, 60)
# significant, so there are some relationships between the variables, hence PCA is appropritate

# Calculate KMO
# Determinant of correlation matrix
det(MusicDataPCA_MRAMatrix)
## greater than necessarily value of 0.0001, so not problematic

# PCA - 3 factors
PC1 <- psych::principal(MusicDataPCA_MRAMatrix, nfactors = 3, rotate = "none")
PC1

# create scree plot
plot(PC1$values, type = "b")

# PCA - 1 factor
PC2 <- psych::principal(MusicDataPCA_MRAMatrix, nfactors = 1, rotate = "none")
PC2

# reproduced correlation matrix
# diagonal --> communalities after extraction for each variable
factor.model(PC2$loadings)

# difference between reproduced correlation matrix and actual correlation matrix (residuals)
# diagonal --> the uniqueness

# check is residuals > .05
residuals <- factor.residuals(MusicDataPCA_MRAMatrix, PC2$loadings)
residuals <- as.matrix(residuals[upper.tri(residuals)])
large.resid <- abs(residuals) > 0.05
sum(large.resid)
sum(large.resid)/nrow(residuals)

sqrt(mean(residuals^2))
hist(residuals)
# varimax rotation
```

```
PC3 <- principal(MusicDataPCA_MRAMatrix, nfactors = 1, rotate = "varimax")
PC3
```

Explore background information (descriptives and frequencies)

```
# Graph - effect Trial x Lexical stress placement
library(scales)
ggplot(data = subset(TotalDataframe, !is.na(Correct.stress)),
    aes(x = Trial,
      fill = Correct.stress)) +
 geom_bar(position = "fill") +
 scale_y_continuous(breaks = seq(0, 1, .2),
          label = percent) +
 scale_fill_brewer(palette = "Set2") +
 labs(y = "Percent",
    fill = "Lexical stress placement",
    x = "Trial") +
 theme_minimal()

# View descriptives for interval data
library(psych)
describe(TotalDataframe$Melody)
describe(TotalDataframe$Rhythm)
describe(TotalDataframe$Accent)
describe(TotalDataframe$WorkingMemory)
describe(TotalDataframe$Age)
describe(TotalDataframe$Language_Home_Amount)

# view frequencies for categorical data
table(TotalDataframe$Condition)
table(TotalDataframe$Trial)
table(TotalDataframe$Rule_Accent_vs_NoAccent)
table(TotalDataframe$Gender)
table(TotalDataframe$Spanish_lessons)
table(TotalDataframe$Musicgroup)
table(TotalDataframe$Audio_Visual_Impairment)
table(TotalDataframe$Dyslexia)
table(TotalDataframe$Gestures_speaker)
table(TotalDataframe$Items)
table(TotalDataframe$Educational.level)
table(TotalDataframe$Student)
table(TotalDataframe$Educational_level_2)
table(TotalDataframe$Student_level)
table(TotalDataframe$Language_Home_Amount)
```

```
# Test whether multilingual subjects do not perform significantly better compared to monolingual subjects
summary(aov(Language_Home_Amount ~ Correct.stress, data = TotalDataframe))

# Test whether subjects who had Spanish lessons do not perform significantly better compared to subjects who have no experience
library(gmodels)
CrossTable(TotalDataframe$Correct.stress, TotalDataframe$Spanish_lessons, fisher= TRUE, chisq = TRUE, expected = TRUE, sresid = TRUE, format = "SPSS")

# Test effect of AV-impairment on outcome
CrossTable(TotalDataframe$Correct.stress, TotalDataframe$Audio_Visual_Impairment, fisher= TRUE, chisq = TRUE, expected = TRUE, sresid = TRUE, format = "SPSS")

# Test effect of dyslexia on outcome
CrossTable(TotalDataframe$Correct.stress, TotalDataframe$Dyslexia, fisher= TRUE, chisq = TRUE, expected = TRUE, sresid = TRUE, format = "SPSS")

# table age by condition
describeBy(TotalDataframe, group = "Condition")

# table age by gender
describeBy(TotalDataframe, group = "Gender")

summary(aov(Age ~ Gender, data = TotalDataframe))

# table gender by condition + chi-square
CrossTable(TotalDataframe$Condition, TotalDataframe$Gender, fisher= TRUE, chisq = TRUE, expected = TRUE, sresid = TRUE, format = "SPSS", simulate.p.value=TRUE, B=100000)

# table student_level by condition + chi-square
CrossTable(TotalDataframe$Condition, TotalDataframe$Student_level, fisher= TRUE, chisq = TRUE, expected = TRUE, sresid = TRUE, format = "SPSS", simulate.p.value=TRUE, B=100000)

# anova age by condition
summary(aov(Age ~ Condition, data = TotalDataframe))

# table amount of languages by condition
describeBy(TotalDataframe, group = "Condition")
```

```
# anova amount of languages by condition
summary(aov(Language_Home_Amount ~ Condition, data = TotalDataframe))

# table AV-impairment by condition + chi-square
CrossTable(TotalDataframe$Condition, TotalDataframe$Audio_Visual_Impairment, fisher= TRUE, chisq =
TRUE, expected = TRUE, sresid = TRUE, format = "SPSS", simulate.p.value=TRUE, B=100000)

# table Dyslexia by condition + chi-square
CrossTable(TotalDataframe$Condition, TotalDataframe$Dyslexia, fisher= TRUE, chisq = TRUE, expected =
TRUE, sresid = TRUE, format = "SPSS", simulate.p.value=TRUE, B=100000)

# table Spanish lessons by condition + chi-square
CrossTable(TotalDataframe$Condition, TotalDataframe$Spanish_lessons, fisher= TRUE, chisq = TRUE,
expected = TRUE, sresid = TRUE, format = "SPSS", simulate.p.value=TRUE, B=100000)
```

Explore influence practice items

```
### PREPARE DATAFILE
# Set working directionary
setwd("C:/Users/Lisette/Desktop/Thesis/Data")

# Importing data into R: Musical Aptitude Scores, Working Memory Scores, and Questionnaire
MusicalAptitudeData = read.csv("Results_Musictask_v2.csv", sep = ";", header = TRUE)
WorkingMemoryData = read.csv("Scoreformulier_WMtaak.csv", sep = ";", header = TRUE)
QuestionnaireData = read.csv("Vragenlijst.csv", sep = ";", header = TRUE)
ReadAloudTaskData = read.csv("results_100%_analyse_NA.csv", sep = ";", header = TRUE)

# Create dataframe
# merge dataframes by participant
MusicalAptitude_WorkingMemory_Dataframe <- merge(MusicalAptitudeData,WorkingMemoryData,by =
"ï..Participant")
MusicalAptitude_WorkingMemory_Questionnaire_Dataframe <-
merge(MusicalAptitude_WorkingMemory_Dataframe,QuestionnaireData,by = "ï..Participant")
TotalDataframe <- merge(ReadAloudTaskData,MusicalAptitude_WorkingMemory_Questionnaire_Dataframe,
by = "ï..Participant" )

# Change numeric into factor values
TotalDataframe$Itemvs <- as.factor(TotalDataframe$Itemvs)
TotalDataframe$Condition <- as.factor(TotalDataframe$Condition)
TotalDataframe$Time <- as.factor(TotalDataframe$Time)
TotalDataframe$Item <- as.factor(TotalDataframe$Item)

# Rename variables
colnames(TotalDataframe)
names(TotalDataframe)[names(TotalDataframe) == "Time"] <- "Trial"

# Attach value labels to factors
TotalDataframe$Itemvs <- factor(TotalDataframe$Itemvs,
                levels = c(0,1),
                labels = c("Untrained Items", "Trained Items"))
TotalDataframe$Condition <- factor(TotalDataframe$Condition,
                levels = c(1,2,3,4,5),
                labels = c("AV", "AVB-Perc", "AVB-Prod", "AVM-Perc", "AVM-Prod"))
TotalDataframe$Trial <- factor(TotalDataframe$Trial,
               levels = c(1,2,3),
               labels = c("T1", "T2", "T3"))
```

```
TotalDataframe$Item <- factor(TotalDataframe$Item,
          levels = c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17),
          labels = c("E1", "E2", "E3", "E4", "E5", "E6", "E7", "E8", "E9", "E10", "E11", "E12", "E13", "E14",
"O1", "O2", "O3"))
TotalDataframe$Correct.stress <- factor(TotalDataframe$Correct.stress,
              levels = c(0,1),
              labels = c("Incorrect", "Correct"))

# Change order levels of factors
TotalDataframe$Trial <- relevel(TotalDataframe$Trial, "T1")
TotalDataframe$Correct.stress <- relevel(TotalDataframe$Correct.stress, "Incorrect")

# Explore effect of practice-items
library(gmodels)
CrossTable(TotalDataframe$Correct.stress, TotalDataframe$Itemvs, fisher= TRUE, chisq = TRUE, expected =
TRUE, sresid = TRUE, format = "SPSS", simulate.p.value=TRUE, B=100000)

library(scales)
ggplot(data = subset(TotalDataframe, !is.na(Correct.stress)),
   aes(x = Itemvs,
     fill = Correct.stress)) +
 geom_bar(position = "fill") +
 scale_y_continuous(breaks = seq(0, 1, .2),
         label = percent) +
 scale_fill_brewer(palette = "Set2") +
 labs(y = "Percentage Correct/Incorrect Lexical Stress Placement",
   fill = "Spanish Lexical Stress Placement",
   x = "") +
 theme_minimal()

# Explore interaction of practice items and trial
item_trial_test <- glm(Correct.stress ~ Itemvs*Trial, data = TotalDataframe, family = binomial)
summary(item_trial_test)

cat_plot(item_trial_test, pred = Trial, modx = Itemvs, geom = "bar",
    legend.main = "",
    x.label = "Time of Testing",
    y.label = "Ratio Correct Spanish Lexical Stress Placement")
```

```
# Explore interaction of practice items and condition
item__condition_test <- glm(Correct.stress ~ Itemvs*Condition, data = TotalDataframe, family = binomial)
summary(item_condition_test)
```

Multilevel logistic regression analysis

```
### PREPARE DATAFILE
# Set working directionary
setwd("C:/Users/Lisette/Desktop/Thesis/Data")

# Importing data into R: Musical Aptitude Scores, Working Memory Scores, and Questionnaire
MusicalAptitudeData = read.csv("Results_Musictask_v2.csv", sep = ";", header = TRUE)
WorkingMemoryData = read.csv("Scoreformulier_WMtaak.csv", sep = ";", header = TRUE)
QuestionnaireData = read.csv("Vragenlijst.csv", sep = ";", header = TRUE)
ReadAloudTaskData = read.csv("results_100%_analyse_NA_zO.csv", sep = ";", header = TRUE)

# Rename variables
names(MusicalAptitudeData)[names(MusicalAptitudeData) == "ï..Participant"] <- "Participant"
names(WorkingMemoryData)[names(WorkingMemoryData) == "ï..Participant"] <- "Participant"
names(QuestionnaireData)[names(QuestionnaireData) == "ï..Participant"] <- "Participant"

# Create dataframe
# merge dataframes by participant
MusicalAptitude_WorkingMemory_Dataframe <- merge(MusicalAptitudeData,WorkingMemoryData,by =
"Participant")
MusicalAptitude_WorkingMemory_Questionnaire_Dataframe <-
merge(MusicalAptitude_WorkingMemory_Dataframe,QuestionnaireData,by = "Participant")
TotalDataframe <- merge(ReadAloudTaskData,MusicalAptitude_WorkingMemory_Questionnaire_Dataframe,
by = "Participant" )

# Change numeric into factor values
TotalDataframe$Participantvs <- as.factor(TotalDataframe$Participantvs)
TotalDataframe$Participant <- as.factor(TotalDataframe$Participant)
TotalDataframe$Versie <- as.factor(TotalDataframe$Versie)
TotalDataframe$Condition <- as.factor(TotalDataframe$Condition)
TotalDataframe$Time <- as.factor(TotalDataframe$Time)
TotalDataframe$Rule <- as.factor(TotalDataframe$Rule)
TotalDataframe$Rule.1.2.vs..3 <- as.factor(TotalDataframe$Rule.1.2.vs..3)
TotalDataframe$Item <- as.factor(TotalDataframe$Item)
TotalDataframe$Correct.stress <- as.factor(TotalDataframe$Correct.stress)
TotalDataframe$Gender <- as.factor(TotalDataframe$Gender)
TotalDataframe$Spanish_lessons <- as.factor(TotalDataframe$Spanish_lessons)
TotalDataframe$Muziek_les <- as.factor(TotalDataframe$Muziek_les)
TotalDataframe$Musicgroup <- as.factor(TotalDataframe$Musicgroup)
TotalDataframe$Audio_Visual_Impairment <- as.factor(TotalDataframe$Audio_Visual_Impairment)
```

```
TotalDataframe$Dyslexia <- as.factor(TotalDataframe$Dyslexia)
TotalDataframe$Gestures_speaker <- as.factor(TotalDataframe$Gestures_speaker)
TotalDataframe$Educational.level <- as.factor(TotalDataframe$Educational.level)
TotalDataframe$Student <- as.factor(TotalDataframe$Student)
TotalDataframe$Educational.level <- as.factor(TotalDataframe$Educational_level_2)
TotalDataframe$Student_level <- as.factor(TotalDataframe$Student_level)

# Rename variables
colnames(TotalDataframe)
names(TotalDataframe)[names(TotalDataframe) == "Laatst.gehaalde.blok"] <- "WorkingMemory"
names(TotalDataframe)[names(TotalDataframe) == "Time"] <- "Trial"
names(TotalDataframe)[names(TotalDataframe) == "Rule.1.2.vs..3"] <- "Rule_Accent_vs_NoAccent"
names(TotalDataframe)[names(TotalDataframe) == "Muziek_les"] <- "Music_lessons"

# Rename variable levels
library("plyr")
TotalDataframe$Rule_Accent_vs_NoAccent <- revalue(TotalDataframe$Rule_Accent_vs_NoAccent,
c("1"="0", "2"="1"))

# Attach value labels to factors
TotalDataframe$Participantvs <- factor(TotalDataframe$Participantvs,
                levels = c(0,1),
                labels = c("met", "not met"))
TotalDataframe$Condition <- factor(TotalDataframe$Condition,
               levels = c(1,2,3,4,5),
               labels = c("AV", "AVB-Perc", "AVB-Prod", "AVM-Perc", "AVM-Prod"))
TotalDataframe$Trial <- factor(TotalDataframe$Trial,
              levels = c(1,2,3),
              labels = c("T1", "T2", "T3"))
TotalDataframe$Rule_Accent_vs_NoAccent <- factor(TotalDataframe$Rule_Accent_vs_NoAccent,
                   levels = c(0,1),
                   labels = c("NoAccent", "Accent"))
TotalDataframe$Rule <- factor(TotalDataframe$Rule,
            levels = c(1,2,3),
            labels = c("Rule1", "Rule2", "Rule3"))
TotalDataframe$Item <- factor(TotalDataframe$Item,
            levels = c(1,2,3,4,5,6,7,8,9,10,11,12,13,14),
            labels = c("E1", "E2", "E3", "E4", "E5", "E6", "E7", "E8", "E9", "E10", "E11", "E12", "E13", "E14"))
TotalDataframe$Correct.stress <- factor(TotalDataframe$Correct.stress,
                levels = c(0,1),
```

```
                labels = c("Incorrect", "Correct"))
TotalDataframe$Gender <- factor(TotalDataframe$Gender,
              levels = c(0,1),
              labels = c("Female", "Male"))
TotalDataframe$Spanish_lessons <- factor(TotalDataframe$Spanish_lessons,
                 levels = c(0,1),
                 labels = c("No", "Yes"))
TotalDataframe$Music_lessons <- factor(TotalDataframe$Music_lessons,
                levels = c(0,1),
                labels = c("No", "Yes"))
TotalDataframe$Musicgroup <- factor(TotalDataframe$Musicgroup,
                levels = c(0,1),
                labels = c("No", "Yes"))
TotalDataframe$Audio_Visual_Impairment <- factor(TotalDataframe$Audio_Visual_Impairment,
                    levels = c(0,1),
                    labels = c("No", "Yes"))
TotalDataframe$Dyslexia <- factor(TotalDataframe$Dyslexia,
              levels = c(0,1),
              labels = c("No", "Yes"))
TotalDataframe$Gestures_speaker <- factor(TotalDataframe$Gestures_speaker,
                levels = c(0,1),
                labels = c("No", "Yes"))
TotalDataframe$Educational.level <- factor(TotalDataframe$Educational.level,
                levels = c(1,2,3,4),
                labels = c("MBO", "HBO", "WO_BA", "WO_MA"))
TotalDataframe$Student <- factor(TotalDataframe$Student,
                 levels = c(0,1),
                 labels = c("No", "Yes"))
TotalDataframe$Educational_level_2 <- factor(TotalDataframe$Educational_level_2,
                 levels = c(2,3,4),
                 labels = c("MBO/HBO", "WO_BA", "WO_MA"))
TotalDataframe$Student_level <- factor(TotalDataframe$Student_level,
              levels = c(0,1,2,3),
              labels = c("No", "HBO", "WO_BA", "WO_MA"))

# Change order levels of factors
TotalDataframe$Trial <- relevel(TotalDataframe$Trial, "T1")
TotalDataframe$Correct.stress <- relevel(TotalDataframe$Correct.stress, "Incorrect")
```

```
# Rescale continuous variables
TotalDataframe$Melodyscale = scale(TotalDataframe$Melody, center = TRUE, scale = TRUE)
TotalDataframe$Rhythmscale = scale(TotalDataframe$Rhythm, center = TRUE, scale = TRUE)
TotalDataframe$Accentscale = scale(TotalDataframe$Accent, center = TRUE, scale = TRUE)
TotalDataframe$WorkingMemoryscale = scale(TotalDataframe$WorkingMemory, center = TRUE, scale =
TRUE)
TotalDataframe$Agescale = scale(TotalDataframe$Age, center = TRUE, scale = TRUE)
TotalDataframe$Language_Home_Amountscale = scale(TotalDataframe$Language_Home_Amount, center =
TRUE, scale = TRUE)

# Explore effect of XP participants
CrossTable(TotalDataframe$Correct.stress, TotalDataframe$Participantvs, fisher= TRUE, chisq = TRUE,
expected = TRUE, sresid = TRUE, format = "SPSS", simulate.p.value=TRUE, B=100000)

library(scales)
ggplot(data = subset(TotalDataframe, !is.na(Correct.stress)),
   aes(x = Participantvs,
     fill = Correct.stress)) +
 geom_bar(position = "fill") +
 scale_y_continuous(breaks = seq(0, 1, .2),
         label = percent) +
 scale_fill_brewer(palette = "Set2") +
 labs(y = "Scores (in Percentages) on the Read-Aloud Tasks",
   fill = "",
   x = "Pre-Defined Requirements") +
 theme_minimal()

### BASELINE MODEL
# Baseline with all control predictors
baseline <- glmer(Correct.stress ~ Gender + Spanish_lessons + Music_lessons + Musicgroup +
Audio_Visual_Impairment + Dyslexia + Gestures_speaker + Student_level + (1 |
Rule_Accent_vs_NoAccent:Item) + (1|Agescale) + (1|Language_Home_Amountscale), data = TotalDataframe,
family = binomial, control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05)))
summary(baseline)

# Baseline which excludes random predictors that explain zero or almost zero variance
baseline_b <- glmer(Correct.stress ~ Gender + Spanish_lessons + Music_lessons + Musicgroup +
Audio_Visual_Impairment + Dyslexia + Gestures_speaker + Student_level + (1 |
Rule_Accent_vs_NoAccent:Item), data = TotalDataframe, family = binomial, control = glmerControl(optimizer
= "bobyqa", optCtrl = list(maxfun = 2e+05)))
```

```
summary(baseline_b)
anova(baseline, baseline_b)

# Baseline which excludes control predictors that are not significant
baseline_c <- glmer(Correct.stress ~ Musicgroup + (1| Rule_Accent_vs_NoAccent:Item), data =
TotalDataframe, family = binomial, control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun =
2e+05)))

summary(baseline_c)
anova(baseline_b, baseline_c)

### Check assumptions baseline model
## Assumption of Linearity (only for continuous predictors)

# interaction of melody score with its log (creates a new variable in TotalDataframe)
TotalDataframe$logMelodyscoreInt <- log(TotalDataframe$Melody)*TotalDataframe$Melody
# interaction of rhythm score with its log
TotalDataframe$logRhythmscoreInt <- log(TotalDataframe$Rhythm)*TotalDataframe$Rhythm
# interaction of accent score with its log
TotalDataframe$logAccentscoreInt <- log(TotalDataframe$Accent)*TotalDataframe$Accent
# interaction of working memory score with its log
TotalDataframe$logWorkingMemoryscoreInt <-
log(TotalDataframe$WorkingMemory)*TotalDataframe$WorkingMemory

# GLM test
LinearityTest <- glm(Correct.stress ~ Melody + Rhythm + Accent + WorkingMemory + logMelodyscoreInt +
logRhythmscoreInt + logAccentscoreInt + logWorkingMemoryscoreInt, data = TotalDataframe,
family=binomial())
summary(LinearityTest)

## Assumption of Multicolinearity
library("rms")

MulticolinearityTest <- glm (Correct.stress ~ Trial + Condition + Melody + Rhythm + Accent +
WorkingMemory + Rule_Accent_vs_NoAccent + Item + Musicgroup, data = TotalDataframe, family =
binomial())

# VIF scores (VIF should be < 10)
rms::vif(MulticolinearityTest)
# Tolerance scores (Tolerance should be > .2)
```

```
1/rms::vif(MulticolinearityTest)

## Checking Overdispersion
library("lme4")
OverdispersionTest <- glmer(Correct.stress ~ Trial + Condition + Melody + Rhythm + Accent +
WorkingMemory + Rule_Accent_vs_NoAccent + Musicgroup + (1|Rule_Accent_vs_NoAccent:Item), data =
TotalDataframe, family = binomial())
summary(OverdispersionTest)

### MODEL 1 - ADD MAIN PREDICTORS
Model1 <- glmer(Correct.stress ~ Trial + Condition + Melodyscale + Rhythmscale + Accentscale +
WorkingMemoryscale + Rule_Accent_vs_NoAccent + Musicgroup + (1| Rule_Accent_vs_NoAccent:Item), data
= TotalDataframe, family = binomial, control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun =
2e+05)))
summary(Model1)

# compare to baseline model
anova(baseline_c, Model1)

## Remove non-significant predictors one-by-one
# Model 1b - remove rhythmscale from model
Model1_b <- glmer(Correct.stress ~ Trial + Condition + Melodyscale + Accentscale + WorkingMemoryscale +
Rule_Accent_vs_NoAccent + Musicgroup + (1| Rule_Accent_vs_NoAccent:Item), data = TotalDataframe,
family = binomial, control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05)))
summary(Model1_b)

# compare to Model with rhythmscale
anova(Model1, Model1_b)

# Model 1c - remove accentscale from model
Model1_c <- glmer(Correct.stress ~ Trial + Condition + Melodyscale + WorkingMemoryscale +
Rule_Accent_vs_NoAccent + Musicgroup + (1| Rule_Accent_vs_NoAccent:Item), data = TotalDataframe,
family = binomial, control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05)))
summary(Model1_c)

#compare to Model with accentscale
anova(Model1_b, Model1_c)

# compare model with all predictors to model without rhythm and accent
anova(Model1, Model1_c)
```

```
# compare baseline model_c with final model 1 (without rhythm and accent)
anova(baseline_c, Model1_c)

### MODEL 2 - ADD TWO-WAY INTERACTIONS
Model2 <- glmer(Correct.stress ~ Trial + Condition + Melodyscale + WorkingMemoryscale +
Rule_Accent_vs_NoAccent
      + Trial:Condition + Trial:WorkingMemoryscale + Trial:Melodyscale + Trial:Rhythmscale +
Trial:Accentscale + Trial:Rule_Accent_vs_NoAccent
      + Musicgroup + (1 | Rule_Accent_vs_NoAccent:Item), data = TotalDataframe, family = binomial, control
= glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05)))
summary(Model2)

# compare final model 1 to model with all two-way interactions
anova(Model1_c, Model2)

## Remove non-significant interactions one-by-one
#Model 2a - removing trial:rhythm
Model2_a <- glmer(Correct.stress ~ Trial + Condition + Melodyscale + WorkingMemoryscale +
Rule_Accent_vs_NoAccent
      + Trial:Condition + Trial:WorkingMemoryscale + Trial:Melodyscale + Trial:Accentscale +
Trial:Rule_Accent_vs_NoAccent
      + Musicgroup + (1 | Rule_Accent_vs_NoAccent:Item), data = TotalDataframe, family = binomial, control =
glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05)))
summary(Model2_a)

# compare model without trial:rhythm to model it
anova(Model2, Model2_a)

#Model2b - removing trial:workingmemory
Model2_b <- glmer(Correct.stress ~ Trial + Condition + Melodyscale + WorkingMemoryscale +
Rule_Accent_vs_NoAccent
      + Trial:Condition + Trial:Melodyscale + Trial:Accentscale + Trial:Rule_Accent_vs_NoAccent
      + Musicgroup + (1 | Rule_Accent_vs_NoAccent:Item), data = TotalDataframe, family = binomial, control
= glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05)))
summary(Model2_b)

# compare model without trial:working memory to model with it
anova(Model2_a, Model2_b)
```

```
#Model2c - removing trial:accent
Model2_c <- glmer(Correct.stress ~ Trial + Condition + Melodyscale + WorkingMemoryscale +
Rule_Accent_vs_NoAccent
      + Trial:Condition + Trial:Melodyscale + Trial:Rule_Accent_vs_NoAccent
      + Musicgroup + (1 | Rule_Accent_vs_NoAccent:Item), data = TotalDataframe, family = binomial, control
= glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05)))
summary(Model2_c)

# compare model without trial:accent to model with it
anova(Model2_b, Model2_c)
# compare model with all interactions to model excluding 3 interactions
anova(Model2, Model2_c)
# compare final model (without non-significant interactions) to Model 1
anova(Model1_c, Model2_c)

### Model3 - ADD THREE-WAY INTERACTIONS
Model3 <- glmer(Correct.stress ~ Trial + Condition + Melodyscale + WorkingMemoryscale +
Rule_Accent_vs_NoAccent
    + Trial:Condition + Trial:Melodyscale + Trial:Rule_Accent_vs_NoAccent
    + Trial:Condition:Rule_Accent_vs_NoAccent + Trial:Condition:WorkingMemoryscale +
Trial:Condition:Melodyscale + Trial:Condition:Rhythmscale + Trial:Condition:Accentscale
    + Trial:Rule_Accent_vs_NoAccent:WorkingMemoryscale + Trial:Rule_Accent_vs_NoAccent:Melodyscale
+ Trial:Rule_Accent_vs_NoAccent:Rhythmscale + Trial:Rule_Accent_vs_NoAccent:Accentscale
    + Musicgroup + (1 | Rule_Accent_vs_NoAccent:Item), data = TotalDataframe, family = binomial, control =
glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05)))
summary(Model3)

# compare model with three-way interactions to Model 2
anova(Model2_c, Model3)

### Model 3a - Remove non-significant three-way interactions (trail:condition:rule + trial:rule:WM +
trial:rule:accent )
Model3_a <- glmer(Correct.stress ~ Trial + Condition + Melodyscale + WorkingMemoryscale +
Rule_Accent_vs_NoAccent
      + Trial:Condition + Trial:Melodyscale + Trial:Rule_Accent_vs_NoAccent
      + Trial:Condition:WorkingMemoryscale + Trial:Condition:Melodyscale + Trial:Condition:Rhythmscale +
Trial:Condition:Accentscale
      + Trial:Rule_Accent_vs_NoAccent:Melodyscale + Trial:Rule_Accent_vs_NoAccent:Rhythmscale
      + Musicgroup + (1 | Rule_Accent_vs_NoAccent:Item), data = TotalDataframe, family = binomial, control =
glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+06)))
```

```
summary(Model3_a)

# compare model without non-significant interaction to model with it
anova(Model3, Model3_a)
# compare final model (without non-significant interaction) to Model 2
anova(Model2_c, Model3_a)

### MODEL 4 - ADD FOUR-WAY INTERACTIONS
Model4 <- glmer(Correct.stress ~ Trial + Condition + Melodyscale + WorkingMemoryscale +
Rule_Accent_vs_NoAccent
    + Trial:Condition + Trial:Melodyscale + Trial:Rule_Accent_vs_NoAccent
    + Trial:Condition:WorkingMemoryscale + Trial:Condition:Melodyscale + Trial:Condition:Rhythmscale +
Trial:Condition:Accentscale
    + Trial:Rule_Accent_vs_NoAccent:Melodyscale + Trial:Rule_Accent_vs_NoAccent:Rhythmscale
    + Trial:Condition:WorkingMemoryscale:Melodyscale + Trial:Condition:WorkingMemoryscale:Rhythmscale
+ Trial:Condition:WorkingMemoryscale:Accentscale
    + Trial:Condition:Rule_Accent_vs_NoAccent:WorkingMemoryscale +
Trial:Condition:Rule_Accent_vs_NoAccent:Melodyscale +
Trial:Condition:Rule_Accent_vs_NoAccent:Rhythmscale +
Trial:Condition:Rule_Accent_vs_NoAccent:Accentscale
    + Musicgroup + (1 | Rule_Accent_vs_NoAccent:Item), data = TotalDataframe, family = binomial, control =
glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05)))
summary(Model4)
# GIVES WARNING --> OPTIMIZE MAX_FUN

# Model 4 - with optimizer
Model4_o <- glmer(Correct.stress ~ Trial + Condition + Melodyscale + WorkingMemoryscale +
Rule_Accent_vs_NoAccent
      + Trial:Condition + Trial:Melodyscale + Trial:Rule_Accent_vs_NoAccent
      + Trial:Condition:WorkingMemoryscale + Trial:Condition:Melodyscale + Trial:Condition:Rhythmscale +
Trial:Condition:Accentscale
      + Trial:Rule_Accent_vs_NoAccent:Melodyscale + Trial:Rule_Accent_vs_NoAccent:Rhythmscale
      + Trial:Condition:WorkingMemoryscale:Melodyscale +
Trial:Condition:WorkingMemoryscale:Rhythmscale + Trial:Condition:WorkingMemoryscale:Accentscale
      + Trial:Condition:Rule_Accent_vs_NoAccent:WorkingMemoryscale +
Trial:Condition:Rule_Accent_vs_NoAccent:Melodyscale +
Trial:Condition:Rule_Accent_vs_NoAccent:Rhythmscale +
Trial:Condition:Rule_Accent_vs_NoAccent:Accentscale
      + Musicgroup + (1 | Rule_Accent_vs_NoAccent:Item), data = TotalDataframe, family = binomial, control =
glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+06)))
```

```
summary(Model4_o)

# compare final model 3 to model with four-way interactions
anova(Model3_a, Model4_o)

# Model 4b - remove non-significant four-way interactions (trial*condition*rule*rhythmscale)
Model4_b <- glmer(Correct.stress ~ Trial + Condition + Melodyscale + WorkingMemoryscale +
Rule_Accent_vs_NoAccent
      + Trial:Condition + Trial:Melodyscale + Trial:Rule_Accent_vs_NoAccent
      + Trial:Condition:WorkingMemoryscale + Trial:Condition:Melodyscale + Trial:Condition:Rhythmscale +
Trial:Condition:Accentscale
      + Trial:Rule_Accent_vs_NoAccent:Melodyscale + Trial:Rule_Accent_vs_NoAccent:Rhythmscale
      + Trial:Condition:WorkingMemoryscale:Melodyscale +
Trial:Condition:WorkingMemoryscale:Rhythmscale + Trial:Condition:WorkingMemoryscale:Accentscale
      + Trial:Condition:Rule_Accent_vs_NoAccent:WorkingMemoryscale +
Trial:Condition:Rule_Accent_vs_NoAccent:Melodyscale +
Trial:Condition:Rule_Accent_vs_NoAccent:Accentscale
      + Musicgroup + (1 | Rule_Accent_vs_NoAccent:Item), data = TotalDataframe, family = binomial, control
= glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+06)))
summary(Model4_b)

# compare model with non-significant interactions to model without them
anova(Model4_o, Model4_b)

# compare final model 3 to new model
anova(Model3_a, Model4_b)

### Create three new datasets by splitting file by Trial (to interpret four-way interactions)
TotalDataframe_split <- split(TotalDataframe, TotalDataframe$Trial)
str(TotalDataframe_split)
TotalDataframe_T1 <- TotalDataframe_split$T1
TotalDataframe_T2 <- TotalDataframe_split$T2
TotalDataframe_T3 <- TotalDataframe_split$T3

Model_T1 <- glmer(Correct.stress ~ Condition + Melodyscale + WorkingMemoryscale +
Rule_Accent_vs_NoAccent
      + Condition:WorkingMemoryscale + Condition:Melodyscale + Condition:Rhythmscale +
Condition:Accentscale
      + Rule_Accent_vs_NoAccent:Melodyscale + Rule_Accent_vs_NoAccent:Rhythmscale
```

```
    + Condition:WorkingMemoryscale:Melodyscale + Condition:WorkingMemoryscale:Rhythmscale +
Condition:WorkingMemoryscale:Accentscale
    + Condition:Rule_Accent_vs_NoAccent:WorkingMemoryscale +
Condition:Rule_Accent_vs_NoAccent:Melodyscale + Condition:Rule_Accent_vs_NoAccent:Accentscale
    + Musicgroup + (1 | Rule_Accent_vs_NoAccent:Item), data = TotalDataframe_T1, family = binomial,
control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+06)))
summary(Model_T1)

Model_T2 <- glmer(Correct.stress ~ Condition + Melodyscale + WorkingMemoryscale +
Rule_Accent_vs_NoAccent
    + Condition:WorkingMemoryscale + Condition:Melodyscale + Condition:Rhythmscale +
Condition:Accentscale
    + Rule_Accent_vs_NoAccent:Melodyscale + Rule_Accent_vs_NoAccent:Rhythmscale
    + Condition:WorkingMemoryscale:Melodyscale + Condition:WorkingMemoryscale:Rhythmscale +
Condition:WorkingMemoryscale:Accentscale
    + Condition:Rule_Accent_vs_NoAccent:WorkingMemoryscale +
Condition:Rule_Accent_vs_NoAccent:Melodyscale + Condition:Rule_Accent_vs_NoAccent:Accentscale
    + Musicgroup + (1 | Rule_Accent_vs_NoAccent:Item), data = TotalDataframe_T2, family = binomial,
control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+06)))
summary(Model_T2)

Model_T3 <- glmer(Correct.stress ~ Condition + Melodyscale + WorkingMemoryscale +
Rule_Accent_vs_NoAccent
    + Condition:WorkingMemoryscale + Condition:Melodyscale + Condition:Rhythmscale +
Condition:Accentscale
    + Rule_Accent_vs_NoAccent:Melodyscale + Rule_Accent_vs_NoAccent:Rhythmscale
    + Condition:WorkingMemoryscale:Melodyscale + Condition:WorkingMemoryscale:Rhythmscale +
Condition:WorkingMemoryscale:Accentscale
    + Condition:Rule_Accent_vs_NoAccent:WorkingMemoryscale +
Condition:Rule_Accent_vs_NoAccent:Melodyscale + Condition:Rule_Accent_vs_NoAccent:Accentscale
    + Musicgroup + (1 | Rule_Accent_vs_NoAccent:Item), data = TotalDataframe_T3, family = binomial,
control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+06)))
summary(Model_T3)
```

Plots regression model

```
### MAIN EFFECTS
# Explore - main effect trial
CrossTable(TotalDataframe$Correct.stress, TotalDataframe$Trial, fisher= TRUE, chisq = TRUE, expected =
TRUE, sresid = TRUE, format = "SPSS", simulate.p.value=TRUE, B=100000)

# Graph - main effect trial
library(gmodels)
ggplot(data = subset(TotalDataframe, !is.na(Correct.stress)),
    aes(x = Trial,
      fill = Correct.stress)) +
 geom_bar(position = "fill") +
 scale_y_continuous(breaks = seq(0, 1, .2),
          label = percent) +
 scale_fill_brewer(palette = "Set2") +
 labs(y = "Scores (in Percentages) on the Read-Aloud Tasks",
    fill = "Spanish Lexical Stress Placement",
    x = "Time of Testing") +
 theme_minimal()

library(dplyr)
data <- subset(TotalDataframe, !is.na(Correct.stress)) %>% group_by(Trial,Correct.stress) %>%
summarize(n=n()) %>% mutate(freq=n/sum(n))

ggplot(data, aes(x=Trial, fill = Correct.stress, group = Correct.stress)) + geom_bar(aes(y=freq), stat="identity",
position = "dodge") +
 scale_y_continuous(breaks = seq(0, 1, .2),
          label = percent) +
 scale_fill_brewer(palette = "Set2") +
 labs(y = "Scores (in Percentages) on Spanish Lexical Stress Placement",
    fill = "Spanish Lexical Stress Placement",
    x = "Time of Testing")

# Contrast analyses - explore difference between T2 and T3 (trial)
library(nlme)
library(emmeans)
em <- emmeans(Model4_b, "Trial")
contrast(em, adjust = "bonferroni", method = "pairwise")
```

```
# Explore - main effect condition
CrossTable(TotalDataframe$Correct.stress, TotalDataframe$Condition, fisher= TRUE, chisq = TRUE, expected
= TRUE, sresid = TRUE, format = "SPSS", simulate.p.value=TRUE, B=100000)

# Graph - main effect condition
ggplot(data = subset(TotalDataframe, !is.na(Correct.stress)),
    aes(x = Condition,
      fill = Correct.stress)) +
 geom_bar(position = "fill") +
 scale_y_continuous(breaks = seq(0, 1, .2),
          label = percent) +
 scale_fill_brewer(palette = "Set2") +
 labs(y = "Scores (in Percentages) on Spanish Lexical Stress Placement",
    fill = "Spanish Lexical Stress Placement",
    x = "Training Condition") +
 theme_minimal()

# Contrast analyses - explore differences between other conditions
em <- emmeans(Model4_b, "Condition")
contrast(em, adjust = "bonferroni", method = "pairwise")

# Graph - main effect WM
library(dplyr)
plotdata <- subset(TotalDataframe, !is.na(Correct.stress)) %>%
 group_by(Correct.stress) %>%
 summarize(n = n(),
      mean = mean(WorkingMemory),
      sd = sd(WorkingMemory),
      se = sd / sqrt(n),
      ci = qt(0.975, df = n - 1) * sd / sqrt(n))

ggplot(plotdata,
    aes(x = Correct.stress,
      y = mean,
      group = 1)) +
 geom_point(size = 3) +
 geom_line() +
 geom_errorbar(aes(ymin = mean - se,
          ymax = mean + se),
        width = .1) +
```

```
 labs(y = "Mean WM scores",
    x = "Lexical stress placement",
    title = "Main effect of WM score on Lexical Stress Placement")

# Graph - main effect rhythm
plotdata <- subset(TotalDataframe, !is.na(Correct.stress)) %>%
 group_by(Correct.stress) %>%
 summarize(n = n(),
      mean = mean(Rhythm),
      sd = sd(Rhythm),
      se = sd / sqrt(n),
      ci = qt(0.975, df = n - 1) * sd / sqrt(n))

ggplot(plotdata,
    aes(x = Correct.stress,
      y = mean,
      group = 1)) +
 geom_point(size = 3) +
 geom_line() +
 geom_errorbar(aes(ymin = mean - se,
          ymax = mean + se),
        width = .1) +
 labs(y = "Mean Rhythm scores",
    x = "Lexical stress placement",
    title = "Main effect of Rhythm on Lexical Stress Placement")

# Graph - main effect melody
plotdata <- subset(TotalDataframe, !is.na(Correct.stress)) %>%
 group_by(Correct.stress) %>%
 summarize(n = n(),
      mean = mean(Melody),
      sd = sd(Melody),
      se = sd / sqrt(n),
      ci = qt(0.975, df = n - 1) * sd / sqrt(n))

ggplot(plotdata,
    aes(x = Correct.stress,
      y = mean,
      group = 1)) +
 geom_point(size = 3) +
```

```
 geom_line() +
 geom_errorbar(aes(ymin = mean - se,
          ymax = mean + se),
        width = .1) +
 labs(y = "Mean Melody scores",
    x = "Lexical stress placement",
    title = "Main effect of Melody on Lexical Stress Placement")

# Graph - main effect accent
plotdata <- subset(TotalDataframe, !is.na(Correct.stress)) %>%
 group_by(Correct.stress) %>%
 summarize(n = n(),
      mean = mean(Accent),
      sd = sd(Accent),
      se = sd / sqrt(n),
      ci = qt(0.975, df = n - 1) * sd / sqrt(n))

ggplot(plotdata,
    aes(x = Correct.stress,
      y = mean,
      group = 1)) +
 geom_point(size = 3) +
 geom_line() +
 geom_errorbar(aes(ymin = mean - se,
          ymax = mean + se),
        width = .1) +
 labs(y = "Mean Accent scores",
    x = "Lexical stress placement",
    title = "Main effect of Accent on Lexical Stress Placement")

# Explore - main effect continuous variables (WM and musical aptitude scores)
describeBy(TotalDataframe, group = "Correct.stress")

# Explore - main effect rule
CrossTable(TotalDataframe$Correct.stress, TotalDataframe$Rule_Accent_vs_NoAccent, fisher= TRUE, chisq =
TRUE, expected = TRUE, sresid = TRUE, format = "SPSS", simulate.p.value=TRUE, B=100000)

# Graph - main effect rule
ggplot(data = subset(TotalDataframe, !is.na(Correct.stress)),
    aes(x = Rule_Accent_vs_NoAccent,
```

```
     fill = Correct.stress)) +
 geom_bar(position = "fill") +
 scale_y_continuous(breaks = seq(0, 1, .2),
          label = percent) +
 scale_fill_brewer(palette = "Set2") +
 labs(y = "Scores (in Percentages) on the Read-Aloud Tasks",
    fill = "Spanish Lexical Stress Placement",
    x = "Stress Rule") +
 theme_minimal()

### TWO-WAY INTERACTIONS
library("jtools")
library("interactions")

# Graph - interaction between trial and condition (condition by trial)
cat_plot(Model4_b, pred = Trial, modx = Condition, geom = "bar",
     legend.main = "Training condition",
     x.label = "Time of testing",
     y.label = "Lexical stress placement")

# Contrast analyses - condition by trail
em <- emmeans(Model4_b, ~ Condition | Trial)
contrast(em, adjust = "bonferroni", method = "pairwise")

# Graph - interaction between trial and condition (trial by condition)
cat_plot(Model4_b, pred = Condition, modx = Trial, geom = "bar",
     legend.main = "Time of Testing",
     x.label = "Training Condition",
     y.label = "Ratio Correct Spanish Lexical Stress Placement")

# Contrast analyses - trial by condition
em <- emmeans(Model4_b, ~ Trial | Condition)
contrast(em, adjust = "bonferroni", method = "pairwise")

# Graph - interaction between trial and rule
cat_plot(Model4_b, pred = Trial, modx = Rule_Accent_vs_NoAccent, geom = "bar",
     legend.main = "Stress Rule",
     x.label = "Time of Testing",
     y.label = "Ratio Correct Spanish Lexical Stress Placement")
```

```
cat_plot(Model4_b, pred = Rule_Accent_vs_NoAccent, modx = Trial , geom = "bar")

# contrast analyses - trial by rule
em <- emmeans(Model4_b, ~ Rule_Accent_vs_NoAccent | Trial)
contrast(em, adjust = "bonferroni", method = "pairwise")

em <- emmeans(Model4_b, ~ Trial | Rule_Accent_vs_NoAccent)
contrast(em, adjust = "bonferroni", method = "pairwise")

# Graph - interaction between WM and trial
interact_plot(Model4_b, pred = WorkingMemoryscale, modx = Trial,
      legend.main = "Time of testing",
      x.label = "WM score (scaled)",
      y.label = "Lexical stress placement",
      main.title = "Effect of WM score on Lexical stress placement by Trial")

# Graph - interaction between melody and trial
interact_plot(Model4_b, pred = Melodyscale, modx = Trial,
      legend.main = "Time of testing",
      x.label = "Melody score (scaled)",
      y.label = "Lexical stress placement",
      main.title = "Effect of Melody score on Lexical stress placement by Trial")

# Graph - interaction between rhythm and trial
interact_plot(Model4_b, pred = Rhythmscale, modx = Trial,
      legend.main = "Time of testing",
      x.label = "Rhythm score (scaled)",
      y.label = "Lexical stress placement",
      main.title = "Effect of Rhythm score on Lexical stress placement by Trial")

# Graph - interaction between accent and trial
interact_plot(Model4_b, pred = Accentscale, modx = Trial,
      legend.main = "Time of testing",
      x.label = "Accent score (scaled)",
      y.label = "Lexical stress placement",
      main.title = "Effect of Accent score on Lexical stress placement by Trial")

### THREE-WAY INTERACTIONS
# Graph - interaction between trial, rule, and condition
cat_plot(Model4_b, pred = Trial, modx = Condition, mod2 = Rule_Accent_vs_NoAccent, geom = "bar",
```

```
    legend.main = "Training condition",
    x.label = "Time of testing",
    y.label = "Lexical stress placement",
    main.title = "Effect Stress Rule by Time of Testing on Lexical stress placement")

# Graph - interaction between trial, condition, and WM
interact_plot(Model4_b, pred = WorkingMemoryscale, modx = Condition, mod2 = Trial,
      legend.main = "Condition",
      x.label = "WM Score (scaled)",
      y.label = "Ratio Correct Spanish Lexical Stress Placement")

# Contrast analyses - interaction between trial, condition, and WM
emtrends(Model4_b, pairwise ~ Trial | Condition, var="WorkingMemoryscale", adjust = "bonferroni")
emtrends(Model4_b, pairwise ~ Condition | Trial , var="WorkingMemoryscale", adjust = "bonferroni")

# Graph - interaction between trial, condition, and melody
interact_plot(Model4_b, pred = Melodyscale, modx = Condition, mod2 = Trial,
      legend.main = "Condition",
      x.label = "Melody Score (scaled)",
      y.label = "Ratio Correct Spanish Lexical Stress Placement")

# Contrast analyses - interaction between trial, condition, and melody
emtrends(Model4_b, pairwise ~ Condition|Trial, var = "Melodyscale", adjust = "bonferroni")
emtrends(Model4_b, pairwise ~ Trial|Condition, var = "Melodyscale", adjust = "bonferroni")



# Graph - interaction between trial, condition, and rhythm
interact_plot(Model4_b, pred = Rhythmscale, modx = Condition, mod2 = Trial,
      legend.main = "Condition",
      x.label = "Rhythm Score (scaled)",
      y.label = "Ratio Correct Spanish Lexical Stress Placement")

# Contrast analyses - interaction between trial, condition, and rhythm
emtrends(Model4_b, pairwise ~ Condition|Trial, var = "Rhythmscale", adjust = "bonferroni")
emtrends(Model4_b, pairwise ~ Trial|Condition, var = "Rhythmscale", adjust = "bonferroni")

# Graph - interaction between trial, condition, and accent
interact_plot(Model4_b, pred = Accentscale, modx = Condition, mod2 = Trial,
      legend.main = "Condition",
      x.label = "Accent Score (scaled)",
```

```
    y.label = "Ratio Correct Spanish Lexical Stress Placement")

# Contrast analyses - interaction between trial, condition, and accent
emtrends(Model4_b, pairwise ~ Trial|Condition, var = "Accentscale", adjust = "bonferroni")
emtrends(Model4_b, pairwise ~ Condition|Trial, var = "Accentscale", adjust = "bonferroni")

# Graph - interaction between trial, rule, and WM
interact_plot(Model4_b, pred = WorkingMemoryscale, modx = Rule_Accent_vs_NoAccent, mod2 = Trial,
    legend.main = "Stress rule",
    x.label = "WM score (scaled)",
    y.label = "Lexical stress placement",
    main.title = "Effect of WM scores on Lexical Stress Placement by Stress Rule")

# Graph - interaction between trial, rule, and melody
interact_plot(Model4_b, pred = Melodyscale, modx = Rule_Accent_vs_NoAccent, mod2 = Trial,
    legend.main = "Stress Rule",
    x.label = "Melody Score (scaled)",
    y.label = "Ratio Correct Spanish Lexical Stress Placement")

# Contrast analyses - interaction between trial, rule, and melody
emtrends(Model4_b, pairwise ~ Rule_Accent_vs_NoAccent|Trial, var = "Melodyscale", adjust = "bonferroni")
emtrends(Model4_b, pairwise ~ Trial|Rule_Accent_vs_NoAccent, var = "Melodyscale", adjust = "bonferroni")

# Graph - interaction between trial, rule, and rhythm
interact_plot(Model4_b, pred = Rhythmscale, modx = Rule_Accent_vs_NoAccent, mod2 = Trial,
    legend.main = "Stress rule",
    x.label = "Rhythm Score (scaled)",
    y.label = "Ratio Correct Spanish Lexical Stress Placement")

# Contrast analyses - interaction between trial, rule, and rhythm
emtrends(Model4_b, pairwise ~ Rule_Accent_vs_NoAccent|Trial, var = "Rhythmscale", adjust = "bonferroni")
emtrends(Model4_b, pairwise ~ Trial|Rule_Accent_vs_NoAccent, var = "Rhythmscale", adjust = "bonferroni")

# Graph - interaction between trial, rule, and accent
interact_plot(Model4_b, pred = Accentscale, modx = Rule_Accent_vs_NoAccent, mod2 = Trial,
    legend.main = "Stress rule",
    x.label = "Accent score (scaled)",
    y.label = "Lexical stress placement",
    main.title = "Effect of Accent scores on Lexical Stress Placement by Training Condition")
```

```
### FOUR-WAY INTERACTIONS
## split file by trial (separate regressions for T1, T2, and T3)
# Graph - interaction between trial, condition, WM, and rule
# T1
interact_plot(Model_T1, pred = WorkingMemoryscale, modx = Rule_Accent_vs_NoAccent, mod2 = Condition,
      legend.main = "Stress Rule",
      x.label = "WM Score (scaled)",
      y.label = "Ratio Correct Spanish Lexical Stress Placement")
# T2
interact_plot(Model_T2, pred = WorkingMemoryscale, modx = Rule_Accent_vs_NoAccent, mod2 = Condition,
      legend.main = "Stress Rule",
      x.label = "WM Score (scaled)",
      y.label = "Ratio Correct Spanish Lexical Stress Placement")
#T3
interact_plot(Model_T3, pred = WorkingMemoryscale, modx = Rule_Accent_vs_NoAccent, mod2 = Condition,
      legend.main = "Stress Rule",
      x.label = "WM Score (scaled)",
      y.label = "Ratio Correct Spanish Lexical Stress Placement")

# Graph - interaction between trial, condition, melody and rule
# T1
interact_plot(Model_T1, pred = Melodyscale, modx = Rule_Accent_vs_NoAccent, mod2 = Condition,
      legend.main = "Stress Rule",
      x.label = "Melody Score (scaled)",
      y.label = "Ratio Correct Spanish Lexical Stress Placement")
# T2
interact_plot(Model_T2, pred = Melodyscale, modx = Rule_Accent_vs_NoAccent, mod2 = Condition,
      legend.main = "Stress Rule",
      x.label = "Melody Score (scaled)",
      y.label = "Ratio Correct Spanish Lexical Stress Placement")
#T3
interact_plot(Model_T3, pred = Melodyscale, modx = Rule_Accent_vs_NoAccent, mod2 = Condition,
      legend.main = "Stress Rule",
      x.label = "Melody Score (scaled)",
      y.label = "Ratio Correct Spanish Lexical Stress Placement")

# Graph - interaction between trial, condition, rhythm, and rule
# T1
interact_plot(Model_T1, pred = Rhythmscale, modx = Rule_Accent_vs_NoAccent, mod2 = Condition,
      legend.main = "Stress Rule",
```

```
    x.label = "Rhythm Score (scaled)",
    y.label = "Ratio Correct Spanish Lexical Stress Placement")

# T2
interact_plot(Model_T2, pred = Rhythmscale, modx = Rule_Accent_vs_NoAccent, mod2 = Condition,
    legend.main = "Stress Rule",
    x.label = "Rhythm Score (scaled)",
    y.label = "Ratio Correct Spanish Lexical Stress Placement")
#T3
interact_plot(Model_T3, pred = Rhythmscale, modx = Rule_Accent_vs_NoAccent, mod2 = Condition,
    legend.main = "Stress Rule",
    x.label = "Rhythm Score (scaled)",
    y.label = "Ratio Correct Spanish Lexical Stress Placement")

# Graph - interaction between trial, condition, accent, and rule
# T1
interact_plot(Model_T1, pred = Accentscale, modx = Rule_Accent_vs_NoAccent, mod2 = Condition,
    legend.main = "Stress Rule",
    x.label = "Accent Score (scaled)",
    y.label = "Ratio Correct Spanish Lexical Stress Placement")
# T2
interact_plot(Model_T2, pred = Accentscale, modx = Rule_Accent_vs_NoAccent, mod2 = Condition,
    legend.main = "Stress Rule",
    x.label = "Accent Score (scaled)",
    y.label = "Ratio Correct Spanish Lexical Stress Placement")
#T3
interact_plot(Model_T3, pred = Accentscale, modx = Rule_Accent_vs_NoAccent, mod2 = Condition,
    legend.main = "Stress Rule",
    x.label = "Accent Score (scaled)",
    y.label = "Ratio Correct Spanish Lexical Stress Placement")

## Change 1 continuous variable (WM) into a categorical variable to graph interaction between 2 continuous and
2 categorical variables
# changing WM variable into a categorical variable on T1
TotalDataframe_T1$WMcat1<-cut(TotalDataframe_T1$WorkingMemory, breaks = c(0,3,5,7), labels=c("low",
"mid", "high"))
summary(TotalDataframe_T1$WMcat1)
# changing WM variable into a categorical variable on T2
TotalDataframe_T2$WMcat1<-cut(TotalDataframe_T2$WorkingMemory, breaks = c(0,3,5,7), labels=c("low",
"mid", "high"))
```

```
summary(TotalDataframe_T2$WMcat1)
# changing WM variable into a categorical variable on T3
TotalDataframe_T3$WMcat1<-cut(TotalDataframe_T3$WorkingMemory, breaks = c(0,3,5,7), labels=c("low",
"mid", "high"))
summary(TotalDataframe_T3$WMcat1)

# Graph - interaction between trial, condition, melody, and WM
# create new interaction terms (with new categorical variable)
glmT1_m <- glm(Correct.stress ~ Condition * WMcat1 * Melodyscale, family = "binomial", data =
TotalDataframe_T1)
glmT2_m <- glm(Correct.stress ~ Condition * WMcat1 * Melodyscale, family = "binomial", data =
TotalDataframe_T2)
glmT3_m <- glm(Correct.stress ~ Condition * WMcat1 * Melodyscale, family = "binomial", data =
TotalDataframe_T3)

# T1
interact_plot(glmT1_m, pred = Melodyscale, modx = WMcat1, mod2 = Condition,
      legend.main = "WM Score",
      x.label = "Melody Score (scaled)",
      y.label = "Ratio Correct Spanish Lexical Stress Placement")
# T2
interact_plot(glmT2_m, pred = Melodyscale, modx = WMcat1, mod2 = Condition,
      legend.main = "WM Score",
      x.label = "Melody Score (scaled)",
      y.label = "Ratio Correct Spanish Lexical Stress Placement")
# T3
interact_plot(glmT3_m, pred = Melodyscale, modx = WMcat1, mod2 = Condition,
      legend.main = "WM Score",
      x.label = "Melody Score (scaled)",
      y.label = "Ratio Correct Spanish Lexical Stress Placement")

# Graph - interaction between trial, condition, rhythm, and WM
# create new interaction terms (with new categorical variable)
glmT1_r <- glm(Correct.stress ~ Condition * WMcat1 * Rhythmscale, family = "binomial", data =
TotalDataframe_T1)
glmT2_r <- glm(Correct.stress ~ Condition * WMcat1 * Rhythmscale, family = "binomial", data =
TotalDataframe_T2)
glmT3_r <- glm(Correct.stress ~ Condition * WMcat1 * Rhythmscale, family = "binomial", data =
TotalDataframe_T3)
```

```
# T1
interact_plot(glmT1_r, pred = Rhythmscale, modx = WMcat1, mod2 = Condition,
      legend.main = "WM Score",
      x.label = "Rhythm score (scaled)",
      y.label = "Ratio Correct Spanish Lexical Stress Placement")
# T2
interact_plot(glmT2_r, pred = Rhythmscale, modx = WMcat1, mod2 = Condition,
      legend.main = "WM Score",
      x.label = "Rhythm Score (scaled)",
      y.label = "Ratio Correct Spanish Lexical Stress Placement")
# T3
interact_plot(glmT3_r, pred = Rhythmscale, modx = WMcat1, mod2 = Condition,
      legend.main = "WM Score",
      x.label = "Rhythm Score (scaled)",
      y.label = "Ratio Correct Spanish Lexical Stress Placement")

# Graph - 4-way interaction between trial, condition, accent, and WM
# create new interaction terms (with new categorical variable)
glmT1_a <- glm(Correct.stress ~ Condition * WMcat1 * Accentscale, family = "binomial", data =
TotalDataframe_T1)
glmT2_a <- glm(Correct.stress ~ Condition * WMcat1 * Accentscale, family = "binomial", data =
TotalDataframe_T2)
glmT3_a <- glm(Correct.stress ~ Condition * WMcat1 * Accentscale, family = "binomial", data =
TotalDataframe_T3)

# T1
interact_plot(glmT1_a, pred = Accentscale, modx = WMcat1, mod2 = Condition,
      legend.main = "WM Score",
      x.label = "Accent Score (scaled)",
      y.label = "Ratio Correct Spanish Lexical Stress Placement")
# T2
interact_plot(glmT2_a, pred = Accentscale, modx = WMcat1, mod2 = Condition,
      legend.main = "WM Score",
      x.label = "Accent Score (scaled)",
      y.label = "Ratio Correct Spanish Lexical Stress Placement")
# T3
interact_plot(glmT3_a, pred = Accentscale, modx = WMcat1, mod2 = Condition,
      legend.main = "WM Score",
      x.label = "Accent Score (scaled)",
      y.label = "Ratio Correct Spanish Lexical Stress Placement")
```

Appendix J - Stepwise procedure regression model

Appendix J.1 - Predictors in regression model

	Predictors
Baseline Model	~ Musicgroup + (1| StressRule:Item)
Model 1	~ TimeofTesting + Condition + Melody + WM + StressRule + Musicgroup + (1| StressRule:Item)
Model 2	~ TimeofTesting + Condition + Melody + WM + StressRule + TimeofTesting:Condition + TimeofTesting:Melody + TimeofTesting:StressRule + Musicgroup + (1 | StressRule:Item)
Model 3	~ TimeofTesting + Condition + Melody + WM + StressRule + TimeofTesting:Condition + TimeofTesting:Melody + TimeofTesting:StressRule + TimeofTesting:Condition:WM + TimeofTesting:Condition:Melody + TimeofTesting:Condition:Rhythm + TimeofTesting:Condition:Accent + TimeofTesting:StressRule:Melody + TimeofTesting:StressRule:Rhythm + Musicgroup + (1 | StressRule:Item)
Model 4	~ TimeofTesting + Condition + Melody + WM + StressRule + TimeofTesting:Condition + TimeofTesting:Melody + TimeofTesting:StressRule + TimeofTesting:Condition:WM + TimeofTesting:Condition:Melody + TimeofTesting:Condition:Rhythm + TimeofTesting:Condition:Accent + TimeofTesting:StressRule:Melody + TimeofTesting:StressRule:Rhythm + TimeofTesting:Condition:WM:Melody + TimeofTesting:Condition:WM:Rhythm + TimeofTesting:Condition:WM:Accent + TimeofTesting:Condition:StressRule:WM +TimeofTesting:Condition:StressRule:Melody + TimeofTesting:Condition:StressRule:Accent + Musicgroup + (1 | StressRule:Item)

Appendix J.2 - Estimates and standard errors of predictors in regression model

		Baseline Model	Model 1	Model 2	Model 3	Model 4
Intercept		-.08 (.33)	-.78 (.50)*	-.81 (.52)	-1.21 (.73)	-1.18 (.83)
Time of testing	*T2*		2.51 (.15)***	2.43 (.35)***	2.37 (.66)***	2.52 (.76)***
	T3		2.46 (.15)***	2.31 (.35)***	2.90 (.67)***	3.35 (.75)***
Condition	*AV-B-Perc*		-.31 (.17).	.37 (.38)	.57 (.62)	-.24 (.82)
	AV-B-Prod		-.45 (.17)**	.23 (.38)	-.16 (.67)	-1.95 (.94)*
	AV-M-Perc		-.21 (.17)	.16 (.36)	.51 (.67)	.47 (.81)
	AV-M-Prod		.11 (.16)	.39 (.37)	-.03 (.68)	-.45 (.95)
Melody score			.22 (.06)***	-.02 (.12)	-.57 (.50)	.15 (.96)
WM score			.18 (.06)***	.18 (.55)***	-.70 (.33)*	-.61 (.71)
Stress rule (written accent)			-1.68 (.69)*	-3.19 (.72)***	-3.50 (.79)***	-3.88 (1.08)***
Time of testing * Condition	*T2 * AV-B-Perc*			-.80 (.46).	-.79 (.75)	.34 (.99)
	*T3 * AV-B-Perc*			-.92 (.46)*	-1.46 (.75).	-1.13 (.98)
	*T2 * AV-B-Prod*			-.73 (.46)	.02 (.79)	1.22 (1.11)
	*T3 * AV-B-Prod*			-.96 (.46)*	-.64 (.79)	-.28 (1.13)
	*T2 * AV-M-Perc*			-.55 (.44)	-.77 (.81)	-.83 (.98)
	*T3 * AV-M-Perc*			-.37 (.44)	-1.05 (.81)	-1.44 (.97)
	*T2 * AV-M-Prod*			-.35 (.45)	.91 (.82)	1.08 (1.12)
	*T3 * AV-M-Prod*			-.31 (.45)	.70 (.85)	-.08 (1.16)
Time of testing * Melody	*T2 * Melody*			.29 (.15).	.29 (.62)	-.42 (1.23)
	*T3 * Melody*			.32 (.15)*	.83 (.62)	1.20 (1.22)
Time of testing * Stress rule	*T2 * Stress rule (written accent)*			1.78 (.37)***	2.00 (.41)***	2.46 (.76)**
	*T3 * Stress rule (written accent)*			1.94 (.37)***	2.16 (.41)***	2.63 (.76)***
Time of testing * Condition * WM score	*T1 * AV * WM score*				.77 (.68)	.60 (.96)
	*T2 * AV * WM score*				1.78 (.54)**	.95 (.91)
	*T3 * AV * WM score*				.98 (.53).	.37 (.89)
	*T1 * AV-B-perc * WM score*				1.42 (.500)**	-.36 (1.30)
	*T2 * AV-B-perc * WM score*				.39 (.43)	1.30 (1.09)
	*T3 * AV-B-perc * WM score*				.33 (.42)	-.20 (1.06)

		Baseline Model	Model 1	Model 2	Model 3	Model 4
	*T1 * AV-B-prod * WM score*				-.05 (.49)	-2.60 (1.20)*
	*T2 * AV-B-prod * WM score*				.80 (.42).	.37 (.99)
	*T3 * AV-B-prod * WM score*				1.17 (.41)**	.44 (1.00)
	*T1 * AV-M-perc * WM score*				.77 (.45).	.95 (.86)
	*T2 * AV-M-perc * WM score*				.88 (.39)*	1.34 (.82)
	*T3 * AV-M-perc * WM score*				.80 (.39)*	-.08 (.84)
	*T1 * AV-M-prod * WM score*				1.82 (.51)***	2.54 (1.25)*
	*T2 * AV-M-prod * WM score*				.45 (.43)	.23 (.94)
	*T3 * AV-M-prod * WM score*				*NA*[a]	*NA*
Time of testing * Condition * Melody score	*T1 * AV * Melody score*				*NA*	*NA*
	*T2 * AV * Melody score*				*NA*	*NA*
	*T3 * AV * Melody score*				*NA*	*NA*
	*T1 * AV-B-perc * Melody score*				.48 (.58)	-.68 (1.05)
	*T2 * AV-B-perc * Melody score*				.82 (.44).	1.64 (.88).
	*T3 * AV-B-perc * Melody score*				-.06 (.43)	-1.35 (.83)
	*T1 * AV-B-prod * Melody score*				1.10 (.66).	2.55 (1.77)
	*T2 * AV-B-prod * Melody score*				.76 (.53)	4.18 (1.33)**
	*T3 * AV-B-prod * Melody score*				.28 (.52)	3.94 (1.38)**
	*T1 * AV-M-perc * Melody score*				.26 (.59)	-.48 (1.10)
	*T2 * AV-M-perc * Melody score*				-.11 (.43)	-.74 (.89)
	*T3 * AV-M-perc * Melody score*				-.88 (.42)*	-1.55 (.88).
	*T1 * AV-M-prod * Melody score*				.77 (.71)	.31 (1.23)
	*T2 * AV-M-prod * Melody score*				.41 (.56)	-.46 (1.08)
	*T3 * AV-M-prod * Melody score*				-1.49 (.64)*	-.83 (1.12)
Time of testing * Condition * Rhythm score	*T1 * AV * Rhythm score*				-.19 (.41)	-.36 (.67)
	*T2 * AV * Rhythm score*				.39 (.32)	.21 (.53)
	*T3 * AV * Rhythm score*				.22 (.32)	-.02 (.51)
	*T1 * AV-B-perc * Rhythm score*				.12 (.33)	.49 (.39)
	*T2 * AV-B-perc * Rhythm score*				-.08 (.24)	.05 (.29)
	*T3 * AV-B-perc * Rhythm score*				-.06 (.23)	.29 (.28)

		Baseline Model	Model 1	Model 2	Model 3	Model 4
	*T1 * AV-B-prod * Rhythm score*				.91 (.34)**	.83 (.57)
	*T2 * AV-B-prod * Rhythm score*				.38 (.25)	1.13 (.36)**
	*T3 * AV-B-prod * Rhythm score*				.71 (.25)**	2.00 (.43)***
	*T1 * AV-M-perc * Rhythm score*				.20 (.37)	.44 (.54)
	*T2 * AV-M-perc * Rhythm score*				-.02 (.28)	.48 (.39)
	*T3 * AV-M-perc * Rhythm score*				.15 (.28)	.14 (.40)
	*T1 * AV-M-prod * Rhythm score*				-.57 (.51)	-.89 (.65)
	*T2 * AV-M-prod * Rhythm score*				1.02 (.42)*	1.24 (.48)*
	*T3 * AV-M-prod * Rhythm score*				1.57 (.49)**	1.52 (.54)**
Time of testing * Condition * Accent score	*T1 * AV * Accent score*				-.64 (.60)	.08 (1.26)
	*T2 * AV * Accent score*				-.01 (.43)	-.39 (1.10)
	*T3 * AV * Accent score*				.47 (.43)	.80 (1.07)
	*T1 * AV-B-perc * Accent score*				-1.24 (.44)**	-.34 (.76)
	*T2 * AV-B-perc * Accent score*				.23 (.32)	-1.32 (.73).
	*T3 * AV-B-perc * Accent score*				.41 (.31)	-.04 (.61)
	*T1 * AV-B-prod * Accent score*				-.20 (.35)	.71 (.91)
	*T2 * AV-B-prod * Accent score*				-.32 (.29)	-1.06 (.64).
	*T3 * AV-B-prod * Accent score*				-.60 (.28)*	-1.98 (.69)**
	*T1 * AV-M-perc * Accent score*				.23 (.54)	1.00 (.77)
	*T2 * AV-M-perc * Accent score*				-.25 (.37)	-1.01 (.69)
	*T3 * AV-M-perc * Accent score*				.19 (.38)	-.76 (.66)
	*T1 * AV-M-prod * Accent score*				.52 (.41)	2.02 (.91)*
	*T2 * AV-M-prod * Accent score*				.37 (.31)	.53 (.57)
	*T3 * AV-M-prod * Accent score*				.61 (.35).	-.63 (.60)
Time of testing * Melody score * Stress rule	*T1 * Melody score * Stress rule (written accent)*				.21 (.35)	2.93 (2.32)
	*T2 * Melody score * Stress rule (written accent)*				.43 (.19)*	1.50 (.58)**
	*T3 * Melody score * Stress rule (written accent)*				.63 (.19)***	.78 (.53)
Time of testing * Rhythm score * Stress rule	*T1 * Rhythm score * Stress rule (written accent)*				-.66 (.44)	-.84 (.64)
	*T2 * Rhythm score * Stress rule (written accent)*				-.37 (.19).	-.46 (.24).
	*T3 * Rhythm score * Stress rule (written accent)*				-.39 (.19)*	-.66 (.24)**

		Baseline Model	Model 1	Model 2	Model 3	Model 4
Time of testing * Condition * WM score * Melody score	*T1 * AV * WM score * Melody score*					1.70 (1.94)
	*T2 * AV * WM score * Melody score*					-.24 (1.52)
	*T3 * AV * WM score * Melody score*					1.51 (1.43)
	*T1 * AV-B-perc * WM score * Melody score*					-.79 (.44).
	*T2 * AV-B-perc * WM score * Melody score*					-.46 (.32)
	*T3 * AV-B-perc * WM score * Melody score*					-.55 (.32).
	*T1 * AV-B-prod * WM score * Melody score*					2.63 (1.34)*
	*T2 * AV-B-prod * WM score * Melody score*					2.50 (.90)**
	*T3 * AV-B-prod * WM score * Melody score*					2.86 (.94)**
	*T1 * AV-M-perc * WM score * Melody score*					.60 (.42)
	*T2 * AV-M-perc * WM score * Melody score*					-.91 (.30)**
	*T3 * AV-M-perc * WM score * Melody score*					.01 (.30)
	*T1 * AV-M-prod * WM score * Melody score*					-.27 (.88)
	*T2 * AV-M-prod * WM score * Melody score*					-1.23 (.58)*
	*T3 * AV-M-prod * WM score * Melody score*					-3.01 (.78)***
Time of testing * Condition * WM score * Rhythm score	*T1 * AV * WM score * Rhythm score*					-1.01 (1.20)
	*T2 * AV * WM score * Rhythm score*					-.14 (1.01)
	*T3 * AV * WM score * Rhythm score*					-1.79 (.91)*
	*T1 * AV-B-perc * WM score * Rhythm score*					1.67 (1.00).
	*T2 * AV-B-perc * WM score * Rhythm score*					.64 (.69)
	*T3 * AV-B-perc * WM score * Rhythm score*					1.22 (.67).
	*T1 * AV-B-prod * WM score * Rhythm score*					.37 (.56)
	*T2 * AV-B-prod * WM score * Rhythm score*					.99 (.33)**
	*T3 * AV-B-prod * WM score * Rhythm score*					1.67 (.40)***
	*T1 * AV-M-perc * WM score * Rhythm score*					.67 (.75)
	*T2 * AV-M-perc * WM score * Rhythm score*					1.02 (.55).
	*T3 * AV-M-perc * WM score * Rhythm score*					-.23 (.56)
	*T1 * AV-M-prod * WM score * Rhythm score*					2.22 (1.41)
	*T2 * AV-M-prod * WM score * Rhythm score*					.04 (.83)
	*T3 * AV-M-prod * WM score * Rhythm score*					2.11 (.88)*

		Baseline Model	Model 1	Model 2	Model 3	Model 4
Time of testing * Condition * WM score * Accent score	*T1 * AV * WM score * Accent score*					2.27 (2.31)
	*T2 * AV * WM score * Accent score*					.61 (1.77)
	*T3 * AV * WM score * Accent score*					2.57 (1.74)
	*T1 * AV-B-perc * WM score * Accent score*					1.00 (1.14)
	*T2 * AV-B-perc * WM score * Accent score*					-2.87 (.79)***
	*T3 * AV-B-perc * WM score * Accent score*					-1.71 (.73)*
	*T1 * AV-B-prod * WM score * Accent score*					2.84 (.97)**
	*T2 * AV-B-prod * WM score * Accent score*					.42 (.65)
	*T3 * AV-B-prod * WM score * Accent score*					.71 (.67)
	*T1 * AV-M-perc * WM score * Accent score*					-.37 (.46)
	*T2 * AV-M-perc * WM score * Accent score*					.30 (.31)
	*T3 * AV-M-perc * Accent score*					.19 (.32)
	*T1 * AV-M-prod * WM score * Accent score*					-3.02 (1.57).
	*T2 * AV-M-prod * WM score * Accent score*					.92 (.46)*
	*T3 * AV-M-prod * WM score * Accent score*					.91 (.51).
Time of testing * Condition * Stress rule * WM score	*T1 * AV * Stress rule (written accent) * WM score*					.41 (1.33)
	*T2 * AV * Stress rule (written accent) * WM score*					1.51 (.75)*
	*T3 * AV * Stress rule (written accent) * WM score*					1.10 (.64).
	*T1 * AV-B-perc * Stress rule (written accent) * WM score*					-.19 (1.20)
	*T2 * AV-B-perc * Stress rule (written accent) * WM score*					-1.01 (.56).
	*T3 * AV-B-perc * Stress rule (written accent) * WM score*					-.19 (.52)
	*T1 * AV-B-prod * Stress rule (written accent) * WM score*					-.89 (.96)
	*T2 * AV-B-prod * Stress rule (written accent) * WM score*					.07 (.52)
	*T3 * AV-B-prod * Stress rule (written accent) * WM score*					-.14 (.47)
	*T1 * AV-M-perc * Stress rule (written accent) * WM score*					1.00 (1.04)
	*T2 * AV-M-perc * Stress rule (written accent) * WM score*					.003 (.39)
	*T3 * AV-M-perc * Stress rule (written accent) * WM score*					.97 (.40)*
	*T1 * AV-M-prod * Stress rule (written accent) * WM score*					.10 (.87)
	*T2 * AV-M-prod * Stress rule (written accent) * WM score*					-.29 (.56)
	*T3 * AV-M-prod * Stress rule (written accent) * WM score*					-.59 (.67)

		Baseline Model	Model 1	Model 2	Model 3	Model 4
Time of testing * Condition * Stress rule * Melody score	*T1 * AV * Stress rule (written accent) * Melody score*					*NA*
	*T2 * AV * Stress rule (written accent) * Melody score*					*NA*
	*T3 * AV * Stress rule (written accent) * Melody score*					*NA*
	*T1 * AV-B-perc * Stress rule (written accent) * Melody score*					-2.25 (2.35)
	*T2 * AV-B-perc * Stress rule (written accent) * Melody score*					-1.73 (.78)*
	*T3 * AV-B-perc * Stress rule (written accent) * Melody score*					.40 (.70)
	*T1 * AV-B-prod * Stress rule (written accent) * Melody score*					-3.20 (2.84)
	*T2 * AV-B-prod * Stress rule (written accent) * Melody score*					-2.05 (1.06).
	*T3 * AV-B-prod * Stress rule (written accent) * Melody score*					-1.51 (.93)
	*T1 * AV-M-perc * Stress rule (written accent) * Melody score*					-3.01 (2.54)
	*T2 * AV-M-perc * Stress rule (written accent) * Melody score*					-.88 (.73)
	*T3 * AV-M-perc * Stress rule (written accent) * Melody score*					-.79 (.69)
	*T1 * AV-M-prod * Stress rule (written accent) * Melody score*					-2.70 (2.55)
	*T2 * AV-M-prod * Stress rule (written accent) * Melody score*					-.43 (.98)
	*T3 * AV-M-prod * Stress rule (written accent) * Melody score*					1.15 (1.04)
Time of testing * Condition * Stress rule * Accent score	*T1 * AV * Stress rule (written accent) * Accent score*					3.80 (3.06)
	*T2 * AV * Stress rule (written accent) * Accent score*					.14 (.83)
	*T3 * AV * Stress rule (written accent) * Accent score*					.72 (.79)
	*T1 * AV-B-perc * Stress rule (written accent) * Accent score*					.07 (1.17)
	*T2 * AV-B-perc * Stress rule (written accent) * Accent score*					1.39 (.69)*
	*T3 * AV-B-perc * Stress rule (written accent) * Accent score*					.80 (.62)
	*T1 * AV-B-prod * Stress rule (written accent) * Accent score*					-.60 (1.13)
	*T2 * AV-B-prod * Stress rule (written accent) * Accent score*					-.48 (.56)
	*T3 * AV-B-prod * Stress rule (written accent) * Accent score*					-.36 (.49)
	*T1 * AV-M-perc * Stress rule (written accent) * Accent score*					-.36 (1.43)
	*T2 * AV-M-perc * Stress rule (written accent) * Accent score*					.47 (.66)
	*T3 * AV-M-perc * Stress rule (written accent) * Accent score*					1.69 (.67)*
	*T1 * AV-M-prod * Stress rule (written accent) * Accent score*					-.53 (1.46)
	*T2 * AV-M-prod * Stress rule (written accent) * Accent score*					-1.06 (.64).
	*T3 * AV-M-prod * Stress rule (written accent) * Accent score*					.51 (.71)

	Baseline Model	Model 1	Model 2	Model 3	Model 4
Music group	.40 (.16)*	.35 (.21).	.35 (.21).	.56 (.27)*	1.19 (.38) **
Deviance	2946.7	2418.6	2370.5	2211.8	1968.1
Variance item (nested under stress rule)	1.57 (1.25)	1.67 (1.29)	1.44 (1.20)	1.74 (1.32)	2.35 (1.53)
-2 Log Likelihood	1473.3	1209.3***	1185.2***	1105.9***	984.0***

Note. .: $p < .1$ *: $p < .05$ **: $p < .01$ ***: $p < .001$.

[a]*NA* values were dropped in the regression analysis and therefore analysed with Bonferroni corrected contrast analyses.